High Dimensional Separable Representations for Statistical Estimation and Controlled Sensing

by

Theodoros Tsiligkaridis

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical Engineering: Systems)
in The University of Michigan
2013

Doctoral Committee:

Professor Alfred O. Hero III, Chair
Professor Jeffrey A. Fessler
Professor Susan A. Murphy
Assistant Professor Rajesh Rao Nadakuditi
Dr. Brian M. Sadler

ABSTRACT

High Dimensional Separable Representations for Statistical Estimation and Controlled Sensing

by
Theodoros Tsiligkaridis

Chair: Alfred O. Hero III

Separable approximations are effective dimensionality reduction techniques for high dimensional data. The statistical estimation performance of separable models is largely unexplored in high dimensions and model mismatch errors need to be accurately controlled. The need for performance bounds associated with statistical estimators in sample starved settings has been a topic of great interest in the field of signal processing and high-dimensional statistics.

Many signal processing methods, including classical filtering, prediction and detection, are intimately linked to the data covariance. In multiple modality and spatio-temporal signal processing, separable models for the underlying covariance may be exploited for dimensionality reduction, improved estimation accuracy and reduction in computational complexity.

In controlled sensing (or inference), estimation performance can be greatly optimized at the expense of query design (or control). Query-based multisensor controlled sensing systems used for target localization consist of a set of sensors (possibly heterogeneous and of different modality) that collaborate (through a fusion center or

by local information sharing) to estimate the location of a target. In the centralized setting, at each time instant, a fusion center designs queries for the sensors on the presence of the target in a given region and noisy responses are obtained. For a large number of sensors and/or high-dimensional targets, separable representations of the query policies can be exploited to maintain tractability. For very large sensor networks, decentralized estimation methods are of primary interest and local message-passing techniques can be exploited to increase flexibility, robustness and scalability.

Motivated by this fundamental set of high dimensional problems, the thesis makes contributions in the following areas: (1) performance bounds for high dimensional estimation for structured Kronecker product covariance matrices, (2) optimal query design for a centralized collaborative controlled sensing system used for target localization, and (3) global convergence theory of decentralized controlled sensing for target localization.

A rich class of covariance models widely applicable to spatio-temporal settings are sums of Kronecker products (KP). For the special case of a single KP model with optional sparsity in the factors, a block-coordinate descent method used to solve the penalized MLE problem is proven to achieve a tight global MSE convergence rate in high dimensions. More generally, under a convex optimization framework, high dimensional MSE convergence rates are derived that show a fundamental tradeoff between estimation error and the approximation error induced by the dimensionality reduction on the space of covariance matrices in terms of KP's. The results improve upon the current state-of-the-art methods.

Under the minimum entropy criterion, the optimality conditions for the joint policy for control of a centralized collaborative system of sensors for target search are

derived and are shown to generalize the probabilistic bisection policy of one player. For high-dimensional targets and/or large number of players, the design of such policies become intractable. A separable bisection policy is introduced and shown to achieve the same expected information gain as the jointly optimal scheme. The MSE performance is characterized and the results are extended to the case of unknown sensor reliabilities. This centralized methodology is extended to decentralized cooperative target search where players are obtaining new noisy information as a function of their current belief and exhange local beliefs among their neighbors at each time instant. Global consistency of the decentralized sequential estimation scheme is proven and it is shown that local information sharing improves estimation performance in low signal-to-noise ratio environments.

ἀεὶ ὁ θεὸς γεωμετρεῖ

Translation: God forever geometrizes.

-Plato

ACKNOWLEDGEMENTS

First and foremost, I would like to say that this thesis would not be possible without the support of certain individuals.

I greatly thank my advisor, Professor Alfred O. Hero III, for his constant support, guidance and advice. I would especially like to thank him for giving me the freedom to tackle challenging and interesting problems, to shape my research and for presenting my work in various venues around the country and overseas. His insight and expertise in many diverse fields has proved to be very useful for my growth as a researcher and a professional. I appreciate all the time invested in our discussions, despite the busy schedules that surround our lives.

I thank my close collaborator and thesis reader, Dr. Brian M. Sadler, for his mentorship, encouragement, and for allowing me to take initiative in the research conducted at Army Research Lab throughout the summers of 2012 and 2013, which culminated in half of my PhD thesis work. I also thank my thesis committee members Prof. Susan Murphy, Prof. Jeff Fessler and Prof. Raj Nadakuditi for keeping a keen interest in my work and improving the final quality of my thesis. I also thank the graduate program coordinator, Becky Turanski, for her prompt responses and simplifications of the processes involved in graduate school and for creating a welcoming environment in the department.

This thesis would not be possible without the love and support of my family and God. I am truly grateful and blessed to have wonderful parents that have served as

role models for me and a loving younger brother for putting a smile on my face even at difficult times. I owe a lot to my dad for introducing me to the exciting direction of mathematics and computers at a young age, through the mathematics periodical Euclid and C programming. I finally thank them for teaching me the value of hard work but also the importance of harmony in life. I thank God for giving me the strength to overcome the difficulties I encountered in my journey.

I am fortunate to have had funding for the work presented in this thesis given partly by the U.S. Army Research Office (ARO), grant No. W911NF-11-1-0391, and a Rackham Engineering Award offered by the Rackham Graduate School.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

CHAPTER I

Introduction

1.1 Overview

Separable approximations are effective dimensionality reduction techniques for high dimensional data. The statistical estimation performance of separable models is largely unexplored in high dimensions and model mismatch errors need to be accurately controlled. A key performance aspect of many signal processing systems is performance bounds associated with statistical estimators in sample starved settings; a line of research that has received considerable attention in the field of high-dimensional statistics.

Many signal processing methods are intimately linked to second order measures of the data, an example being the data covariance. In multiple modality data sets and spatio-temporal signal processing, separable models for the underlying covariance may be exploited for dimensionality reduction, and as a result they can improve estimation accuracy and reduce computational complexity in the algorithms.

In controlled sensing (or controlled inference), a field that has recently gained attention in the signal processing community, estimation performance can be greatly optimized at the expense of query design (or control). A multisensor controlled sensing system used for target localization consists of a set of sensors (possibly het-

erogeneous and of different modality) that collaborate (through a fusion center or local information sharing) to estimate the location of a target, taking into account the quality of the sensors. In the centralized setting, at each time instant, a central authority (i.e., a fusion center) sequentially designs queries for these sensors on the presence of the target in a given region and the sensors yield noisy responses. This iterative process continuously refines the posterior distribution of the target such that it concentrates fast towards its true location X^*. For a large number of sensors and/or high-dimensional targets, optimal query policies become intractable and separable representations of the policies can be exploited to maintain tractability and ease of implementation. Furthermore, the sensor responses may be based on a statistic computed as a function of the data covariance. Since the covariance is generally unknown, each sensor estimates the covariance using data it collected from a certain region over a period of n time instants. Using further processing, sensors make a decision on the presence of the target (with some error). Choosing too many samples n introduces delays and consumes resources such as energy and storage, while too few samples lead to poor estimates due to high dimensionality of the covariance matrix. Such an active tracking system is illustrated in Figure 1.1.

For large-scale sensor networks, it is impractical for all sensors to transmit information to a fusion center due to finite bandwidth or power constraints. Further, sensors that are far apart from the fusion center might not be able to transmit their information reliably to the fusion center due to a combination of factors including environmental constraints, interference conditions and limited resources. In these cases, decentralized methods for active estimation of an unknown target become necessary. Here, a separable representation of the information in the network up to the current time takes the form of a collection of posterior distributions (one per agent). Our

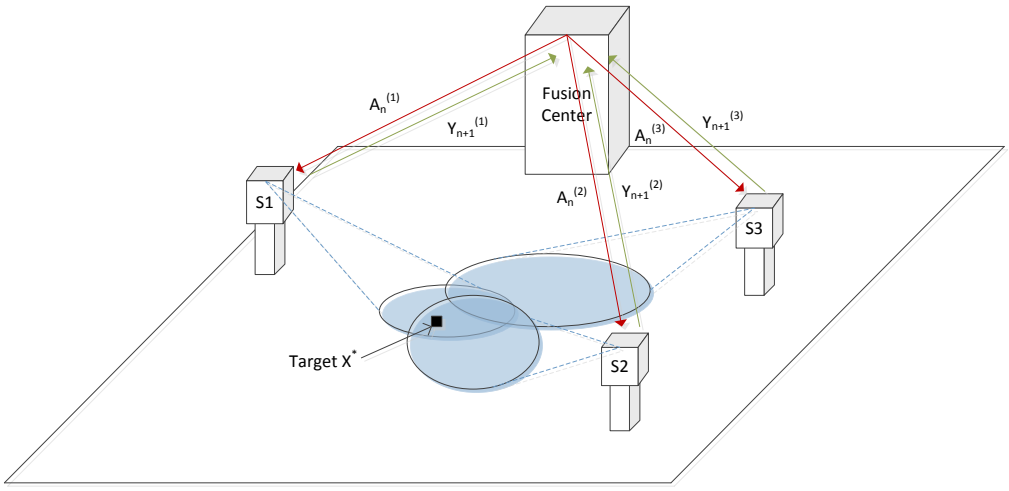


Figure 1.1: Illustration of basic centralized collaborative tracking system. At each time instant n, the controller (here, fusion center) designs queries $\{A_n^{(m)} : 1 \leq m \leq M\}$ and the sensors focus a beam in each region and provide a noisy response $Y_{n+1}^{(m)}$ after doing some covariance-based target detection. The responses $\{Y_{n+1}^{(m)} : 1 \leq m \leq M\}$ are transmitted back to the fusion center and the posterior distribution of the target is refined. This process is repeated until the target is localized to within an acceptable accuracy.

approach iteratively refines this separable representation through repeated querying and belief sharing. In the decentralized setting, at each time instant, each agent in the network first designs a query using a low-complexity controller (i.e., bisection method) as a function of its local current belief and yields a noisy response which is used to update its local belief. Second, the belief of each agent is updated as a convex combination of its refined belief (from the first step) and its neighbors' beliefs at the previous time instant (before they were updated). An illustration of such a decentralized active tracking system in shown in Figure 1.2.

In a parameter estimation setting, it is a common theme that exploiting the structure of the data distribution often yields superior estimation performance as compared to naive estimators. Often, even though the data may lie in a high dimensional space, most of the relevant information lies in a much lower dimensional space. The search for good low-dimensional representations of high dimensional data sets has

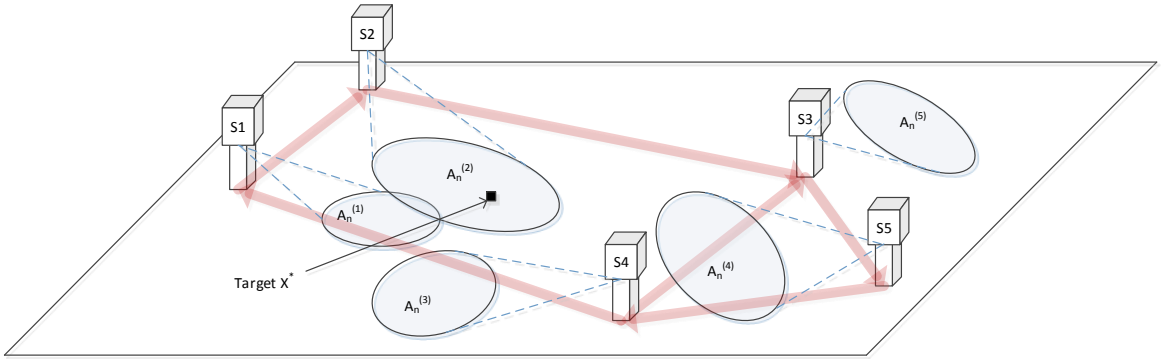


Figure 1.2: Illustration of basic decentralized collaborative tracking system. At each time instant n, each sensor in the network designs a query $A_n^{(m)}$ and focuses a beam in a region and provides a noisy response $Y_{n+1}^{(m)}$ after doing some covariance-based target detection. Each sensor uses its response $Y_{n+1}^{(m)}$ to refine the posterior distribution of the target and updates its belief using a convex combination of its refined belief and the beliefs of its neighbors at the previous time instant n (the incoming neighbors are shown as directed red arrows). This process is repeated until the target is localized to within an acceptable accuracy, and a consensus on the target location is reached across the network.

recently yielded breakthroughs in multivariate statistics and signal processing. This modern theme of studying high-dimensional objects having small intrinsic dimension, has sparked new results and methodologies in signal processing, an excellent example being compressed sensing, where s-sparse vectors of dimension d can be recovered with $n = \Omega(s \log(d/s))$ appropriately designed measurements [28, 6, 26, 47]. Similar results have appeared for the matrix completion problem, where a low-rank $d \times d$ matrix $\mathbf{C}$ can be recovered by nuclear norm minimization given only $n = \Omega(rd \log^2(d))$ observed entries, assuming $r = \text{rank}(\mathbf{C})$ and $\mathbf{C}$ satisfies an incoherence condition [24, 25, 27, 102].

Covariance estimation is a fundamental problem in multivariate statistics and finds application in many diverse areas, including economics, geostatistics and signal processing. It can be a very challenging problem when the number of samples n is fewer than the number of variables d, which is increasingly true in applications where

resources are limited. Sparsity is one of the most well-studied constraints imposed on the inverse covariance (i.e. precision) matrix. The graphical lasso (Glasso) estimator is a convex optimization approach proposed in [136, 5] for estimating a sparse inverse covariance, under an i.i.d. Gaussian observation model. The high dimensional convergence rate of Glasso was established by Ravikumar *et al* [99] and by Rothman *et al* [103] for a slight variant, showing that $n = \Omega((d+s)\log(d))$ samples are sufficient for accurate covariance estimation (wrt. Frobenius norm), where s is the number of nonzero off-diagonal entries in the underlying $d \times d$ precision matrix. Low-rank structure is another covariance constraint that comes up in factor analysis [50], random effect models [51] and spiked covariance models [74]. A convex optimization problem was proposed in [85] to derive a consistent estimate of the low-rank covariance in high dimensions. The high dimensional convergence rate for low-rank approximations was established in considerable generality in Lounici [85] and includes the case of missing observations. It was shown that $n = \Omega(rd\log(d))$ suffices for recovering the rank r covariance matrix $\mathbf{\Sigma}_0$ of size $d \times d$ (wrt. Frobenius norm).

A class of covariance models that finds applications in multimodal data are Kronecker product (KP) models. In their simplest form, these separable models assume that the covariance can be represented as the Kronecker product of two lower dimensional covariance matrices, i.e. $\mathbf{\Sigma}_0 = \mathbf{A}_0 \otimes \mathbf{B}_0$, with $p \times p$ p.d. matrix $\mathbf{A}_0$ and $q \times q$ p.d. matrix $\mathbf{B}_0$ [49, 130]. Chapter II considers the statistical estimation of covariance matrices constrained to obey the Kronecker product factorization, with possibly sparse structure in the precision matrices $\mathbf{X}_0 = \mathbf{A}_0^{-1}, \mathbf{Y}_0 = \mathbf{B}_0^{-1}$. It is shown that the separable structure of the covariance $\mathbf{\Sigma}_0$ effectively reduces the high dimensionality of the ambient space. For the case of sparse inverse of the factors, ℓ_1-penalized MLE estimators are proposed. In [111], a global ℓ_1 penalty on the pre-

cision matrix is imposed under Kronecker product structure and a block coordinate optimization method is shown to be empirically fast. In [115, 116], two additive ℓ_1 penalties are imposed on the MLE, each one for each Kronecker factor. To keep the algorithm computationally tractable, a block coordinate method was proposed to solve the underlying nonconvex (nonsmooth) optimization problem. Under mild conditions on the sample size, it was shown that this method (KGlasso) converges to a local minimum of the objective. For a fixed number of iterations, it was shown that the structured covariance estimate greatly outperforms previous state-of-the-art estimators in the high dimensional setting in terms of mean-square-error (MSE). The analysis yields considerably improved scaling laws for minimal sample size requirements for accurate estimation of these structured covariance matrices and aids the choice of regularization parameters.

In Chapter III, a more general class of KP covariance models is considered that finds applications in spatiotemporal signal processing where the covariance admits an additive decomposition of Kronecker products. This modeling approach allows any covariance matrix to be arbitrarily approximated by such a representation and as a result, it offers a dimensionality reduction when the covariance has a low dimensional representation on a Kronecker product basis. The number of components on this Kronecker product basis will be called the separation rank [112]. Product separability is imposed through the Kronecker product models and additive separability is obtained through the addition of Kronecker product forms. This decomposition has strong analogies to low rank matrix expansions. In [112], a convex optimization framework is proposed for obtaining asymptotically consistent estimators for this type of covariance structures. The objective is based on a model-free least-squares approach with nuclear norm penalization. The nuclear norm regularization implic-

itly projects the sample covariance matrix into a space of covariances admitting low separation rank. The computational complexity of the estimation algorithm, PRLS, remains scalable in terms of the separation rank (i.e., the model order). High dimensional MSE convergence rates are derived that generalize the convergence rates obtained for the single term unstructured Kronecker product. The results further show a fundamental tradeoff between approximation error (i.e., the error induced by model mismatch between the true covariance and the model) and estimation error (i.e., the error due to finite sample size). When the model order is exactly known, the estimation performance can be explicitly characterized in terms of the true separation rank of the underlying model. For models where the model order is approximately known or unknown, to obtain a desired estimation accuracy ϵ, the minimal separation rank needs to be calculated to arrive at a meaningful MSE convergence rate. More details on this are given in Section 1.2.

Returning to the original target localization problem, at each time instant n a set of M sensors need to make a decision about the presence of a target in a given set of regions $\mathbf{A}_n = \{A_n^{(m)} : 1 \leq m \leq M\}$. Given estimates of the covariance formed using N samples from a certain region $A_n^{(m)}$, a decision about the presence of a target in a region $A_n^{(m)}$ is made by the mth sensor. Due to finite sample noise in the covariance estimate, nonstationarities in the data caused by moving targets or other factors in the environment, the mth sensor will make an error with some probability $\epsilon_m \in (0, 1/2)$. In an active sensing problem (e.g. frequency agile radar), each sensor may choose to focus a beampattern on an region $A_n^{(m)}$ and obtain a small number N of samples that are locally stationary, which can be used to form an empirical estimate of the covariance.

Chapter IV considers the problem of optimal query design for a set of sensors;

a problem that arises in centralized active collaborative target search. In this context, each query would be a question of the form "Is the target in the region $\mathbf{A}_n$?". Through this lens, the search for jointly optimal policies for control of a collaborative set of sensors can be viewed as a collaborative 20 questions game. A 20 questions game aims to locate a target after asking a set of carefully designed questions. It can be formulated as the sequential design of questions such that entropy of the posterior distribution of the target's location after asking n questions is minimized [73]. Under the minimum expected entropy criterion and conditional independence between players, the optimality conditions for the joint policy (i.e., the policy that asks all players questions in parallel at each time instant) were derived in [119] and shown to generalize the probabilistic bisection policy of one player. For targets lying in a high dimensional space or for a large number of players, the design of such jointly optimal policies becomes intractable. A separable bisection policy is introduced for constructing questions and queries each player in sequence (after intermediately refining the posterior of the target location).

In [119], it is proven that this separable approach achieves the same expected information gain (or entropy loss) as the jointly optimal scheme. Upper and lower bounds on the MSE are also derived in [118]. This equivalence was also generalized to cover the case of unknown sensor reliabilities in [117, 118] under a joint Bayesian setting. This framework allows a mathematical model for incorporating a human in the loop in active machine learning systems. More details on the value of the human in the loop are included in Section 1.3.

Chapter V considers the problem of decentralized collaborative stochastic search. In this context, each sensor m in the network is faced with queries of the form "Is the target in the region $A_n^{(m)}$?", where the query region $A_n^{(m)}$ is obtained as a

function of the current local belief $p_n^{(m)}(\cdot)$. Once a (noisy) response is obtained to this query, first, the local belief (i.e., the posterior distribution) of each sensor m is updated using Bayes' rule and second, the local belief is further updated by linearly combining the Bayesian updated local belief with its neighbors' beliefs at the previous time instant. This scheme combines new information through repeated querying and shares information throughout the network by local belief sharing between neighbors.

In [120], under mild assumptions on the network structure (i.e., strong connectivity and strictly positive self-reliances), it is proven that this decentralized sequential estimation scheme yields a sequence of posterior distributions that globally converge almost surely to the true target location. Thus, asymptotically all agents in the network will reach a consensus on the space of beliefs and the limiting belief will be centered at the true target location. More details on this decentralized scheme and results are contained in Section 1.4.

The thesis is organized as follows. This chapter summarizes the main results in the thesis. More specifically, Section 1.2 displays the main results of the Kronecker product covariance models in high dimensions and Section 1.3 contains the main results of the centralized collaborative 20 questions model for target localization. Chapter 2 explores how Kronecker product structure and sparsity affect the mean-square-error (MSE). Using a greedy alternating minimization method, it is shown that significantly higher convergence rates can be obtained by exploiting both of these constraints. Chapter 3 generalizes these fast convergence rates to models with additive Kronecker product structure using a convex formulation. Chapter 4 studies the problem of optimal query selection for multiple collaborative observers that arises in the context of centralized active collaborative target localization. A basic equivalence is established between jointly optimal query policies and separable

query policies. Chapter 5 introduces and studies the decentralized collaborative target localization problem and convergence theory is presented that shows successful aggregation of information is guaranteed under mild assumptions on the network structure.

1.2 High Dimensional Covariance Estimation under Kronecker Product Structure

1.2.1 Introduction

Covariance estimation is a fundamental problem in several disciplines including signal processing, economics, and geostatistics. For the special case of jointly Gaussian zero mean observations, the sample covariance matrix is the minimal sufficient statistic and summarizes the necessary information for inference. Several applications that involve covariance matrices are filtering, prediction, detection, and inference on graphical models. Often, better estimates of the covariance lead to improved task performance.

One of the seminal papers on structured covariance matrix estimation is the work by Burg *et al* [19], in which the general constrained maximum-likelihood estimator (MLE) problem was studied in the multivariate Gaussian setting. In [19], general optimality conditions for structured MLE's were derived. Assuming $n \geq d$, a variational principle was derived that characterizes the solution of the constrained MLE problem. For certain low dimensional special cases, closed form expressions for the constrained MLE are obtained. A general iterative method for finding a solution to the necessary conditions is presented. Although the framework in [19] is fairly general, it does not give insight into mean-square-error (MSE) performance of the estimator and the role of inherent dimensionality is unclear. In addition, the algorithm proposed to solve for the constrained MLE boils down to solving a sequence of

linear systems of the form $\mathbf{Ax} = \mathbf{b}$, which can be of high dimension and the coefficients depending on a basis of the constraint space, so a basis for the constraint space needs to be obtained. Thus, the computational complexity of the proposed method in [19] largely depends on the constraint space and the complexity of obtaining the solution to the linear inverse problems.

With computational tractability and MSE performance of the estimator (in high dimensions) as key motivators, the high dimensional statistics and signal processing community has recently shifted focus on convex relaxations of nonconvex optimization problems that arise from constraining the MLE. Popular methods involve penalizing a data-fit term (i.e., log-likelihood or loss function) to enforce constraints on the structure of the estimator in some manner. Through proper regularization, consistent estimators for the covariance can be obtained for high dimensional settings.

A good example is sparse inverse covariance matrix estimation, where a covariance estimate with a sparse inverse is desired. In the multivariate Gaussian setting, this can be written as the nonconvex constrained optimization problem:

$$\min_{\mathbf{\Theta} \in S_{++}^d} \operatorname{tr}(\mathbf{\Theta}\hat{\mathbf{S}}_n) - \log\det(\mathbf{\Theta}) \text{ subject to } \|\operatorname{vec}(\mathbf{\Theta})\|_0 \leq C$$

where $C \geq d$ controls the number of nonzero entries in the estimate and $\hat{\mathbf{S}}_n$ is the sample covariance matrix (SCM). Since the ℓ_0 norm often leads to combinatorially difficult problems, the convex relaxation to ℓ_1 norm has been proposed and has been shown to yield great theoretical and practical results. The approach is known as the Graphical Lasso method [136, 5]:

$$\min_{\mathbf{\Theta} \in S_{++}^d} \operatorname{tr}(\mathbf{\Theta}\hat{\mathbf{S}}_n) - \log\det(\mathbf{\Theta}) + \lambda|\mathbf{\Theta}|_1 \tag{1.1}$$

Efficient ways to optimize (1.1) have been proposed in the literature [52, 66] and have worst case computational complexity $O(d^4)$. For reasonably sparse problems,

the block coordinate method in [52] is roughly $O(d^3)$.

Gaussian graphical models encode the conditional independence relationships between random variables [80]. For the special case of jointly Gaussian distribution on the observation $\mathbf{Z} \in \mathbb{R}^d$, zeros in the (i,j)th element of the precision matrix $\mathbf{\Theta}_0$ are equivalent to having variables i, j conditionally independent given the rest of the variables, corresponding to no edge joining the variables i and j in the underlying graphical model [80]. Thus, the zero pattern of the inverse covariance $\mathbf{\Theta}_0$ determines the sparsity of the Gaussian graphical model. There has been much work related to Gaussian graphical model estimation and model selection [99, 88, 5, 140, 79].

For variables with a natural ordering (i.e., in time series modeling), an estimator based on maximum likelihood with ℓ_p regularization on the generalized autoregressive parameters has been proposed that exploit the banded structure of the modified Cholesky decomposition of the precision matrix [67]. In addition, covariance banding, thresholding and tapering techniques have also been proposed for the high dimensional setting to exploit sparsity or banded forms on the covariance matrix [13, 14, 104].

For data sets of spatiotemporal or multimodal character, Kronecker product models for the covariance have been proven useful for obtaining a reduction in the number of model parameters and obtaining superior estimation and task performance than other naive estimators (e.g. SCM) [131, 16, 137, 54, 39, 116, 114].

The high dimensional MSE convergence rate of Glasso was originally derived by Rothman *et. al* [103]:

$$\|\hat{\mathbf{\Theta}}_{Glasso,n} - \mathbf{\Theta}_0\|_F^2 = O_P\left(\frac{(d + s_{\mathbf{\Theta}_0})\log(d)}{n}\right) \tag{1.2}$$

where

$$s_{\mathbf{\Theta}_0} = \text{card}\left(\{(i,j) : [\mathbf{\Theta}_0]_{i,j} \neq 0, i \neq j\}\right)$$

is twice the number of edges in the underlying graph. The rate in (1.2) offers an improvement over the sample covariance matrix (SCM) rate:

$$\|\hat{\mathbf{\Theta}}_n - \mathbf{\Theta}_0\|_F^2 = O_P\left(\frac{d^2}{n}\right) \tag{1.3}$$

It has been shown in [95, 97] that the SCM suffers in the high dimensional regime from large eigenvalue spread. This phenomenon makes the SCM singular for d larger than n. It was also shown that the estimation of eigenvectors of the SCM becomes impossible if the ratio n/d is below a critical threshold.

While much is known about the convergence of the SCM and the Graphical Lasso estimator, it is largely unknown what type of high dimensional convergence rates one can expect from the Kronecker product covariance estimators. This is the subject of Chapters 2 and 3; the inherent dimensionality of the Kronecker product structure plays a dominant role in the high dimensional MSE convergence rate. The next subsections summarize the main results that characterize the benefits of Kronecker product-based covariance models.

1.2.2 Single Kronecker Product Covariance Model

Chapter 3 considers covariance estimation under the assumption that the data are i.i.d. zero mean multivariate Gaussian with covariance:

$$\mathbf{\Sigma}_0 = \mathbf{A}_0 \otimes \mathbf{B}_0 \tag{1.4}$$

where $\mathbf{A}_0$ and $\mathbf{B}_0$ are $p \times p$ and $q \times q$ p.d. matrices, respectively. We let the precision matrices be $\mathbf{X}_0 = \mathbf{A}_0^{-1}$ and $\mathbf{Y}_0 = \mathbf{B}_0^{-1}$. The model (1.4) is relevant to channel modeling for MIMO wireless communications, where $\mathbf{A}_0$ is a transmit covariance matrix and $\mathbf{B}_0$ is a receive covariance matrix [131]. The model is also relevant to other transposable models arising in recommendation systems like NetFlix and in gene expression analysis [2].

Using the KP constraint (1.4) under the maximum likelihood objective function, we seek to solve the nonconvex optimization problem:

$$J(\mathbf{X}, \mathbf{Y}) = \mathrm{tr}((\mathbf{X} \otimes \mathbf{Y})\hat{\mathbf{S}}_n) - q \log \det(\mathbf{X}) - p \log \det(\mathbf{Y}) \tag{1.5}$$

where $\hat{\mathbf{S}}_n$ is the SCM. Since the problem (1.5) is biconvex, a block coordinate approach is adopted [49, 130] that yields closed-form updates for $\mathbf{X}$ and $\mathbf{Y}$, known as the Flip-Flop (FF) algorithm.

Further, assuming the precision matrices $\mathbf{X}_0$ and $\mathbf{Y}_0$ are sparse, a pair of ℓ_1 penalties is added to the smooth objective (1.5):

$$J_\lambda(\mathbf{X}, \mathbf{Y}) = J(\mathbf{X}, \mathbf{Y}) + \bar{\lambda}_X |\mathbf{X}|_1 + \bar{\lambda}_Y |\mathbf{Y}|_1. \tag{1.6}$$

where $\bar{\lambda}_X$ and $\bar{\lambda}_Y$ are nonnegative regularization parameters. Due to the biconvexity of (1.6), a block coordinate descent method decomposes into first computing the FF solution and then sparsifying the resulting precision matrix using the Glasso framework. The resulting alternating minimization algorithm is called the KGlasso algorithm [116, 114, 115].

It is proven that under mild conditions on the sample size n, the sequence of iterations convergences to a local minimum of the objective function (1.6). In addition, in Chapter 2 it is proven that for a fixed number of iterations, the MSE convergence rate for the FF algorithm is [116, 115]:

$$\|\boldsymbol{\Theta}_{FF,n} - \boldsymbol{\Theta}_0\|_F^2 = O_P\left(\frac{(p^2+q^2)\log\max(p,q,n)}{n}\right) \tag{1.7}$$

offering a dramatic improvement in MSE performance over the unstructured SCM rate (1.3). The same rate holds for the estimation of the covariance matrix. We note that the inherent dimensionality of the unstructured Kronecker product model is of the order $O(p^2+q^2))$ since there are at most $p(p+1)/2+q(q+1)/2$ unknown covariance

parameters that characterize the model. The MSE convergence rate in (1.7) implies that to get an accurate covariance estimate, the sample size needs to scale in terms of the inerent dimensionality of the KP model, i.e., $n = \Omega((p^2+q^2)\log\max(p,q,n))$.

For sparse precision matrices, i.e., $s_{\mathbf{X}_0} = O(p), s_{\mathbf{Y}_0} = O(q)$, the inherent dimensionality of the Kronecker product model is of the order $O(p+q)$. In Chapter 2 it is shown that KGLasso offers a better MSE convergence rate [116, 115] in the case of sparse precision matrices:

$$\|\mathbf{\Theta}_{KGlasso,n} - \mathbf{\Theta}_0\|_F^2 = O_P\left(\frac{(p+q)\log\max(p,q,n)}{n}\right) \tag{1.8}$$

Thus, for accurate covariance estimation under the sparse Kronecker product model, $n = \Omega(p+q)$ samples suffice. KGLasso outperforms the Glasso estimator, the Flip-Flop estimator and the SCM. The results are supported by several synthetic simulations.

1.2.3 Series of Kronecker Products Covariance Model

There are applications where the model (1.4) does not suffice to model the data-i.e., it is too rigid of a model. To this end, we consider a nontrivial extension of the single Kronecker product model (1.4) and represent the covariance as a series of Kronecker products of two lower dimensional factor matrices, where the number of terms in the summation may depend on the factor dimensions:

$$\mathbf{\Sigma}_0 \approx \sum_{\gamma=1}^{r} \mathbf{A}_{0,\gamma} \otimes \mathbf{B}_{0,\gamma} \tag{1.9}$$

where $\{\mathbf{A}_{0,\gamma}\}$ are $p \times p$ linearly independent matrices and $\{\mathbf{B}_{0,\gamma}\}$ are $q \times q$ linearly independent matrices. We note $1 \leq r \leq r_0 = \min(p^2, q^2)$ and refer to r as the *separation rank*. The subject of Chapter 3 is to obtain consistent covariance estimators for the model (1.9) and derive tight MSE rates in high dimensions. We note that

the coordinate descent techniques of Chapter 2 do not easily apply to the additive model (1.9) since the log-determinant of a summation of bilinear forms is a difficult term to deal with. Moreover, the issue of local minima is not well understood in high dimensions, unlike in the separable structure of (1.4) (see Chapter II for more details).

The model (1.9) has been applied to several applications. In spatiotemporal MEG/EEG covariance modeling [41, 40, 15, 75], the model (1.9) is used as a general model for the spatiotemporal covariance matrix of MEG residuals. Different terms in the sum describe different independent phenomena related to background activity, which can further be interpreted as generated by randomly distributed dipoles with a certain spatial and temporal distribution. The model (1.9) also find concrete applications in synthetic aperture radar (SAR) data [110, 105]. In [110], each term in the summation was used to recover a different scattering mechanism present in the signal. In that setting of polarimetric SAR imaging, the left Kronecker factors $\mathbf{A}_{0,\gamma}$ are polarimetric signatures and $\mathbf{B}_{0,\gamma}$ are interferometric coherences and backscattered powers of the corresponding scattering mechanism.

The model (1.9) is analogous to separable approximation of continuous functions [12]. It is evocative of a type of low rank principal component decomposition where the components are Kronecker products. Van Loan and Pitsianis [84] have derived a correspondence that shows low separation rank is equivalent to low rank in a permuted space defined by a reshaping operator $\mathcal{R}(\cdot)$. Using the singular value decomposition (SVD) as the main tool, Van Loan and Pitsianis [84] showed that any $pq \times pq$ matrix can be written as an orthogonal expansion of Kronecker products. Thus, it follows that *any* covariance matrix can be arbitrarily approximated by a bilinear decomposition of the form (1.9).

Recent work on high dimensional covariance estimation by Lounici [85] has shown that a simple convex optimization program can be used to give optimal MSE rates of convergence for low rank covariance matrices. Specifically, the singular value thresholding (SVT) problem was proposed:

$$\hat{\mathbf{\Sigma}}_n^{\lambda} \in \arg\min_{\mathbf{S} \in S_{++}^d} \|\hat{\mathbf{S}}_n - \mathbf{S}\|_F^2 + \lambda \mathrm{tr}(\mathbf{S}) \tag{1.10}$$

where $\lambda > 0$ is a regularization parameter. For $\lambda = C' \|\mathbf{\Sigma}_0\|_2 \sqrt{\frac{r(\mathbf{\Sigma}_0)\log(2d)}{n}}$, where $C' > 0$ is large enough, and $n \geq c r(\mathbf{\Sigma}_0) \log^2(\max(2d, n))$ for some constant $c > 0$ sufficiently large, Corollary 1 in [85] establishes a tight Frobenius norm error bound, which states that with probability $1 - \frac{1}{2d}$:

$$\|\hat{\mathbf{\Sigma}}_n^{\lambda} - \mathbf{\Sigma}_0\|_F^2 \leq \inf_{\mathbf{S} \succ 0} \left\{ \|\mathbf{\Sigma}_0 - \mathbf{S}\|_F^2 + C \|\mathbf{\Sigma}_0\|_2^2 \mathrm{rank}(\mathbf{S}) \frac{r(\mathbf{\Sigma}_0)\log(2d)}{n} \right\}$$

where $r(\mathbf{\Sigma}_0) = \frac{\mathrm{tr}(\mathbf{\Sigma}_0)}{\|\mathbf{\Sigma}_0\|_2} \leq \min\{\mathrm{rank}(\mathbf{\Sigma}_0), d\}$ is the effective rank [85].

Motivated by the correspondence between Kronecker product series decomposition and low rank series decompositions [84], and the high dimensional rates obtained by Lounici [85], we propose the permuted rank-penalized least-squares (PRLS) estimator:

$$\hat{\mathbf{\Sigma}}_n^{\lambda} \in \mathcal{R}^{-1} \left(\arg\min_{\mathbf{R}} \|\mathcal{R}(\hat{\mathbf{S}}_n) - \mathbf{R}\|_F^2 + \lambda \|\mathbf{R}\|_* \right) \tag{1.11}$$

where $\mathbf{R}(\cdot)$ is a permutation reshaping operator (see Chapter 3 for more details), $\hat{\mathbf{S}}_n$ is the SCM, and $\|\cdot\|_*$ is the nuclear norm. Since the nuclear norm of a matrix is the sum of absolute values of singular values, it enforces low rank structure in the permuted space; thus enforcing low separation rank in the original $pq \times pq$ domain. The solution of the nuclear norm penalized least squares problem

$$\min_{\mathbf{R}} \|\mathcal{R}(\hat{\mathbf{S}}_n) - \mathbf{R}\|_F^2 + \lambda \|\mathbf{R}\|_* \tag{1.12}$$

is given by:

$$\hat{\mathbf{R}}_n^{\lambda} = \sum_{j=1}^{r_0} \left(\sigma_j(\hat{\mathbf{R}}_n) - \frac{\lambda}{2} \right)_+ \mathbf{u}_j \mathbf{v}_j^T \tag{1.13}$$

Thus, the singular value spectrum is regularized through the nuclear norm penalty in a way to enforce low separation rank solutions. The permuted singular value thresholding problem (1.12) can be efficiently solved using fast optimization methods, without computing the full SVD [21, 22]. Although empirically observed to be fast, the computational complexity of the algorithms presented in [21] and [22] is unknown. Standard computation of the rank r SVD in the permuted space requires on the order $O(p^2q^2r)$ floating point operations. However, faster probabilistic-based methods for truncated SVD take $O(p^2q^2\log(r))$ computational time [61]. Thus, the computational complexity of solving (3.5) scales well with respect to the designed separation rank.

The convex optimization problem (1.12) is a convex relaxation of the rank constrained least-squares problem:

$$\min_{\mathbf{R}:\text{rank}(\mathbf{R})\leq r} \|\mathcal{R}(\hat{\mathbf{S}}_n) - \mathbf{R}\|_F^2 \tag{1.14}$$

The estimator arising from the solution of (1.14) will be called the covariance matching (CM) estimator. Working with the convex optimization problem (1.12) instead of (1.14) makes the MSE analysis more tractable and the solution can be efficiently computed using machinery from convex optimization [18], without computing the full SVD. Interestingly enough, even if the true model order r is known, the PRLS estimator outperforms the CM estimator in the small sample regime [113, 112].

In Chapter III, the positive definiteness of the PRLS estimator $\hat{\mathbf{\Sigma}}_n^{\lambda}$ is established under mild assumptions on the sample size n [113, 112]. In Chapter III, the high dimensional MSE convergence rate associated with the PRLS estimator is shown to

be [112]:

$$\|\hat{\mathbf{\Sigma}}_n^{\lambda} - \mathbf{\Sigma}_0\|_F^2 \leq \inf_{\mathbf{R}:\text{rank}(\mathbf{R})\leq r} \|\mathbf{R} - \mathbf{R}_0\|_F^2 + O_P\left(\frac{r(p^2+q^2+\log\max(p,q,n))}{n}\right) \tag{1.15}$$

The rate (1.15) shows a fundamental tradeoff between approximation error (i.e., the error induced by model mismatch between the true covariance and the model) and estimation error (i.e., the error due to finite sample size). For exactly separation rank r covariances, in the large p, q, n regime where $p^2 + q^2 + \log M = O(n)$, the convergence rate simplifies to:

$$\|\hat{\mathbf{\Sigma}}_n^{\lambda} - \mathbf{\Sigma}_0\|_F^2 = O_P\left(\frac{r(p^2+q^2+\log\max(p,q,n))}{n}\right)$$

In this scenario, the PRLS rate (1.15) reflects the inherent dimensionality of the model, which is of the order of $O(r(p^2+q^2))$. Finally, the rate generalizes the high dimensional rates obtained in Chapter II for the single Kronecker product model, i.e., for $r = 1$.

For covariance models characterized by singular value spectra that have no sharp cutoff at some $k = r$ point, the approximation error will be nonzero for any r. In some cases, the singular value spectrum of $\mathbf{R}_0 = \mathcal{R}(\mathbf{\Sigma}_0)$ may follow a power law decay. In that case, we can hope to estimate the covariance up to some bounded approximation error ϵ. To maintain this bounded approximation error as $p, q \to \infty$, we need to ensure the approximation error $\inf_{\mathbf{R}:\text{rank}(\mathbf{R})\leq r} \|\mathbf{R} - \mathbf{R}_0\|_F^2$ stays bounded above by $\epsilon > 0$ as p, q grow to infinity.

From least-squares approximation theory, we have the relation:

$$\inf_{\mathbf{R}:\text{rank}(\mathbf{R})\leq r} \|\mathbf{R} - \mathbf{R}_0\|_F^2 = \sum_{k=r+1}^{r_0} \sigma_k^2(\mathbf{R}_0)$$

where $r_0 = \min(p^2, q^2)$ grows to infinity, which can make the approximation error infinite. To ensure the sum remains finite as $p, q \to \infty$, the singular values of the permuted covariance $\mathbf{R}_0$ must decay to zero fast enough.

A nontrivial example where the approximation error can be explicitly controlled is the case of block-Toeplitz matrices. Such covariance matrices naturally arise as covariance matrices of multivariate stationary random processes $\mathbf{y}$ of dimension m and take a block-form:

$$\underbrace{\mathbf{\Sigma}_0}_{(N+1)m\times(N+1)m} = \begin{bmatrix} \mathbf{\Sigma}(0) & \mathbf{\Sigma}(1) & \dots & \mathbf{\Sigma}(N) \\ \mathbf{\Sigma}(-1) & \mathbf{\Sigma}(0) & \dots & \mathbf{\Sigma}(N-1) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{\Sigma}(-N) & \mathbf{\Sigma}(-N+1) & \dots & \mathbf{\Sigma}(0) \end{bmatrix} \tag{1.16}$$

where each submatrix is of size $m \times m$. For a zero-mean vector process $\mathbf{y} = \{\mathbf{y}(0), \dots, \mathbf{y}(N)\}$, the submatrices are given by $\mathbf{\Sigma}(\tau) = \mathbb{E}[\mathbf{y}(0)\mathbf{y}(\tau)^T]$. For concreteness, consider a block-Toeplitz p.d. matrix $\mathbf{\Sigma}_0$ of size $(N+1)m \times (N+1)m$ [1]. Under a mild assumption on the off-diagonal decay of the covariance, namely $\|\mathbf{\Sigma}(\tau)\|_F^2 \leq C' u^{2|\tau|} q$ for all $\tau = -N, \dots, N$ and constant $u \in (0, 1)$, the minimal separation rank can be obtained as a function of the desired approximation accuracy ϵ:

$$r \geq \frac{\log(pq/\epsilon)}{\log(1/u)}.$$

Then, the PRLS algorithm estimates $\mathbf{\Sigma}_0$ up to an absolute tolerance $\epsilon \in (0, 1)$ with convergence rate guarantee:

$$\|\hat{\mathbf{\Sigma}}_n^\lambda - \mathbf{\Sigma}_0\|_F^2 \leq \epsilon + C' r \frac{p^2 + q^2 + \log M}{n}$$

for appropriately scaled regularization parameter $\lambda > 0$, and absolute constant $C' > 0$.

[1] Here, the factor dimensions are $p = N + 1$ and $q = m$.

Several synthetic simulations are presented in Chapter III to show the benefits of PRLS over the standard SCM, the SVT estimator of Lounici [85] (which is tailored for low rank covariance matrices) and the covariance matching (CM) estimator (i.e, the solution of (1.14)). The PRLS estimator also displayed superior performance on a wind speed prediction task using real data collected by the NCEP [113, 112].

1.3 Centralized Collaborative Stochastic Search

1.3.1 Introduction

To locate a target quickly and efficiently, tools from stochastic search and optimal control are used. A series of sequentially designed questions about the location of the target are asked and noisy responses are obtained that are used to refine the estimate of the target location. The sequential aspect of the problem is key to speeding up the location estimation of targets, and this framework can be applied to various target localization applications. In our framework, asking a set of questions in the collaborative setting requires, the mth sensor to collect N data samples from a region $A_n^{(m)}$ and performing a detection task that yields a noisy response about the presence of a target within that region $A_n^{(m)}$.

The roots of the techniques for optimal query design lie in stochastic control [96, 11]. Applications of this methodology include active learning [108, 31, 29, 30] and sequential experimental design [42, 132]. The adaptation of decision increases efficiency at the expense of cost for finding the optimal decision policy. More specifically, for Bayesian formulations, it is known that the Bayes-optimal policy that arises is the solution to a partially observed Markov decision process (POMDP) which is described by a dynamic programming recursion. While sometimes it is possible to obtain explicit solutions to this recursion [56, 10], in many cases it is intractable. As a result, when the globally optimal policy is too difficult to compute, a one-step

lookahead heuristic is often used as a greedy approximation [138].

A key motivator for our work is the paper by Jedynak *et al* [73], where a Bayesian formulation is considered for sequential estimation of the target location. The problem was formulated in the context of a 20 questions game and it was shown that the greedy policy is Bayes-optimal under the minimum expected entropy criterion. In addition, under noisy responses with a symmetric noise model, bisecting the posterior yields globally optimal policies after a finite number of questions. This posterior bisection policy has been previously known as the probabilistic bisection policy, or Horstein's scheme, and has its roots in information theory [65], where sequential encoding of a message through a binary symmetric channel (BSC) was considered. The origins of 20 questions lie in information theory and binary search [35]. The binary search procedure was generalized in [92], where under some incoherence conditions, the generalized binary search (GBS) can learn a "correct" binary-valued function through a sequence of $O(\log N)$ queries in a space of N hypothesized functions. This method was also applied to the problem of learning halfspaces in machine learning.

Another related problem to the stochastic target search problem is stochastic root-finding. In this problem, the target is the zero of a decreasing function f, and the task is to locate the root of f given noisy observations of the function. The controller chooses the query points $x_1, x_2, \ldots$ and observes noisy versions of $f(x_1), f(x_2), \ldots$. The queries in this setting are questions of the form 'Is $f(x) < 0$?', and rates of convergence are well known. In [125], it was shown that under mild conditions on the noisy response models, a probabilistic bisection method converges to the root of f almost surely. In addition, for the constant error rate case, it was also shown that it converges exponentially fast; contrary to the best stochastic approximation rate of $n^{-1/2}$ [100, 78].

All the above mentioned works consider the single player case-i.e., there is one query to be designed at each time instant and one noisy response on the target's location is obtained. Next, let us consider the collaborative setting. In this setting, the joint controller sequentially selects a set of queries for the M players and uses the noisy responses to formulate the next set of questions. The questions are chosen such that an information criterion is maximized in order to extract as much information as possible about the target from the players. Chapter IV addresses this design problem for the criterion of minimizing expected entropy after asking n questions (i.e., finite horizon); with entropy quantifying the uncertainty in the target's location.

An application of this collaborative setting is human-in-the-loop active learning. For simplicity, a machine (player 1) and a human (player 2) may be available to collaborate in order to learn a target's location more efficiently. At each time instant, the queries have to be chosen such that the value of adding the human in the loop is maximized. The human-machine interaction can be modeled as a noisy collaborative 20 questions game and the design of queries can be addressed using the minimum expected entropy formulation (see Chapter IV).

1.3.2 Structure of Jointly Optimal Queries

To quantitatively describe the structure of optimal policies, the criterion of minimum expected entropy is adopted as in Jedynak *et al* [73]. In [73], a noisy oracle is repeatedly queried about the presence of a target X^* in a measurable set $A_n \subseteq \mathbb{R}^d$. At time n, the noisy response Y_{n+1} is a probabilistic function of the indicator function $Z_n = I(X^* \in A_n)$ [2]. Starting with a prior distribution $p_0(\cdot)$ on the target's location, the objective is to minimize the expected entropy of the distribution after asking N

[2]The channel is also assumed to be memoryless and time-invariant.

questions:

$$\inf_{\pi} \mathbb{E}^{\pi} \left[H(p_N) \right] \tag{1.17}$$

where $\pi = (\pi_0, \pi_1, \dots)$ denotes the policy and the entropy is the standard differential entropy [35]. The posterior distribution $p_N(\cdot)$ is the distribution of the target X^* given the history of the previous questions $\{A_n\}_{n=0}^{N-1}$ and responses $\{Y_{n+1}\}_{n=0}^{N-1}$. The median of the posterior distribution $p_N(\cdot)$ can be used to estimate the target location after N questions. It was shown that the bisection policy is optimal under the minimum entropy criterion. Assuming the noisy channel is a binary symmetric channel (BSC), all optimal policies are characterized by:

$$\mathbb{P}_n(A_n) := \int_{A_n} p_n(x) dx = 1/2 \tag{1.18}$$

In the collaborative setting, M collaborating players can be asked questions at each time instant n. The mth player's query at time n is of the form 'Does X^* lie in the region $A_n^{(m)} \subseteq \mathbb{R}^d$?'. These queries can be summarized as binary variables $Z_n^{(m)} = I(X^* \in A_n^{(m)})$ and the mth player yields a noisy response $Y_{n+1}^{(m)} \in \{0, 1\}$. for simplicity, let us define the M-tuples $\mathbf{Y}_{n+1} = (Y_{n+1}^{(1)}, \dots, Y_{n+1}^{(M)})$ and $\mathbf{A}_n = \{A_n^{(1)}, \dots, A_n^{(M)}\}$.

Under the assumptions of conditional independence of players and binary symmetric channels (BSC) for each player (with crossover probabilities $\epsilon_m \in (0, 1/2)$), the structure of the optimal policy can be fully characterized. Define the set of subsets of $\mathbb{R}^d$:

$$\gamma(A^{(1)}, \dots, A^{(M)}) = \left\{ \bigcap_{m=1}^{M} (A^{(m)})^{i_m} : i_m \in \{0, 1\} \right\}$$

where $(A)^0 := A^c$ and $(A)^1 := A$. The cardinality of this set of subsets is 2^M and these subsets partition $\mathbb{R}^d$. All optimal policies under this criterion must satisfy the

following set of equalities:

$$\mathbb{P}_n(R) = \int_R p_n(x)dx = 2^{-M}, \quad \forall R \in \gamma(\mathbf{A}_n) \tag{1.19}$$

As in Jedynak *et al* [73], the one-step lookahead policy is the Bayes-optimal policy for the multisensor setup also. At each time instant, the expected entropy decreases by the sum of the capacities of all BSC's-i.e., $C = \sum_{m=1}^{M} C(\epsilon_m)$ [119]. Surprisingly, despite the fact that all players are conditionally independent, the joint policy does not decouple into separate single player optimal policies. This is analogous to the non-separability of the optimal vector-quantizer in source coding even for independent sources [55].

According to the optimality condition (1.19), jointly optimal policies require overlapping, but non-identical queries. Thus, there is a nontrivial structure associated with the optimal set of questions. This structure becomes increasingly more intricate as the number of players M grows.

1.3.3 Equivalence Principle

Although the structure of the optimal policies is explicitly given by (1.19), the problem of designing optimal queries becomes intractable as the dimensions of the target d or number of players M become large.

As an alternative, a separable sequential coordinate-by-coordinate design is introduced: ask an optimal query to the first player, then update posterior density and ask an optimal query to the second player, and so on. The optimal query of each player is given by the probabilistic bisection policy [73]. This sequential bisection scheme has access to a more refined filtration (e.g., since the query and response of the first player are used to design the query of the second player), but requires more intermediate posterior updates. In Chapter IV, it is shown that this separable

scheme achieves the same expected entropy loss as the joint optimal design.

Thus, the complexity is transferred from the joint controller design to the posterior updates, since the posterior density needs to be re-updated after obtaining each player's response for the separable bisection scheme. The separable scheme effectively allows for an implementation of the joint scheme, without any performance loss on average.

1.3.4 Performance Bounds

The value of the 20 questions game is measured by the expected entropy reduction, which is the sum of capacities of all players. This quantity provides a fundamental limit of the MSE performance of the sequential Bayesian estimator [119]. Assuming $H(p_0)$ is finite, the MSE of the joint or sequential query policies satisfies:

$$\frac{K}{2\pi e} d e^{-2nC/d} \leq \mathbb{E}[\| X^* - X_n \|_2^2] \tag{1.20}$$

where $K = e^{2H(p_0)}$ and X_n is the posterior median [3]. The expected entropy loss per iteration is $C = \sum_m C(\epsilon_m)$.

The performance analysis of the bisection method is difficult primarily due to the continuous nature of the posterior [30]. A discretized version of the probabilistic bisection method was proposed in [20], using the Burnashev-Zingagirov (BZ) algorithm, which imposes a piecewise constant structure on the posterior. Using this framework, in Chapter IV, an upper bound on the MSE for the case of one-dimensional targets is obtained for the separable scheme:

$$\mathbb{E}[(X^* - \hat{X}_n)^2] \leq (2^{-2/3} + 2^{1/3}) \exp\left(-\frac{2}{3} n \bar{C}\right) \tag{1.21}$$

where $\bar{C} = \sum_{m=1}^{M} \bar{C}(\epsilon_m)$, $\bar{C}(\epsilon) = 1/2 - \sqrt{\epsilon(1-\epsilon)}$ is a measure of channel quality different from the capacity. The combination of the lower bound (1.20) and the upper

[3] For general dimensions $d \geq 1$, the posterior median X_n is defined as $P_n([0, X_{n,1}] \times \cdots \times [0, X_{n,d}]) = 1/2$.

bound (1.21) imply that the MSE converges to zero at an exponential rate with rate constant between $2C$ and $2/3\bar{C}$. Almost sure convergence is also shown for both the discretized-space and continuous-space (standard) versions of the separable scheme.

As an application, in Chapter IV, this methodology is applied to different noisy response models, e.g., human-like error models in which the human is more prone to making errors as the estimate X_n gets closer to the target X^*. It is shown that under the human error model, the value of including the human-in-the-loop for a sequential target localization task provides the largest gain in the beginning few question iterations and the additional contribution of the human decreases as the number of iterations grow to infinity. Synthetic simulations are also presented to validate the analysis.

1.3.5 Unknown Error Probabilities

We also extend the equivalence principle between the joint optimal query gain and the sequential bisection query gain to the case of unknown error probabilities. In this setting, the probabilistic bisection algorithm cannot be directly used since the Bayesian update is not well-defined. In the most generic setting of having unknown $\epsilon_m \in (0, 1/2)$, we propose a joint estimation scheme to estimate the target X^* and the error probabilities $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_M)$.

We consider the evolution of the joint posterior distribution of the joint random vector $(X^*, \boldsymbol{\epsilon})$ in time given the designed queries and noisy responses, because the error probabilities are coupled with the target through the Bayesian update. In this case, we prove in [117, 118] that the maximum entropy loss that can be achieved by the joint optimal design at time n is the expected sum of the capacities of the players conditioned on the information up to the current time instant n. This implies that the entropy loss is time-varying across iterations. The equivalence principle

states that the joint and sequential schemes are equivalent in the sense that the maximum entropy loss for each scheme is the sum of the capacities conditioned on the information available at each time instant (or at every set of sub-instants for the sequential scheme).

Since the maximum entropy loss is time-varying, we also consider a sensor selection scheme where at each time instant, the sensor with the highest information gain is selected. Finally, it is shown [117, 118] that even in the one-dimensional case with one sensor, for the setup of unknown probabilities, the optimal query policy is not equivalent to the (marginalized with respect to the noise) probabilistic bisection policy. Specifically, the optimal query policy can still be determined by a one-dimensional optimization (similar to the median search) as the solution x_n to:

$$\max_{x\in[0,1]} h_B(g_{1,n}(x)) \tag{1.22}$$

where $h_B(\cdot)$ is the binary entropy function [35] and $g_{1,n}(x)$ is a function dependent on the posterior distribution $p_n(x,\epsilon)$ given by

$$\begin{aligned} g_{1,n}(x) &= \int_0^x \mu_n(x')dx' + \int_x^1 (p_n(x') - \mu_n(x'))dx' \\ \mu_n(x') &= \int_{\epsilon=0}^{1/2} \epsilon p_n(x',\epsilon)d\epsilon \\ p_n(x') &= \int_{\epsilon=0}^{1/2} p_n(x',\epsilon)d\epsilon \end{aligned}$$

The solution to (1.22) is clearly not equivalent to the median of the marginalized distribution $p_n(x)$.

Simulations are also provided to illustrate the effectiveness of the methodology. It is empirically observed that the target estimation performance continues to be fast, robust to the simultaneous learning of the sensor error probabilities.

1.4 Decentralized Collaborative Stochastic Search

1.4.1 Introduction

Due to limited resources (e.g. power, bandwidth, hardware constraints), environmental factors (e.g. occlusions, geographic distance) and synchronization issues, it may be impractical to assume that all sensors can reliably transmit information to a fusion center at each time instant. In the decentralized setting, there is no fusion center available for centralized sequential Bayesian estimation of the target location as studied in [118].

To locate a target in a decentralized manner implies that sensors have to collaborate between each other in order to achieve the common objective of estimating the target location. In this setting, each sensor in the network has access to its own local belief (i.e., distribution) on the target location. Due to finite bandwidth, power, delay constraints and other environmental factors, sensors cannot communicate with all other sensors in the network, but may be able to communicate reliably with a few sensors close to them. This topology constraint can be described mathematically by a graph $G = (\mathcal{N}, E)$, where the vertex set $\mathcal{N} = \{1, \ldots, M\}$ indexes the sensors in the network and the edge set E contains all allowable directed links of information. Sharing the beliefs with its neighbors using a linear combination and repeating indefinitely may lead to a common limiting belief across all sensors in a network if convergence occurs. However, the limiting belief may not converge to the true target location. Thus, new information needs to be injected into the system in order to guide the beliefs to a unique belief centered at the true target location X^*.

The basis of collaboration for a common task can be traced back to the early work of Tsitsiklis [121] on distributed estimation and detection. Works on distributed averaging consensus have followed since in fields including the social sciences and

engineering. Althrough consensus is usually presented in the form of distributed averaging, consensus has broad applications including distributed optimization [121, 122], load-balancing [37] and distributed detection [106].

Consensus has appeared in the literature under different facets, including gossip algorithms and distributed averaging. Gossip algorithms are being recently studied primarily due to their robustness and flexibility; when link failures or packet losses occur randomly due to the unreliable and/or dynamic nature of wireless channels, the distributed information schemes may still converge, while centralized counterparts may fail or be impractical to implement. Even under specialized routing schemes to the fusion center, aggregating data towards a fusion center may require significant overhead (e.g. to maintain routes) causing communications bottlenecks and creates a single source of failure. Gossip algorithms require no specialized routing protocols; at each iteration, subsets of nodes exchange information and local updates are made at the receiving agents.

One of the first works that studied the convergence rate of these scheme in detail is the randomized gossip formulation presented by Boyd et al. [17], in which it was shown that the convergence rate of the averaging problem under randomized gossip (i.e., choose a pair of agents in the graph and do averaging) is controlled by the second largest eigenvalue of a doubly stochastic matrix defining the algorithm, making evident a natural relation between mixing times of random walks on the graph defined by a matrix of transition probabilities and averaging time of a gossip algorithm. One of the drawbacks of randomized gossip on random graphs was slow convergence. Further works, including geographic gossip [46], where nodes pair up with geographically distant nodes and exhange informaiton via multihop routing methods, and randomized path averaging [8], where routing nodes contributed their

own estimates along the way, requiring only a number of tranismission on ther order of the number of nodes in the network, offered faster convergence rates.

The survey paper by Dimakis et al. [45] reviews applications of gossip algorithms in sensor networks used for distributed estimation, source localization and compression. Recent work has also extended randomized gossip method to broadcast setting for consensus [3]. Futher, for the problem of gossip distributed estimation for linear parameter estimation, it was shown that under appropriate conditions on the network structure and observation models, the distributed estimator achieves the same performance as the best centralized linear estimator (in terms of asymptotic variance) [77]. Recently in [91], a new consensus scheme called hierarchical averaging was proposed to improve tradeoffs between resource consumption and quantization error for wireless links. Our work differs from this literature as the agents' observations are controlled through the query-response models, since the queries are functions of agents' local beliefs (and thus time-varying).

Another related large body of literature includes opinion dynamics over networks, spanning engineering to social sciences. The recent advances in sensor networks, social networks, etc. have sparked interest in convergence-related issues over networks with different assumptions on the observation models of the agents and network structure. There have been works proposed on learning over networks when agents follow simple updates to learn parameters including the works of DeGroot [43], Golub and Jackson [58] and Acemoglu et al. [1]. In these simple models, agents have an initial belief on the unknown state and agents aggregate this information in the presence of biases. Problems of agent coordination have been studied in detail in the distributed control literature [69, 89], where conditions for reaching consensus were studied for networks with changing topologies and time-dependent communication

links. In addition, problems of distributed estimation and detection have been studied in engineering by Tsitsiklis [121], where the problem of convergence to a common posterior (about a common unknown parameter) was studied under a belief exchnage scheme over a network with possible communication delays. In this early work, the models considered assumed that the network and observation structuers of all agents is common knowledge across all agents. This strong assumption will be abandoned in our work.

In our methodology, each agent acts greedily as an individual; he only knows his own error probability ϵ_m and the query is only a function of his own local belief. Although the estimation scheme is not fully Bayesian, this decentralized model probvides a tractable mathematical model for studying the evolution of dynamics of agents that repeately make new measurements based on designed queries in addition to observing beliefs of their neighbors.

1.4.2 Intractability of a fully Bayesian decentralized approach

The difficulty with the fully Bayesian approach stem from limited observability (i.e., observations of an agent are not observable by other agents) combined with the interactions of beliefs spread around the network. These two factors render the Bayesian approach impractical. In scenarios where agents have only partial information on the networks structure and the probability distribution of the signals observed by other agents (i.e., the observation densities of neighbors), the Bayesian approach becomes more complicated because agents would need to form and update beliefs on the states of the while, in addition to the networks struture and the rest of the agents' signal structures. This would increase the computational burden of the estimation scheme considerably, and given the assumptions that agents are naive (i.e., may act greedily and afford minimal computation), the scheme would simply

be impractical. Even if the network structure is known, agents would still need to update beliefs on the information of every other agent in the network, given only the neighbors' beliefs at each iteration. These complexities of a fully Bayesian scheme make it prohibitive, except for a few special cases [90] which do not apply to our decentralized estimation problem driven by local active queries and responses. Thus, Bayesian social learning has focused on simple networks [53, 34].

A key motivator for our semi-Bayesian (or non-Bayesian) approach is the recent work of Jadbabaie et al. [70]. In [70], it was shown that under a simple non-Bayesian scheme, all agents in the network asymptotically learn the true state of the world (i.e., the true parameter) even though agents and their neighbors may not have enough information to infer the true parameter by themselves. Contrary to this line of thought, in our problem, each agent in the network eventually has enough information to estimate the target X^* up to an arbitrary accuracy[4]. The surprising contribution is that consistency is maintained globally across all agents in the network under a non-Bayesian learning rule. In addition, in low signal-to-noise ratio settings, local belief sharing improves estimation performance.

1.4.3 Decentralized Estimation

Define the neighborhood of sensor m as $\mathcal{N}_m = \{m' \in \mathcal{N} : (m', m) \in E\}$. The weights $\{a_{i,j}\}$ for weighing the neighbor's beliefs at each iteration in the algorithm are summarized in a row-stochastic matrix A.

Starting with a collection of prior distributions $p_{i,0}(\cdot)$ on $\mathcal{X}$, the objective is to iteratively refine these distributions to reach global consensus towards the true target location across the network through repeated querying and information sharing. Motivated by the optimality of the bisection rule for symmetric channels proved in

[4]This follows from the strong consistency results recently derived for the one player setting in [126].

[73] and the simple non-Bayesian learning rule from [70], the decentralized estimation algorithm consists of two stages; at the beginning of each time step n, each agent m designs a query $A_{m,n}$ solely as a function of its local belief $p_{n,m}(\cdot)$ and yields a noisy response to this query, say $Y_{m,n+1}$. The semi-Bayesian scheme proposed consists of: (1) first, each agent forms the Bayesian posterior given his response $Y_{m,n+1}$ as an intermediate step, and (2) second, updates his belief to the convex combination of his Bayesian posterior and his neighbors' beliefs. This can be described mathematically as, for each agent $m \in \mathcal{N}$:

$$p_{m,n+1}(x) = a_{m,m} p_{m,n}(x) \frac{l_m(Y_{m,n+1}|x, A_{m,n})}{\mathcal{Z}_{m,n}(Y_{m,n+1})} + \sum_{m' \in \mathcal{N}_m} a_{m,m'} p_{m',n}(x), \quad x \in \mathcal{X} \tag{1.23}$$

where $l_m(Y_{m,n+1}|x, A_{n,m})$ is the observation likelihood which is a function of the controls (i.e., query region) $A_{n,m}$. The denominator $\mathcal{Z}_{m,n}(\cdot)$ is a normalizing factor to ensure that $\tilde{p}_{m,n}(x) := p_{m,n}(x) \frac{l_m(Y_{m,n+1}|x, A_{n,m})}{\mathcal{Z}_{m,n}(Y_{m,n+1})}$ is a valid density over $\mathcal{X}$.

1.4.4 Asymptotic Convergence of Beliefs

In [120], it is proven that in strongly connected networks (i.e., the matrices A and P are irreducible), the deterministic and randomized decentralized estimation schemes motivated by a semi-Bayesian approach enables successful aggregation of dispersed information. Although agents act greedily and the scheme is not fully Bayesian, the controlled dynamical system (1.23) is shown to converge to a common limiting belief centered at the true target location.

While similar results on non-Bayesian asymptotic learning have been derived in [70], our work differs significantly because (1) we consider a continuous-valued target space, (2) we make no assumptions on the collective identifiability of the true state of the world (i.e., the true target location), and (3) we consider observations that are based on active queries and responses of each agent. We remark that in

[70], the strong connectivity assumption on the network was made in order to prove convergence of the densities towards the true state of the world. In the absense of this assumption, agents in each strongly connected component of the network will asymptotically reach consensus to the true belief. Thus, this is a mild assumption.

The first principal result in [120] shows that consensus is achieved over intervals of the domain $\mathcal{X} = [0, 1]$. For any pair of agents (i, j) in the network, it is shown that for any $B = [0, b]$, $0 \leq b \leq 1$:

$$\max_i \mathbb{P}_{i,t}(B) - \min_i \mathbb{P}_{i,t}(B) \xrightarrow{p.} 0$$

as $t \to \infty$.

The second principal result in [120] is a consistency result for the target estimates and characterizes the structure of the limiting belief. Specifically, it is shown that for any interval $B = [0, b]$:

$$\mathbb{P}_{i,t}([0, b]) \xrightarrow{p.} \begin{cases} 0 & , b < X^* \\ 1 & , b > X^* \end{cases}$$

Thus, asymptotically as $t \to \infty$ all the mass is concentrated on the point $x = X^*$ (i.e. a Dirac measure). This further implies consistency of the estimates to X^*.

Simulations are also provided to show that the proposed methodology is valid under different graph topologies.

1.5 List of relevant publications

Journal Publications

1. T. Tsiligkaridis, B. M. Sadler and A. O. Hero III, "On Decentralized Estimation with Active Queries", in preparation.

2. T. Tsiligkaridis, B. M. Sadler and A. O. Hero III, "Collaborative 20 Questions for Target Localization", submitted to *IEEE Transactions on Information Theory*, in revision.

   Extended technical report available online at *arXiv: 1306.1922*

3. T. Tsiligkaridis and A. O. Hero III, "Covariance Estimation in High Dimensions via Kronecker Product Expansions", *IEEE Transactions on Signal Processing*, Vol. 61, No. 21, pp. 5347-5360, November 2013. doi: 10.119/TSP.2013.2279355

   Extended technical report available online at *arXiv: 1302.2686*

4. T. Tsiligkaridis, A. O. Hero III and S. Zhou, "On Convergence of Kronecker Graphical Lasso Algorithms", *IEEE Transactions on Signal Processing*, Vol. 61, No. 7, pp. 1743-1755, April 2013. doi: 10.1109/TSP.2013.2240157

   Extended technical report available online at *arXiv: 1204.0585*

Conference Publications

1. T. Tsiligkaridis, B. M. Sadler and A. O. Hero III, "Blind Collaborative 20 Questions for Target Localization", in Proceedings of *2013 IEEE GlobalSIP - Symposium on Controlled Sensing for Inference: Applications, Theory and Algorithms (GlobalSIP), Austin, TX, December 2013*

2. T. Tsiligkaridis and A. O. Hero III, "Low Separation Rank Covariance Estimation using Kronecker Product Expansions", in Proceedings of *2013 IEEE International Symposium on Information Theory (ISIT)*, pp. 1202-1206, Istanbul, Turkey, July 2013

3. T. Tsiligkaridis, B. M. Sadler and A. O. Hero III, "A Collaborative 20 Questions Model for Target Search with Human-Machine Interaction", in Proceedings of

2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 6516-6520, Vancouver, CA, March 2013

4. T. Tsiligkaridis, A. O. Hero III and S. Zhou, "Kronecker Graphical Lasso", in Proceedings of *2012 IEEE Statistical Signal Processing (SSP) Workshop*, pp. 884-887, Ann Arbor, MI, August 2012

5. T. Tsiligkaridis and A. O. Hero III, "Sparse covariance estimation under Kronecker product structure", in Proceedings of *2012 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 3633-3636, Kyoto, Japan, March 2012

CHAPTER II

Kronecker Graphical Lasso

This chapter studies iteration convergence of Kronecker graphical lasso (KGLasso) algorithms for estimating the covariance of an i.i.d. Gaussian random sample under a sparse Kronecker-product covariance model and MSE convergence rates. The KGlasso model, originally called the transposable regularized covariance model by Allen *et al* [2], implements a pair of ℓ_1 penalties on each Kronecker factor to enforce sparsity in the covariance estimator. The KGlasso algorithm generalizes Glasso, introduced by Yuan and Lin [136] and Banerjee *et al* [5], to estimate covariances having Kronecker product form. It also generalizes the unpenalized ML flip-flop (FF) algorithm of Dutilleul [49] and Werner *et al* [130] to estimation of sparse Kronecker factors. We establish that the KGlasso iterates converge pointwise to a local maximum of the penalized likelihood function. We derive high dimensional rates of convergence to the true covariance as both the number of samples and the number of variables go to infinity. Our results establish that KGlasso has significantly faster asymptotic convergence than Glasso and FF. Simulations are presented that validate the results of our analysis. For example, for a sparse $10,000 \times 10,000$ covariance matrix equal to the Kronecker product of two 100×100 matrices, the root mean squared error of the inverse covariance estimate using FF is 2 times larger than

that obtainable using KGlasso for sample size of $n = 100$.

2.1 Introduction

Covariance estimation is a problem of great interest in many different disciplines, including machine learning, signal processing, economics and bioinformatics. In many applications the number of variables is very large, e.g., in the tens or hundreds of thousands, leading to a number of covariance parameters that greatly exceeds the number of observations. To address this problem constraints are frequently imposed on the covariance to reduce the number of parameters in the model. For example, the Glasso model of Yuan and Lin [136] and Banerjee *et al* [5] imposes sparsity constraints on the covariance. The Kronecker product model of Dutilleul [49] and Werner *et al* [130] assumes that the covariance can be represented as the Kronecker product of two lower dimensional covariance matrices. The transposable regularized covariance model of Allen *et al* [2] imposes a combination of sparsity and Kronecker product form on the covariance. When there is no missing data, an extension of the alternating optimization algorithm of [130], that the authors call the flip flop (FF) algorithm, can be applied to estimate the parameters of this combined sparse and Kronecker product model. In this chapter, we call this extension the Kronecker Glasso (KGlasso) and we analyze pointwise convergence (Theorem II.2) and MSE convergence (Lemma II.7 and Thm. II.13) analyze convergence of the algorithm in the high dimensional ($d \gg n$) setting.

We adopt the notation of [130] and assume that there are pf variables whose covariance $\mathbf{\Sigma}_0$ has the separable positive definite Kronecker product representation:

$$\mathbf{\Sigma}_0 = \mathbf{A}_0 \otimes \mathbf{B}_0 \tag{2.1}$$

where $\mathbf{A}_0$ is a $p \times p$ positive definite matrix and $\mathbf{B}_0$ is an $f \times f$ positive definite

matrix. When the variables are multivariate Gaussian with covariance following the Kronecker product model (2.1) the variables are said to follow a matrix normal distribution [38, 49, 60]. As shown by [131] the Kronecker product model is relevant to channel modeling for MIMO wireless communications, where $\mathbf{A}_0$ is a transmit covariance matrix and $\mathbf{B}_0$ is a receive covariance matrix. The model has been applied to many other domains including: geostatistics [36], genomics [134], multi-task learning [16], face recognition [137], recommendation systems [2] and collaborative filtering [135]. The Kronecker product model (2.1) can easily be generalized to the k-fold case, where $\mathbf{\Sigma}_0 = \mathbf{A}_1 \otimes \mathbf{A}_2 \otimes \cdots \otimes \mathbf{A}_k$.

When there are n i.i.d. measurements from a matrix normal distribution with covariance factorization (2.1), the maximum likelihood (ML) estimator of $\mathbf{\Sigma}_0$ can be formulated [86]. While the ML estimator has no known closed-form solution, an approximation to the solution can be iteratively computed via an alternating algorithm [49, 86] called the flip-flop (FF) algorithm in [130]. As compared to the standard saturated (unstructured) covariance model, the number of unknown parameters in (2.1) is reduced from order $\Theta(p^2 f^2)$ to order $\Theta(p^2) + \Theta(f^2)$. This results in a significant reduction in estimator mean squared error (MSE) and in the computational complexity of the maximum likelihood (ML) covariance estimator. This chapter establishes that further reductions MSE are achievable when the inverse of the covariance (2.1) is known to be sparse, i.e., the measurements obey a Kronecker structured Gaussian graphical model.

The graphical lasso (Glasso) estimator was originally proposed in [136, 5] for estimating a sparse inverse covariance, also called the precision matrix, under an i.i.d. Gaussian observation model. An algorithm for efficiently solving the nonsmooth optimization problem that arises in the Glasso estimator, based on ideas from [5], was

proposed in [52]. Glasso has been applied to the time-varying coefficients setting in Zhou *et al* [139] using the kernel estimator for covariances at a target time. Rothman *et al* [103] derived high dimensional convergence rates for a slight variant of Glasso, i.e., only the off-diagonal entries of the estimated precision matrix were penalized using an ℓ_1-penalty. The high dimensional convergence rate of Glasso was established by Ravikumar *et al* [99]. This chapter extends their analysis to the case that the covariance has the Kronecker structure of (2.1) and shows that significantly higher rates of convergence are achievable.

The main contribution of this chapter is the derivation of tight high-dimensional MSE convergence rates for KGlasso as n, p and f go to infinity. When both Kronecker factors are sparse, it is shown that KGlasso *strictly* outperforms FF and Glasso in terms of MSE convergence rate. More specifically, we show KGlasso achieves a convergence rate of $O_P\left(\frac{(p+f)\log\max(p,f,n)}{n}\right)$ and FF achieves a rate of $O_P\left(\frac{(p^2+f^2)\log\max(p,f,n)}{n}\right)$ as $n\to\infty$, while it is known [103, 139] that Glasso achieves a rate of $O_P\left(\frac{(pf+s)\log\max(p,f,n)}{n}\right)$, where s denotes the number of off-diagonal nonzero elements in the true precision matrix $\mathbf{\Theta}_0$. Simulations show that the performance improvements predicted by the high-dimensional analysis continue to hold for small sample size and moderate matrix dimension. For the example studied in Sec. 2.9 the empirical MSE of KGlasso is significantly lower than that of Glasso and FF for $p=f=100$ over the range of n from 10 to 100.

The starting point for the MSE convergence analysis is the large-sample analysis of the FF algorithm (Thm. 1 in [130]). The KGlasso convergence proof uses a novel large deviation inequality (Lemma II.7) that shows that the dimension of one estimated Kronecker factor, say $\mathbf{A}$, acts as a multiplier on the number of independent samples when performing inference on the other factor $\mathbf{B}$. This result is then used

to obtain tight MSE rates in terms of Frobenius norm error between the KGlasso estimated matrix and the ground truth. The asymptotic MSE convergence analysis is useful since it can be used to guide the selection of sparsity regularization parameters and to determine minimum sample size requirements.

Independently, in the related work of Yin and Li [134], published after submission of our paper [116] for publication, high-dimensional MSE bounds for the same matrix normal estimation problem were considered. However, our MSE bounds are tighter than the bounds given in Yin and Li. In particular, neglecting terms of order $\log(pf)$, our bounds are of order $p+f$ as compared to Yin and Li's bounds of order pf, which is significantly weaker for large p, f. We obtain improved bounds due to the use of a tighter concentration inequality, established in Lemma II.7. While our paper [116] was being reviewed similar results to Thm. II.13 were published, but using a different method of proof, by Leng and Tang [83].

2.2 Notation

For a square matrix $\mathbf{M}$, define $|\mathbf{M}|_1 = \|\mathrm{vec}(\mathbf{M})\|_1$ and $|\mathbf{M}|_\infty = \|\mathrm{vec}(\mathbf{M})\|_\infty$, where $\mathrm{vec}(\mathbf{M})$ denotes the vectorized form of $\mathbf{M}$ (concatenation of columns into a vector). $\|\mathbf{M}\|_2$ is the spectral norm of $\mathbf{M}$. $\mathbf{M}_{i,j}$ and $[\mathbf{M}]_{i,j}$ are the (i,j)th element of $\mathbf{M}$. Let the inverse transformation (from a vector to a matrix) be defined as: $\mathrm{vec}^{-1}(\mathbf{x}) = \mathbf{X}$, where $\mathbf{x} = \mathrm{vec}(\mathbf{X})$. Define the $pf \times pf$ permutation operator $\mathbf{K}_{p,f}$ such that $\mathbf{K}_{p,f}\mathrm{vec}(\mathbf{N}) = \mathrm{vec}(\mathbf{N}^T)$ for any $p \times f$ matrix $\mathbf{N}$. For a symmetric matrix $\mathbf{M}$, $\lambda(\mathbf{M})$ will denote the vector of real eigenvalues of $\mathbf{M}$ and define $\lambda_{max}(\mathbf{M}) = \|\mathbf{M}\|_2 = \max \lambda_i(\mathbf{M})$ for p.d. symmetric matrix, and $\lambda_{min}(\mathbf{M}) = \min \lambda_i(\mathbf{M})$. Define the sparsity parameter associated with $\mathbf{M}$ as $s_M = \mathrm{card}(\{(i_1, i_2) : [\mathbf{M}]_{i_1,i_2} \neq 0, i_1 \neq i_2\})$. Let $\kappa(\mathbf{M}) := \frac{\lambda_{max}(\mathbf{M})}{\lambda_{min}(\mathbf{M})}$ denote the condition number of a symmetric matrix $\mathbf{M}$.

For a matrix $\mathbf{M}$ of size $pf \times pf$, let $\{\mathbf{M}(i,j)\}_{i,j=1}^{p}$ denote its $f \times f$ block submatrices, where each block submatrix is $\mathbf{M}(i,j) = [\mathbf{M}]_{(i-1)f+1:if,(j-1)f+1:jf}$. Also let $\{\overline{\mathbf{M}}(k,l)\}_{k,l=1}^{f}$ denote the $p \times p$ block submatrices of the permuted matrix $\overline{\mathbf{M}} = \mathbf{K}_{p,f}^{T}\mathbf{M}\mathbf{K}_{p,f}$.

Define the set of symmetric matrices $S^p = \{\mathbf{A} \in \mathbb{R}^{p\times p} : \mathbf{A} = \mathbf{A}^T\}$, the set of symmetric positive semidefinite (psd) matrices S_+^p, and the set of symmetric positive definite (pd) matrices S_{++}^p. $\mathbf{I}_d$ is a $d \times d$ identity matrix. It can be shown that S_{++}^p is a convex set, but is not closed [18]. Note that S_{++}^p is simply the interior of the closed convex cone S_+^p.

Statistical convergence rates will be denoted by the $O_P(\cdot)$ notation, which is defined as follows. Consider a sequence of real random variables $\{X_n\}_{n\in\mathbb{N}}$ defined on a probability space $(\Omega, \mathcal{F}, P)$ and a deterministic (positive) sequence of reals $\{b_n\}_{n\in\mathbb{N}}$. By $X_n = O_P(1)$ is meant: $\sup_{n\in\mathbb{N}} \Pr(|X_n| > K) \to 0$ as $K \to \infty$, where X_n is a sequence indexed by n, for fixed p, f. The notation $X_n = O_P(b_n)$ is equivalent to $\frac{X_n}{b_n} = O_P(1)$. By $X_n = o_p(1)$ is meant $\Pr(|X_n| > \epsilon) \to 0$ as $n \to \infty$ for any $\epsilon > 0$. By $\lambda_n \asymp b_n$ is meant $c_1 \leq \frac{\lambda_n}{b_n} \leq c_2$ for all n, where $c_1, c_2 > 0$ are absolute constants. The asymptotic notation $a_n = O(b_n)$ means $\limsup_{n\to\infty} |\frac{a_n}{b_n}| \leq C$ for some constant $C > 0$, while $c_n = \Omega(d_n)$ means $\liminf_{n\to\infty} |\frac{c_n}{d_n}| \geq C'$ for some constant $C' > 0$.

2.3 Graphical Lasso Framework

For simplicity, we assume the number of Kronecker components is $k = 2$. Available are n i.i.d. multivariate Gaussian observations $\{\mathbf{z}_t\}_{t=1}^{n}$, where $\mathbf{z}_t \in \mathbb{R}^{pf}$, having zero-mean and covariance equal to $\mathbf{\Sigma} = \mathbf{A}_0 \otimes \mathbf{B}_0$. Then, ignoring irrelevant constants, the log-likelihood $l(\mathbf{\Sigma})$ is:

$$l(\mathbf{\Sigma}) = \log\det(\mathbf{\Sigma}^{-1}) - \mathrm{tr}(\mathbf{\Sigma}^{-1}\hat{\mathbf{S}}_n), \tag{2.2}$$

where $\mathbf{\Sigma}$ is the positive definite covariance matrix and $\hat{\mathbf{S}}_n = \frac{1}{n}\sum_{t=1}^{n} \mathbf{z}_t \mathbf{z}_t^T$ is the sample covariance matrix. Recent work [5, 52, 99] has considered ℓ_1-penalized maximum likelihood estimators for the saturated model where $\mathbf{\Sigma}$ belongs to the unrestricted cone of positive definite matrices. These estimators are known as graphical lasso (Glasso) estimators and are obtained as the solution to the ℓ_1-penalized minimization problem:

$$\hat{\mathbf{\Sigma}}_n \in \arg\min_{\mathbf{\Sigma} \in S_{++}^p} \{-l(\mathbf{\Sigma}) + \lambda |\mathbf{\Sigma}^{-1}|_1\}, \tag{2.3}$$

where $\lambda \geq 0$ is a regularization parameter. If $\lambda > 0$ and $\hat{\mathbf{S}}_n$ is positive definite, then $\hat{\mathbf{\Sigma}}_n$ in (2.3) is the unique minimizer.

A fast iterative algorithm, based on a block coordinate descent approach, exhibiting a computational complexity $\mathcal{O}((pf)^4)$, was developed in [52] to solve the convex program (2.3). A fast algorithm, based on an active-set second-order method, with the same computational complexity was developed in [66] to solve the convex program. The Glasso mapping (2.3) is written as $\mathbf{G}(\cdot, \lambda) : S^d \to S^d$,

$$\mathbf{G}(\mathbf{T}, \lambda) = \arg\min_{\mathbf{\Theta} \in S_{++}^d} \Big\{ \mathrm{tr}(\mathbf{\Theta}\mathbf{T}) - \log\det(\mathbf{\Theta}) + \lambda |\mathbf{\Theta}|_1 \Big\}. \tag{2.4}$$

Under the assumption $\lambda \asymp \sqrt{\frac{\log(pf)}{n}}$ solution of (2.3) was shown to have high dimensional convergence rate [103, 139]:

$$\|\mathbf{G}(\hat{\mathbf{S}}_n, \lambda) - \mathbf{\Theta}_0\|_F = O_P\left(\sqrt{\frac{(pf + s)\log(pf)}{n}}\right) \tag{2.5}$$

where s is an upper bound on the number of non-zero off-diagonal elements of $\mathbf{\Theta}_0$. When $s = O(pf)$, this rate is better than that achieved in the case of the standard sample covariance estimator ($\lambda = 0$):

$$\|\hat{\mathbf{S}}_n - \mathbf{\Sigma}_0\|_F = O_P\left(\sqrt{\frac{p^2 f^2}{n}}\right). \tag{2.6}$$

2.4 Kronecker Graphical Lasso

Let $\mathbf{\Sigma}_0 = \mathbf{A}_0 \otimes \mathbf{B}_0$ denote the true covariance matrix, where $\mathbf{A}_0 = \mathbf{X}_0^{-1}$ and $\mathbf{B}_0 = \mathbf{Y}_0^{-1}$ are the true Kronecker factors. Let $\mathbf{A}_{init}$ denote an initial guess of $\mathbf{A}_0 = \mathbf{X}_0^{-1}$.

Define $J(\mathbf{X}, \mathbf{Y})$ as the negative log-likelihood

$$J(\mathbf{X}, \mathbf{Y}) = \operatorname{tr}((\mathbf{X} \otimes \mathbf{Y})\hat{\mathbf{S}}_n) - f \log\det(\mathbf{X}) - p \log\det(\mathbf{Y}) \tag{2.7}$$

Although the objective function (2.7) is not jointly convex in $(\mathbf{X}, \mathbf{Y})$, it is biconvex. This motivates the flip-flop algorithm [49, 130]. Adapting the notation from [130], define the mappings $\hat{\mathbf{A}}(\cdot), \hat{\mathbf{B}}(\cdot)$:

$$\underbrace{\hat{\mathbf{A}}(\mathbf{B})}_{p\times p} = \frac{1}{f}\sum_{k,l=1}^{f} [\mathbf{B}^{-1}]_{k,l}\overline{\hat{\mathbf{S}}_n}(l,k), \tag{2.8}$$

$$\underbrace{\hat{\mathbf{B}}(\mathbf{A})}_{f\times f} = \frac{1}{p}\sum_{i,j=1}^{p} [\mathbf{A}^{-1}]_{i,j}\hat{\mathbf{S}}_n(j,i), \tag{2.9}$$

where $\overline{\hat{\mathbf{S}}}_n = \mathbf{K}_{p,f}^T \hat{\mathbf{S}}_n \mathbf{K}_{p,f}$ (see Sec. 3.2 for definition of $\mathbf{K}_{p,f}$). For fixed $\mathbf{B} \in S_{++}^f$, $\hat{\mathbf{A}}(\mathbf{B})$ in (2.8) is the minimizer of $J(\mathbf{A}^{-1}, \mathbf{B}^{-1})$ over $\mathbf{A} \in S_{++}^p$. A similar interpretation holds for (2.9). The flip-flop algorithm starts with some arbitrary p.d. matrix $\mathbf{A}_{init}$ and computes $\mathbf{B}$ using (2.9), then $\mathbf{A}$ using (2.8), and repeats until convergence. This algorithm does not account for sparsity.

If $\mathbf{\Theta}_0 = \mathbf{X}_0 \otimes \mathbf{Y}_0$ is a sparse matrix, which implies that at least one of $\mathbf{X}_0$ or $\mathbf{Y}_0$ is sparse, one can penalize the outputs of the flip-flop algorithm and minimize

$$J_\lambda(\mathbf{X}, \mathbf{Y}) = J(\mathbf{X}, \mathbf{Y}) + \bar{\lambda}_X|\mathbf{X}|_1 + \bar{\lambda}_Y|\mathbf{Y}|_1. \tag{2.10}$$

where $\lambda_X = \bar{\lambda}_X/f$ and $\lambda_Y = \bar{\lambda}_Y/p$. This leads to an algorithm that we call KGlasso (see Algorithm 1), which sparsifies the Kronecker factors in proportion to the parameters $\bar{\lambda}_X, \bar{\lambda}_Y > 0$. This is the same objective function that was proposed in [2]

when specialized to the case that there is no missing data. A similar algorithm was presented in [134], where only the off-diagonal elements of the precision matrices were penalized.

Algorithm 1 Kronecker Graphical Lasso (KGlasso)

Input: $\hat{\mathbf{S}}_n$, p, f, n, $\bar{\lambda}_X > 0$, $\bar{\lambda}_Y > 0$
Output: $\hat{\mathbf{\Theta}}_{KGlasso}$
Initialize $\mathbf{A}_{init}$ to be positive definite satisfying Assumption II.10.
$\check{\mathbf{X}} \leftarrow \mathbf{A}_{init}^{-1}$
repeat
  $\hat{\mathbf{B}} \leftarrow \frac{1}{p}\sum_{i,j=1}^{p}[\check{\mathbf{X}}]_{i,j}\hat{\mathbf{S}}_n(j,i)$ (see Eq. (2.8))
  $\check{\mathbf{Y}} \leftarrow \mathbf{G}(\hat{\mathbf{B}}, \frac{\bar{\lambda}_Y}{p})$, where $\mathbf{G}(\cdot,\cdot)$ is defined in (2.4)
  $\hat{\mathbf{A}} \leftarrow \frac{1}{f}\sum_{k,l=1}^{f}[\check{\mathbf{Y}}]_{k,l}\overline{\hat{\mathbf{S}}_n}(l,k)$ (see Eq. (2.9))
  $\check{\mathbf{X}} \leftarrow \mathbf{G}(\hat{\mathbf{A}}, \frac{\bar{\lambda}_X}{f})$
until convergence
$\hat{\mathbf{\Theta}}_{KGlasso} \leftarrow \check{\mathbf{X}} \otimes \check{\mathbf{Y}}$

As compared to the $\mathcal{O}(p^4 f^4)$ computational complexity of Glasso, KGlasso has a computational complexity of only $\mathcal{O}(p^4 + f^4)$ [1].

2.5 Convergence of KGlasso Iterations

In this section, we provide an alternative characterization of the KGlasso algorithm (Algorithm 1) and show the iterations converge pointwise to a local minimum of the objective.

2.5.1 Block-Coordinate Reformulation of KGlasso

The following lemma shows that exploiting the property that the KGlasso algorithm is a block-coordinate optimization of the penalized objective function (2.10), each subproblem takes the form of standard Glasso applied on a compressed version of the SCM that is relevant for inference in each step.

Lemma II.1. *The KGlasso objective function (2.10) has the following properties:*

[1] In the sparse Kronecker factor case, this cost can be reduced to $\mathcal{O}(p^3 + f^3)$.

1. *Assume* $\bar{\lambda}_X, \bar{\lambda}_Y \geq 0$ *and* $\mathbf{X} \in S^p_{++}, \mathbf{Y} \in S^f_{++}$. *When one argument of* $J_\lambda(\mathbf{X}, \mathbf{Y})$ *is fixed, the objective function (2.10) is convex in the other argument.*

2. *Assume* $\hat{\mathbf{S}}_n$ *is positive definite. Consider* $J_\lambda(\mathbf{X}, \mathbf{Y})$ *in (2.10) with matrix* $\mathbf{X} \in S^p_{++}$ *fixed. Then, the dual subproblem for minimizing* $J_\lambda(\mathbf{X}, \mathbf{Y})$ *over* $\mathbf{Y}$ *is:*

$$\max_{|\mathbf{W} - \frac{1}{p}\sum_{i,j=1}^p \mathbf{X}_{i,j}\hat{\mathbf{S}}_n(j,i)|_\infty \leq \lambda_Y} \log\det(\mathbf{W}) \tag{2.11}$$

where $\lambda_Y := \bar{\lambda}_Y / p$.

On the other hand, consider (2.10) with matrix $\mathbf{Y} \in S^f_{++}$ *fixed. Then, the dual problem for minimizing* $J_\lambda(\mathbf{X}, \mathbf{Y})$ *over* $\mathbf{X}$ *is:*

$$\max_{|\mathbf{Z} - \frac{1}{f}\sum_{k,l=1}^f \mathbf{Y}_{k,l}\overline{\hat{\mathbf{S}}_n}(l,k)|_\infty \leq \lambda_X} \log\det(\mathbf{Z}) \tag{2.12}$$

where $\overline{\hat{\mathbf{S}}_n} := \mathbf{K}^T_{p,f}\hat{\mathbf{S}}_n\mathbf{K}_{p,f}$ *and* $\lambda_X := \bar{\lambda}_X / f$.

3. *Strong duality holds for (2.11) and (2.12).*

4. *The solutions to (2.11) and (2.12) are positive definite.*

Proof. See Appendix. □

Since the dual subproblems (2.11) and (2.12) are maximizations of a strictly concave function over a closed convex set they have unique solution attaining the maximum. Lemma II.1 is similar to the result obtained in [5], but with the pair $(\frac{1}{p}\sum_{i,j=1}^p \mathbf{X}_{i,j}\hat{\mathbf{S}}_n(j,i), \lambda_Y)$ playing the role of $(\hat{\mathbf{S}}_n, \lambda)$, for the fixed $\mathbf{X}$ subproblem.

2.5.2 Limit Point Characterization of KGlasso

The following theorem establishes that KGlasso converges to a local minimum of the penalized likelihood function (2.2).

Theorem II.2. *Assume* $n \geq \max(\frac{p}{f}, \frac{f}{p})+1$. *Then the KGlasso iterations converge to a local minimum of the negative penalized likelihood function (2.10) and if* $(\mathbf{X}^{(0)}, \mathbf{Y}^{(0)})$ *is not a local minimum, strict descent follows.*

Proof. This theorem is a specialization of the more general Theorem II.6. See Appendix of Theorem II.6 for proof. □

The proof of Thm. II.2 is built on several lemmas included in the Appendix. The main line of argument is as follows. For $n \geq \max(\frac{p}{f}, \frac{f}{p})+1$, the Kronecker structured MLE exists [49], and this implies that the objective function is bounded below. This can be used to show that the iterates generated by Algorithm 1 converge to a critical point. The coordinate convexity and continuity properties of the objective rule out existence of local maxima and saddle points. Combining this result with the KKT optimality conditions and the strict descent property of the algorithm, we arrive at the claim in Thm. II.2. A similar limit theorem was obtained in [2] but they only established convergence to a stationary point of (2.10).

The details follow next. We will first show that KGlasso converges to a fixed point. Let $J_\lambda(\mathbf{X}, \mathbf{Y})$ be as defined in (2.10) and define $J_\lambda^{(k)} = J_\lambda(\mathbf{X}^{(k)}, \mathbf{Y}^{(k)})$ for $k = 0, 1, 2, \ldots$.

Theorem II.3. *If* $n \geq \max(\frac{p}{f}, \frac{f}{p}) + 1$, *KGlasso converges to a fixed point. Also, we have* $J_\lambda^{(k)} \searrow J_\lambda^{(\infty)}$.

Proof. See Appendix. □

The following analysis uses Theorem II.3 to prove convergence of the KGlasso algorithm to a local minimum. To do this, we consider a more general setting. The KGlasso algorithm is a special case of Algorithm 2. Assuming a k-fold Kronecker product structure for the covariance matrix, the optimization problem (2.10) can be

written in the form:

$$J_\lambda(\mathbf{X}_1,\ldots,\mathbf{X}_k) = J_0(\mathbf{X}_1,\ldots,\mathbf{X}_k) + \sum_{i=1}^{k} J_i(\mathbf{X}_i) + \bar{\lambda}_i\eta_1(\mathbf{X}_i) \tag{2.13}$$

where $\mathbf{X}_i \in S_{++}^{d_i}$, $\eta_1(\mathbf{X}_i) := |\mathbf{X}_m|_1$, $J_0(\mathbf{X}_1,\ldots,\mathbf{X}_k) := \mathrm{tr}((\mathbf{X}_1 \otimes \mathbf{X}_2 \otimes \cdots \otimes \mathbf{X}_k)\hat{\mathbf{S}}_n)$ and $J_i(\mathbf{X}_i) = -\prod_{i'\neq i} d_{i'} \cdot \log\det(\mathbf{X}_i)$ for $i = 1,\ldots,k$.

Without loss of generality, by reshaping matrices into appropriate vectors, (2.13) can be rewritten as:

$$J_\lambda(\mathbf{x}_1,\ldots,\mathbf{x}_k) = J_0(\mathbf{x}_1,\ldots,\mathbf{x}_k) + \sum_{i=1}^{k} J_i(\mathbf{x}_i) + \bar{\lambda}_i\eta_i(\mathbf{x}_i) \tag{2.14}$$

where the optimization variable is $\mathbf{x} := [\mathbf{x}_1^T, \mathbf{x}_2^T, \ldots, \mathbf{x}_k^T]^T \in \mathbb{R}^{d'}$, where $\mathbf{x}_i \in \mathbb{R}^{d_i^2}$ and $d' = \sum_{i=1}^{k} d_i^2$. For example, $\eta_i(\mathbf{X}_i) = |\mathbf{X}_i|_1 = \|\mathrm{vec}(\mathbf{X}_i)\|_1 = \|\mathbf{x}_i\|_1 = \eta_i(\mathbf{x}_i)$. The mapping $\{J_i\}_{i=0}^k$ can be similarly written in terms of the vectors $\mathbf{x}_i$ instead of the matrices $\mathbf{X}_i$.

The reader can verify that the objective function (2.13) satisfies the properties (for $n \geq \max(\frac{p}{f}, \frac{f}{p}) + 1$) in Appendix 2.11.4.

The general optimization problem of interest here is:

$$\min_{\mathbf{x}\in\mathbb{R}^{d'}} J_\lambda(\mathbf{x}) \text{ subject to } \mathrm{vec}^{-1}(\mathbf{x}_i) = \mathbf{X}_i \in S_{++}^{d_i}, i = 1,\ldots,k \tag{2.15}$$

The positive definiteness constraints are automatically taken care of by the construction of the algorithm (see Lemma II.1.4). Let the dimension of the covariance matrix be denoted by $d := \prod_{i=1}^{k} d_i$. We assume $n > d$. To solve (2.15), a block coordinate-descent penalized algorithm is constructed:

Remark II.4. The positive definiteness constraint at each coordinate descent iteration of Algorithms 1 and 2 need not be explicit since the objective function $J_\lambda(\cdot)$ acts as a logarithmic barrier function.

Algorithm 2 Block Coordinate-Descent Penalized Algorithm

Input: $\hat{\mathbf{S}}_n$, d_i, n, $\epsilon > 0$, $\lambda_i > 0$
Output: $\hat{\mathbf{\Theta}}$
Initialize $\mathbf{X}_1^0, \mathbf{X}_2^0, \ldots, \mathbf{X}_k^0$ matrices as positive definite matrices, e.g., scaled identity.
$\hat{\mathbf{\Theta}}_0 \leftarrow \mathbf{X}_1^0 \otimes \mathbf{X}_2^0 \otimes \cdots \otimes \mathbf{X}_k^0$
$m \leftarrow 0$
repeat
  $\hat{\mathbf{\Theta}}_{\text{prev}} \leftarrow \hat{\mathbf{\Theta}}$
  $\mathbf{X}_1^m \leftarrow \arg\min_{\mathbf{A}_1 \succ 0} J_\lambda(\mathbf{A}_1, \mathbf{X}_2^{m-1}, \ldots, \mathbf{X}_k^{m-1})$
  $\mathbf{X}_2^m \leftarrow \arg\min_{\mathbf{A}_2 \succ 0} J_\lambda(\mathbf{X}_1^m, \mathbf{A}_2, \ldots, \mathbf{X}_k^{m-1})$
  $\vdots$
  $\mathbf{X}_k^m \leftarrow \arg\min_{\mathbf{A}_k \succ 0} J_\lambda(\mathbf{X}_1^m, \mathbf{X}_2^m, \ldots, \mathbf{A}_k)$
  $\hat{\mathbf{\Theta}} \leftarrow \mathbf{X}_1^m \otimes \mathbf{X}_2^m \otimes \cdots \otimes \mathbf{X}_k^m$
  $m \leftarrow m + 1$
until $\|\hat{\mathbf{\Theta}}_{\text{prev}} - \hat{\mathbf{\Theta}}\| \leq \epsilon$

Note that Algorithm 1 is a special case of Algorithm 2. An extension of Theorem II.3, assuming $n > d$ or $J_\lambda^* > -\infty$, based on induction, can be used to show that the limit points of the sequence of iterates $(\mathbf{x}^m)_{m\geq 0} = (\mathbf{x}_1^m, \ldots, \mathbf{x}_k^m)_{m\geq 0}$ are fixed points.

Remark II.5. Note that a necessary condition for $\mathbf{x}^*$ to minimize J_λ is $0 \in \partial J_\lambda(\mathbf{x}^*)$. This is not sufficient however.

We next show that the limit point(s) of $(\mathbf{x}^m)_{m\geq 0}$ are nonempty and are local minima.

Theorem II.6. *Let $(\mathbf{x}^m) = (\mathbf{x}_1^m, \ldots, \mathbf{x}_k^m)_{m\geq 0}$ be a sequence generated by Algorithm 2. Assume $n > d$* [2].

1. *The algorithm converges to a local minimum.*

2. *If $\mathbf{x}^0$ is not a local minimum, strict descent follows.*

Proof. See Appendix. □

As a consequence of Theorem II.6, we have Theorem II.2.

[2] This requirement on the sample size $n > d$ can be significantly relaxed. For the two-fold case, this can be relaxed to $n \geq \max(\frac{p}{f}, \frac{f}{p}) + 1$.

2.6 Large-Deviation Bound for Linear Combination of SCM submatrices

Since the FF and KGLasso algorithms involve matrices formed from linear combinations of submatrices of the sample covariance, it is important to understand how the concentration of measure behaves for updates of the form (2.8) and (2.9). We note that to get tight bounds on the rate of concentration is not an easy task since submatrices of the SCM are non-symmetric in general and can be highly correlated. The following theorem derives a tight bound for this rate and will be used in the proofs of Thm. II.11 and Thm. II.13.

Lemma II.7. *Let* $\mathbf{X}$ *be a* $p \times p$ *data-independent matrix. Define the linear operator* $\mathbf{T}$ *as* $\mathbf{T}(\mathbf{X}) = \hat{\mathbf{B}}(\mathbf{X}^{-1})$*, where* $\hat{\mathbf{B}}(\cdot)$ *is defined in (2.9). Assume* $\max_k[\mathbf{B}_0]_{k,k}, \|\mathbf{X}\|_2, \|\mathbf{A}_0\|_2$ *are uniformly bounded constants as* $p, f \to \infty$*. Define* $\mathbf{B}_* := \frac{tr(\mathbf{X}\mathbf{A}_0)}{p}\mathbf{B}_0$*. Let* $c, \tau > 0$*. Define* $\psi(u) = \sum_{m=0}^{\infty} \frac{(2m+2)!!}{m!} u^m$ [3]*. Let* $\bar{C} := \frac{4(2+\tau)^2 \max(2,c)}{\psi(\frac{1}{2+\tau})} < \frac{np}{\log(\max(f,n))}$ [4]*. Then, with probability* $1 - \frac{2}{\max(f,n)^c}$*,*

$$|\mathbf{T}(\mathbf{X}) - \mathbf{B}_*|_\infty \leq K(c,\tau)\sqrt{\frac{\log(\max(f,n))}{np}}$$

where $K(c,\tau) = \overline{k} \cdot \sqrt{4\psi(\frac{1}{2+\tau})\max(2,c)}$*,* $\overline{k} = \max_k[\mathbf{B}_0]_{k,k} \cdot \|\mathbf{X}\|_2 \|\mathbf{A}_0\|_2$*.*

Remark II.8. Choosing $c \leq 2$ in Lemma II.7, the best relative constant is obtained by taking τ to infinity, which yields $\sqrt{4\psi(\frac{1}{2+\tau})\max(2,c)} \to 4$.

Remark II.9. For the case of symmetric matrices $\mathbf{X} \in S^p$, the constant $\overline{k}$ can be improved to $\max_k[\mathbf{B}_0]_{k,k} \cdot \|\mathbf{X}\mathbf{A}_0\|_2$.

We provide some intuition on this bound below. Assume that $\mathbf{X}_{init} = \mathbf{X}_0$, or $\mathbf{A}_{init} = \mathbf{X}_{init}^{-1} = \mathbf{A}_0$. Define $\mathbf{W} = \mathbf{X}_0^{1/2} \otimes \mathbf{I}_p$ and $\tilde{\mathbf{z}}_t = \mathbf{W}\mathbf{z}_t$, with i.i.d. $\mathbf{z}_t \sim$

[3] The double factorial notation is defined as

$$m!! = \begin{cases} m \cdot (m-2) \cdot \cdots \cdot 3 \cdot 1 & \text{if } m > 0 \text{ is odd} \\ m \cdot (m-2) \cdot \cdots \cdot 4 \cdot 2 & \text{if } m > 0 \text{ is even} \\ 1 & \text{if } m = -1 \text{ or } m = 0 \end{cases}$$

.

[4] If $p = f = n^{c'}$ for some $c' > 0$, this condition will hold for n large enough.

$N(\mathbf{0}, \mathbf{A}_0 \otimes \mathbf{B}_0)$, $t = 1, \ldots, n$. Then, $\tilde{\mathbf{z}}_t$ has block-diagonal covariance

$$\mathrm{Cov}(\tilde{\mathbf{z}}_t) = \mathbf{I}_p \otimes \mathbf{B}_0.$$

When $\mathbf{W}$ is applied to the transformed $pf \times pf$ sample covariance matrix, $\hat{\mathbf{S}}_n^W := \mathbf{W}\hat{\mathbf{S}}_n\mathbf{W}^T$, the first step of KGlasso produces an iterate $\hat{\mathbf{Y}}_n^{(1)} = \mathbf{G}(\hat{\mathbf{B}}, \lambda_Y)$ with $\hat{\mathbf{B}} = \frac{1}{p}\sum_{i=1}^{p} \hat{\mathbf{S}}_n^W(i,i)$ (recall (2.9)). For suitable $\lambda_Y = \lambda_Y^{(1)}$, $\hat{\mathbf{Y}}_n^{(1)}$ converges to $\mathbf{Y}_0$ with respect to maximal elementwise norm at a rate $O_P\left(\sqrt{\frac{\log M}{np}}\right)$. The convergence of $\hat{\mathbf{Y}}_n^{(1)}$ is easily established by applying the Chernoff bound and invoking the jointly Gaussian property of the measurements and the block diagonal structure of $\mathrm{Cov}(\tilde{\mathbf{z}}_t)$. Lemma II.7 establishes that this rate holds even if $\mathbf{X}_{init} \neq \mathbf{X}_0$ in Assumption II.10. In view of the rate of convergence of $\hat{\mathbf{Y}}^{(1)}$, to achieve a reduction in the MSE of $\mathbf{Y}$, either the sample size n or the dimension p must increase.

2.7 High Dimensional Consistency of FF

In this section, we show that the flip-flop (FF) algorithm achieves the optimal (non-sparse) statistical convergence rate of $O_P\left(\sqrt{\frac{(p^2+f^2)\log M}{n}}\right)$. This result (see Thm. II.11) will be compared to the statistical convergence rate of KGlasso (see Thm. II.13) to establish that KGlasso has lower asymptotic MSE than FF. We make the following boundedness assumptions on the spectra of the Kronecker factors.

Assumption II.10. *Uniformly Bounded Spectra*

There exist absolute constants $\underline{k}_A, \overline{k}_A, \underline{k}_B, \overline{k}_B, \underline{k}_{A_{init}}, \overline{k}_{A_{init}}$ *such that:*

1a. $0 < \underline{k}_A \leq \lambda_{min}(\mathbf{A}_0) \leq \lambda_{max}(\mathbf{A}_0) \leq \overline{k}_A < \infty$

1b. $0 < \underline{k}_B \leq \lambda_{min}(\mathbf{B}_0) \leq \lambda_{max}(\mathbf{B}_0) \leq \overline{k}_B < \infty$

2. $0 < \underline{k}_{A_{init}} \leq \lambda_{min}(\mathbf{A}_{init}) \leq \lambda_{max}(\mathbf{A}_{init}) \leq \overline{k}_{A_{init}} < \infty$

Let $\mathbf{\Sigma}_{FF}(3) := \hat{\mathbf{A}}(\hat{\mathbf{B}}(\mathbf{A}_{init})) \otimes \hat{\mathbf{B}}(\hat{\mathbf{A}}(\hat{\mathbf{B}}(\mathbf{A}_{init})))$ denote the 3-step (noniterative) version of the flip-flop algorithm [130].

Theorem II.11. *Let* $\mathbf{A}_0, \mathbf{B}_0$*, and* $\mathbf{A}_{init}$ *satisfy Assumption II.10 and define* $M = \max(p, f, n)$*. Assume* $p \geq f \geq 2$ *and* $p \log M \leq C'' n$ *for some finite constant* $C'' > 0$*. Finally, assume* $n \geq \frac{p}{f} + 1$*. Then,*

$$\left\| \mathbf{\Theta}_{FF}(3) - \mathbf{\Theta}_0 \right\|_F = O_P\left(\sqrt{\frac{(p^2 + f^2) \log M}{n}} \right) \tag{2.16}$$

as $n \to \infty$*.*

Proof. See Appendix. □

Remark II.12. The sufficient conditions are symmetric with respect to p and f-i.e. for $f \geq p$, the corresponding conditions would become $f \log M \leq C'' n$ for some constant $C'' > 0$, and $n \geq \frac{f}{p} + 1$.

For the special case of $p = f$, the sufficient conditions of Thm. II.11 become $p \log M = O(n)$. The relation (2.16) indicates that the error is asymptotically bounded as long as n is of order $\Omega((p^2 + f^2) \log M)$. The relation (2.16) specifies the rate of reduction of the estimation error for the three step FF algorithm ($k = 3$) [130]. This relation will also hold for the multi-step FF as long as the number of steps are finite. Note that (2.16) specifies a faster rate than that of the ordinary ML sample covariance matrix estimator (2.6).

2.8 High Dimensional Consistency of KGlasso

Here a relation like (2.16) is established for KGlasso. Recall that a $p \times p$ matrix is called sparse if its number of nonzero elements is of order p. Recall $\bar{\lambda}_X = \lambda_X f$ and $\bar{\lambda}_Y = \lambda_Y p$, as in (2.10).

Theorem II.13. *Assume* $\mathbf{X}_0$ *and* $\mathbf{Y}_0$ *are sparse-i.e.,* $s_{X_0} = O(p)$ *and* $s_{Y_0} = O(f)$*. Let* $\mathbf{A}_0, \mathbf{B}_0, \mathbf{A}_{init}$ *satisfy Assumptions II.10. Let* $M = \max(p, f, n)$*. Let* $\lambda_Y^{(1)} \asymp$

$\sqrt{\frac{\log M}{np}}$, *and* $\lambda_X^{(2)}, \lambda_Y^{(3)} \asymp \left(\frac{1}{\sqrt{p}} + \frac{1}{\sqrt{f}}\right)\sqrt{\frac{\log M}{n}}$. *Then, if* $\max\left(\frac{p}{f}, \frac{f}{p}\right)\log M = o(n)$,

$$\|\mathbf{\Theta}_{KGlasso}(3) - \mathbf{\Theta}_0\|_F = O_P\left(\sqrt{\frac{(p+f)\log M}{n}}\right) \tag{2.17}$$

as $n \to \infty$.

Proof. See Appendix. □

Theorem II.13 offers a strict improvement over standard Glasso [103, 5] and generalizes Thm. 1 in [103] to the case of sparse Kronecker product structure. Thm. II.13 generalizes Thm. II.11 to the case of sparse Kronecker structure. The rate in (2.17) offers a significant improvement over the rate $O_P(\sqrt{\frac{(p+f)pf\log(pf)}{n}})$ shown in Theorem 3 in [134], as the dimensions p, f grow to infinity.

Comparison between the error expressions (2.5), (2.16) and (2.17) show that, by exploiting both Kronecker structure and sparsity, KGlasso can attain significantly lower estimation error than standard Glasso [103] and FF [130]. To achieve accurate covariance estimation for the sparse Kronecker product model, the minimal sample size needed is $n = \Omega((p+f)\log M)$.

The minimal sample size required to achieve accurate covariance estimation is graphically depicted in Fig. 2.1 for the special case $p = f$. The regions below the lines are the MSE convergence regions-i.e., the MSE convergence rate goes to zero as p, n grow together to infinity at a certain growth rate controlled by these regions. It is shown that KGlasso allows the dimension p to grow almost linearly in n and still achieve accurate covariance estimation (see (2.17)) and thus, uniformly outperforms FF, Glasso and the naive SCM estimators in the case both Kronecker factors are sparse. Although Thm. II.13 shows a rate on the inverse covariance matrix, this asymptotic rate can be shown to hold for the covariance matrix as well (see proof of Thm. II.13 in Appendix). Lemma II.7 provides a tight bound that makes the

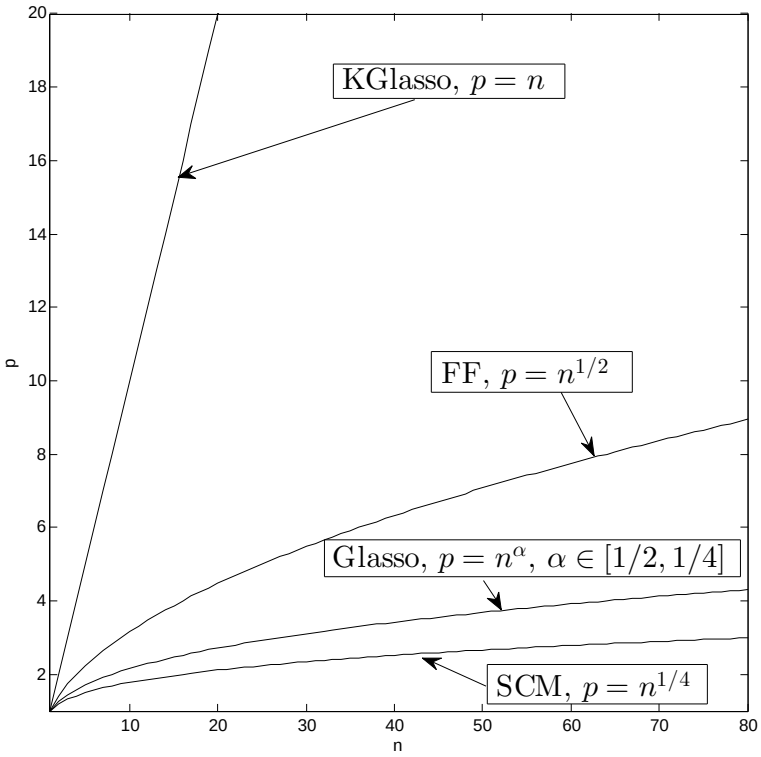


Figure 2.1: Regions of convergence for KGlasso (below upper curve), FF (below second highest curve), Glasso (below third highest curve), and standard sample covariance matrix estimator (SCM) (bottom curve). These regions are obtained from the analytical expressions in equations (2.17), (2.16), (2.5) and (2.6), respectively. The simulation shown in Fig. 2.5 establishes that the FF algorithm indeed diverges when the parameters p and n fall inbetween the KGlasso and FF curves in the above figure.

dependence of the convergence rate explicit in p, f and n. Theorem II.13 uses Lemma II.7 to show that KGlasso converges to $\mathbf{X}_0 \otimes \mathbf{Y}_0$ with rate $O_P\left(\sqrt{\frac{(p+f)\log M}{n}}\right)$ with respect to Frobenius norm.

2.9 Simulation Results

In this section, we empirically validate the convergence rates established in previous sections using Monte Carlo simulation.

Each iteration of the KGlasso involves solving an ℓ_1 penalized covariance estimation problem of dimension 100×100 (Step 6 and Step 8 of KGlasso specified by Algorithm 1). To solve these small sparse covariance estimation problems we used the Glasso algorithm of Hsieh *et al* [66] where the Glasso stopping criterion was determined by monitoring when the duality gap falls below a threshold of 10^{-3}.

In each of the simulations the true covariance matrix factors $\mathbf{X}_0 = \mathbf{A}_0^{-1}$ and $\mathbf{Y}_0 = \mathbf{B}_0^{-1}$ were unstructured randomly generated positive definite matrices. First, p random nonzero elements were placed on the diagonal of a square $p \times p$ matrix C. Then, on average p nonzero elements were placed on the off-diagonal and symmetry was imposed. On average, a total of $3p$ elements were nonzero. The resulting matrix $\tilde{\mathbf{C}}$ was regularized to produce the sparse positive definite inverse covariance $\mathbf{Y}_0 = \tilde{\mathbf{C}} + \rho \mathbf{I}_f$, where $\rho = 0.5 - \lambda_{min}(\tilde{\mathbf{C}})$. A total of $N_{MC} = 50$ simulation runs were performed for each sample size n, where n ranged from 10 to 100. Performance assessment was based on normalized Frobenius norm error in the covariance and precision matrix estimates. The normalized error was calculated using

$$\sqrt{\frac{1}{N_{MC}} \sum_{i=1}^{N_{MC}} \frac{\|\mathbf{\Sigma}_0 - \hat{\mathbf{\Sigma}}(i)\|_F^2}{\|\mathbf{\Sigma}_0\|_F^2}}$$

where $\hat{\mathbf{\Sigma}}(i)$ is the covariance estimate for the i-th simulation. The same formula was used to calculate the normalized error in the precision matrix $\hat{\mathbf{\Theta}}_0$. In the implementa-

tion of KGlasso, the regularization parameters were chosen as follows. The initialization was $\mathbf{X}_{init} = \mathbf{I}_p$. The regularization parameters were selected as $\lambda_Y^{(1)} = c_y\sqrt{\frac{\log M}{np}}$, $\lambda_X^{(2)} = c_x\sqrt{\frac{\log M}{nf}} + \lambda_Y^{(1)}$, $\lambda_Y^{(2)} = \lambda_X^{(2)}$, $\lambda_X^{(3)} = \lambda_X^{(2)}$ and so on. We set $c_x = c_y = 0.4$. In real-world applications, the constants c_x and c_y can be chosen via cross-validation or by optimizing an information criterion on the training data.

We considered the setting where $\mathbf{X}_0$ and $\mathbf{Y}_0$ are large sparse matrices of dimension $p = f = 100$ (see Fig. 2.2) yielding a covariance matrix $\mathbf{\Theta}_0$ of dimension $10,000 \times 10,000$. This dimension was too large for implementation of Glasso even when implemented using the state-of-the-art algorithm by Hsieh *et al* [66]. Approximately 2% of the elements of each precision matrix are nonzero and approximately 0.04% of the elements of the full precision matrix $\mathbf{\Theta}_0$ are nonzero. Figures 2.3 and 2.4 compare the root-mean squared error (RMSE) performance in precision and covariance matrices as a function of n. As expected, KGlasso outperforms FF over the range of n for both covariance and inverse covariance estimation problems. KGlasso outperforms FF in the small-sample regime since it exploits sparsity in addition to Kronecker structure.

We also compare KGlasso to a natural extension of the FF algorithm that accounts for both sparsity and Kronecker structure. The flip-flop thresholding method (FF/Thres) that we consider consists of first computing the FF solution and then thresholding each estimated precision matrix. To ensure a fair comparison we set the threshold level of FF/Thres that yields exactly the same sparsity factor as the KGLasso estimated precision matrices.

From Fig. 2.3 and 2.4, we observe that KGLasso outperforms all methods uniformly across all n. For $n = 10$, there is a 72% ($\approx$ 5.53 dB) RMSE reduction for the precision matrix and 41% RMSE reduction for the covariance matrix when using

KGLasso instead of FF. For $n = 10$, there is a 70% (≈ 5.26 dB) RMSE reduction for the precision matrix and 62% RMSE reduction for the covariance matrix when using KGLasso instead of FF/Thres. For $n = 100$, there is a 53% (≈ 3.28 dB) RMSE reduction for the precision matrix and 33% RMSE reduction for the covariance matrix when using KGLasso instead of FF. For $n = 100$, there is a 50% (≈ 3.01 dB) RMSE reduction for the precision matrix and 41% RMSE reduction for the covariance matrix when using KGLasso instead of FF/Thres. For the small sample regime, there is approximately a 5.53 dB reduction for the precision matrix, which is a significant performance gain. Next, we show a borderline case $p = f = \lceil n^{0.6} \rceil$. In this case, according to Thm. II.11 and Thm. II.13, the FF diverges (MSE increases in n), while the KGlasso converges (MSE decreases in n). This is illustrated in Fig. 2.5. Our predicted rates are plotted on top of the empirical curves. Dutilleul initially developed the MLE algorithm for the matrix normal model and it was first applied to detect periodicities in multivariate time series [49]. More recent real-data experiments for the Kronecker-structured covariance model have been performed for spatiotemporal data [54], recommendation systems [2] and multi-tissue gene expression data [134].

2.10 Conclusion

We established high dimensional consistency for Kronecker Glasso algorithms that use iterative ℓ_1-penalized likelihood optimization to exploit both Kronecker structure and sparsity of the covariance. A tight MSE convergence rate was derived for KGlasso, showing significantly better MSE performance than standard Glasso [103, 5] and FF [130]. In addition, our rate for KGlasso in (2.17) offers a significant improvement over the rate derived in [134] (independently from ours) in the high dimensional regime, thereby yielding a smaller sample size requirement for accurate

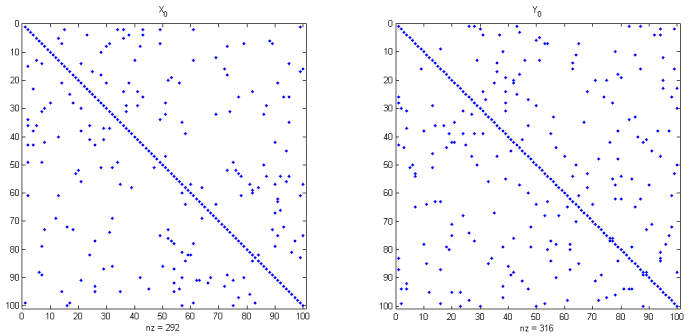


Figure 2.2: Sparse Kronecker matrix representation. Left panel: left Kronecker factor. Right panel: right Kronecker factor. As the Kronecker-product covariance matrix is of dimension $10,000 \times 10,000$, standard Glasso is not practically implementable for this example. The sparsity factor for both precision matrices is approximately 200.

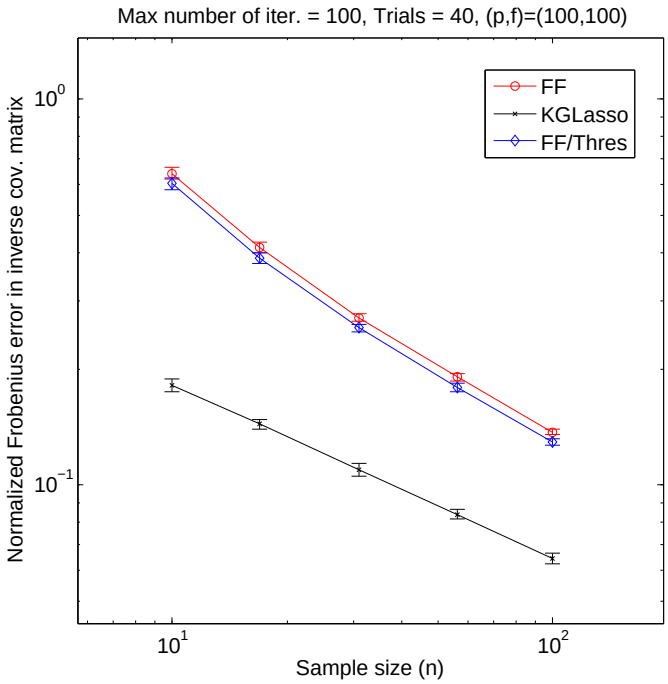


Figure 2.3: Normalized RMSE performance for precision matrix as a function of sample size n. KGlasso (Kronecker graphical lasso) uniformly outperforms FF (flip-flop) algorithm and FF/Thres (flip-flop thresholding) for all n. Here, $p = f = 100$ and $N_{MC} = 40$. The error bars are centered around the mean with $\pm$ one standard deviation. For $n = 10$, there is a 72% RMSE reduction from the FF to KGLasso solution and a 70% RMSE reduction from the FF/Thres to KGLasso.

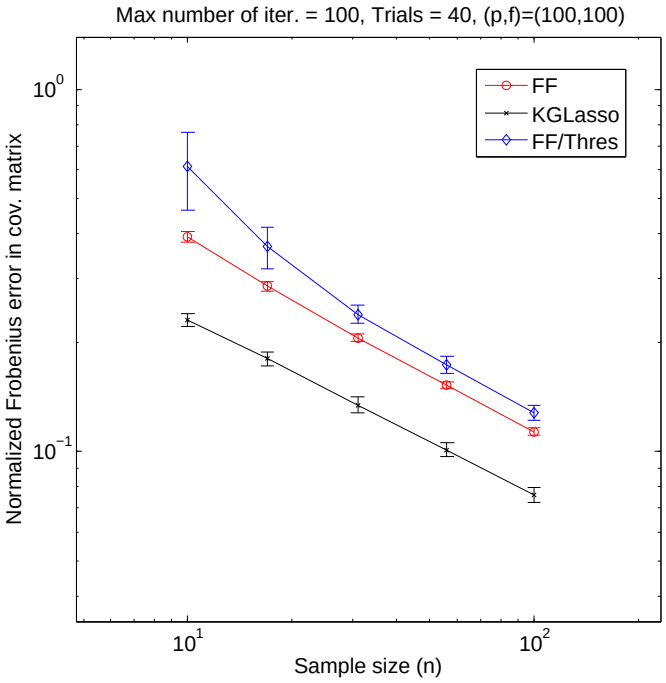


Figure 2.4: Normalized RMSE performance for covariance matrix as a function of sample size n. KGlasso (Kronecker graphical lasso) uniformly outperforms FF (flip-flop) algorithm for all n. Here, $p = f = 100$ and $N_{MC} = 40$. The error bars are centered around the mean with $\pm$ one standard deviation. For $n = 10$, there is a 41% RMSE reduction from the FF to KGLasso solution and a 62% RMSE reduction from the FF/Thres to KGLasso.

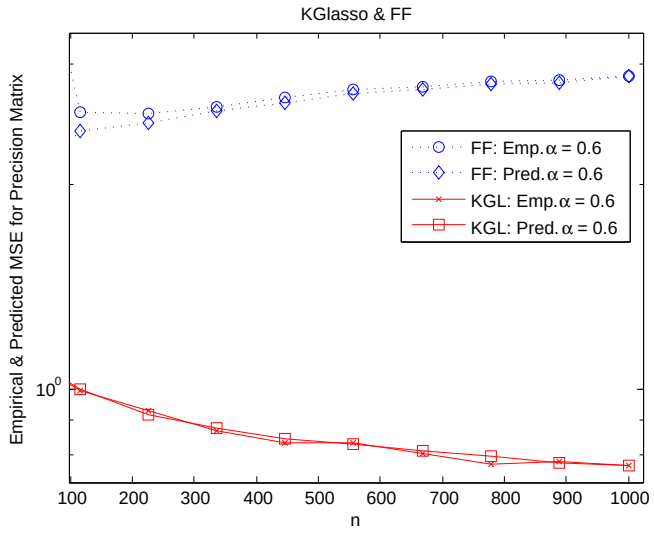


Figure 2.5: Precision Matrix MSE as a function of sample size n for FF and KGlasso. The dimensions of the Kronecker factor matrices grow as a function of n as: $p(n) = f(n) = \lceil n^{0.6} \rceil$. The true Kronecker factors were set to identity (so their inverses are fully sparse). The predicted MSE curves according to Thm. II.11 and Thm. II.13 are also shown. As predicted by our theory, and by the predicted convergent regions of (n, p) for FF and KGlasso in Fig. 2.1, the MSE of the FF diverges while the MSE of the KGlasso converges as n increases.

covariance estimation under the sparse Kronecker covariance model. Simulations validated our theoretical predictions.

As expected, the proposed KGlasso algorithm outperforms both algorithms (Glasso, FF) that do not exploit all prior knowledge about the covariance matrix, i.e., sparsity and Kronecker product structure, that KGlasso exploits. The theory and experiments in this chapter establish that this performance gain is substantial, more so as the variable dimension increases. Furthermore, as compared to a simple thresholded FF algorithm, which does account for both sparsity and Kronecker structure, KGlasso has significantly better estimation performance.

2.11 Appendix

2.11.1 Proof of Lemma II.1

Proof. 1. Without loss of generality, fix $\mathbf{Y} \in S^f_{++}$. The function $\mathrm{tr}((\mathbf{X} \otimes \mathbf{Y})\hat{\mathbf{S}}_n)$ is linear in $\mathbf{X}$. The function $g(\mathbf{X}_1) := -\log\det(\mathbf{X}_1)$ is a convex function in $\mathbf{X}_1$ over the set S^p_{++} [18]. The triangle inequality implies $|\cdot|_1$ is convex. Finally, the sum of convex functions is convex. The set S^p_{++} is a convex set for any $p \in \mathbb{N}$.

2. By symmetry we only need prove that (2.12) is the dual of $\min_{\mathbf{Y}\in S^f_{++}} J_\lambda(\mathbf{X}, \mathbf{Y})$. By standard duality relations between ℓ_1 and ℓ_∞ norms [18] and symmetry of $\mathbf{Y}$ [5]:

$$|\mathbf{Y}|_1 = \max_{\mathbf{U}\in S^f : |\mathbf{U}|_\infty \le 1} \mathrm{tr}(\mathbf{Y}\mathbf{U})$$

Using this in (2.10) and invoking the saddlepoint inequality:

$$\begin{aligned} &\min_{\mathbf{Y}\in S^f_{++}} \mathrm{tr}((\mathbf{X}\otimes\mathbf{Y})\hat{\mathbf{S}}_n) - p\log\det(\mathbf{Y}) + p\lambda_Y|\mathbf{Y}|_1 \\ &\ge \max_{|\mathbf{U}|_\infty\le\lambda_Y} \min_{\mathbf{Y}\in S^f_{++}} \Big\{\mathrm{tr}((\mathbf{X}\otimes\mathbf{Y})\hat{\mathbf{S}}_n) - p\log\det(\mathbf{Y}) + p\mathrm{tr}(\mathbf{Y}\mathbf{U})\Big\} \end{aligned} \tag{2.18}$$

[5] The maximum is attained at $\mathbf{U}_{i,j} = \frac{\mathbf{Y}_{i,j}}{|\mathbf{Y}_{i,j}|}$ for $\mathbf{Y}_{i,j} \neq 0$ and at $\mathbf{U}_{i,j} = 0$ for $\mathbf{Y}_{i,j} = 0$.

When the equality in (2.18) is achieved, $(\mathbf{U}, \mathbf{Y})$ is a saddlepoint and the duality gap is zero. Rewrite the objective function, denoted $\tilde{J}_\lambda(\cdot,\cdot)$, in the minimax operation (2.18):

$$\tilde{J}_\lambda(\mathbf{X}, \mathbf{Y}) := \mathrm{tr}((\mathbf{X} \otimes \mathbf{Y})(\hat{\mathbf{S}}_n + \tilde{\mathbf{U}}(\mathbf{X}))) - p \log\det(\mathbf{Y})$$

where $\tilde{\mathbf{U}}(\mathbf{X}) = p\frac{\mathbf{I}_p \otimes \mathbf{U}}{\mathrm{tr}(\mathbf{X})}$. Define $\mathbf{M} = \hat{\mathbf{S}}_n + \tilde{\mathbf{U}}(\mathbf{X})$. To evaluate $\min_{\mathbf{Y} \in S^f_{++}} \tilde{J}_\lambda(\mathbf{X}, \mathbf{Y})$ in (2.18), we invoke the KKT conditions to obtain the solution

$$\mathbf{Y} = \left(\frac{1}{p} \sum_{i,j=1}^{p} \mathbf{X}_{i,j} \mathbf{M}(j,i) \right)^{-1}.$$

Define $\mathbf{W} = \mathbf{Y}^{-1}$ as the dual space variable. Using this in (2.18):

$$\max_{|\mathbf{W} - \frac{1}{p}\sum_{i,j=1}^{p} \mathbf{X}_{i,j} \hat{\mathbf{S}}_n(j,i)|_\infty \leq \lambda_Y} p \log\det(\mathbf{W}) + pf \tag{2.19}$$

where the constraint set was obtained after noting that $\tilde{\mathbf{U}}(\mathbf{X})(j,i) = \frac{p\mathbf{U}}{\mathrm{tr}(\mathbf{X})} I(j = i)$, and $I(\cdot)$ is the indicator function. It is evident that (2.19) is equivalent to (2.11).

3. It suffices to verify that the duality induced by the saddle point formulation is equivalent to Lagrangian duality (see Section 5.4 in [18]). Slater's constraint qualification (see Section 5.3.2 in [18]) trivially holds for the convex problem $\min_{\mathbf{Y} \in S^f_{++}} \tilde{J}_\lambda(\mathbf{X}, \mathbf{Y})$, and thus for the corresponding convex problem $\min_{\mathbf{Y} \in S^f_{++}} J_\lambda(\mathbf{X}, \mathbf{Y})$. Since the objective function of each dual problem has an optimal objective that is bounded below, Slater's constraint qualification also implies that the dual optimal solution is attained.

4. From [130], it follows that if $\hat{\mathbf{S}}_n$ is p.d., each "compression step" (see lines 6 and 8 in Algorithm 1) yields a p.d. matrix. Combining this with the positive definiteness of the Glasso estimator [5], we conclude that the first subiteration

of KGlasso yields a p.d. matrix. A simple induction, combined with the fact that the Kronecker product of p.d. matrices is p.d., establishes that (2.11) and (2.12) are p.d.

□

2.11.2 Proof of Theorem II.3

Proof. Recall that the basic optimization problem (2.3) is

$$\min_{\mathbf{X}\in S^p_{++},\mathbf{Y}\in S^f_{++}} J_\lambda(\mathbf{X},\mathbf{Y})$$

Let $J^* := \inf_{\mathbf{X}\in S^p_{++},\mathbf{Y}\in S^f_{++}} J_\lambda(\mathbf{X},\mathbf{Y})$ be the optimal primal value. Note that $J^*_\lambda > -\infty$ when $n \geq \max(\frac{p}{f},\frac{f}{p})+1$. Now, consider the first step in Algorithm 1. Fix $\mathbf{X} = \mathbf{X}^{(k-1)}$ and optimize over $\mathbf{Y} \in S^f_{++}$. Invoking Lemma II.1, we have $\mathbf{Y}^{(k)} = \arg\min_{\mathbf{Y}\in S^f_{++}} J_\lambda(\mathbf{X}^{(k-1)},\mathbf{Y})$. Note, by induction $\mathbf{Y}^{(k)}$ remains positive definite if $\mathbf{X}^{(0)}$ is positive definite. Considering the second step in Algorithm 1, we fix $\mathbf{Y} = \mathbf{Y}^{(k)}$ and obtain $\mathbf{X}^{(k)} = \arg\min_{\mathbf{X}\in S^p_{++}} J_\lambda(\mathbf{X},\mathbf{Y}^{(k)})$, so that

$$J_\lambda(\mathbf{X}^{(k)},\mathbf{Y}^{(k)}) \leq J_\lambda(\mathbf{X}^{(k-1)},\mathbf{Y}^{(k)}) \leq J_\lambda(\mathbf{X}^{(k-1)},\mathbf{Y}^{(k-1)}) \tag{2.20}$$

By induction on the number of iterations of the penalized flip-flop algorithm, we conclude that the iterates yield a nonincreasing sequence of objective functions. Since $\lambda_X|\mathbf{X}|_1, \lambda_Y|\mathbf{Y}|_1 \geq 0$, we see that the objective function evaluated at the Kronecker structured MLE provides a lower bound to the optimal primal value [6]

$$J_\lambda(\mathbf{X}_{KGlasso},\mathbf{Y}_{KGlasso}) \geq J^*_\lambda \geq J_\lambda(\mathbf{X}_{MLE},\mathbf{Y}_{MLE}) > -\infty \tag{2.21}$$

Thus, the sequence $\{J^{(k)}_\lambda : k \geq 0\}$ forms a nonincreasing sequence bounded below (since for $n > pf$, the log-likelihood function is bounded above by the log-likelihood evaluated at the sample mean and sample covariance matrix). The mono-

[6] The Kronecker structured MLE $(\mathbf{X}_{MLE},\mathbf{Y}_{MLE})$ exists for $n \geq \max(\frac{p}{f},\frac{f}{p})+1$.

tone convergence theorem for sequences [7] implies that $\{J_\lambda^{(k)}\}$ converges monotonically to $J_\lambda^{(\infty)} = \inf_k J_\lambda^{(k)}$. By the alternating minimization, we conclude that the sequence of iterates $\{(\mathbf{X}^{(k)}, \mathbf{Y}^{(k)})\}_k$ converges since the minimizer at each Glasso step is unique. □

2.11.3 Subdifferential Calculus Review

As sparse Kronecker Glasso involves non-smooth objective functions, we review a few definitions and facts from subdifferential calculus [101].

Definition II.14. By J-attentive convergence denoted as, $\mathbf{x}^n \xrightarrow{J} \mathbf{x}$, we mean that: $\mathbf{x}^n \to \mathbf{x}$ with $J(\mathbf{x}^n) \to J(\mathbf{x})$ as $n \to \infty$.

The role of J-attentive convergence is to make sure that subgradients at a point $\overline{\mathbf{x}}$ reflect no more than the local geometry of $epi(J)$ around $(\overline{\mathbf{x}}, J(\overline{\mathbf{x}}))$.

Definition II.15. Consider a proper lower semicontinuous (LSC) function $g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$. Let $\overline{\mathbf{x}}$ be such that $J(\overline{\mathbf{x}}) < \infty$.

For $\mathbf{v} \in \mathbb{R}^d$,

a) $\mathbf{v}$ is a *regular subgradient* of J at $\overline{\mathbf{x}}$ (i.e., $\mathbf{v} \in \hat{\partial} J(\overline{\mathbf{x}})$) if

$$\liminf_{\mathbf{x} \neq \overline{\mathbf{x}}, \mathbf{x} \to \overline{\mathbf{x}}} \frac{J(\mathbf{x}) - J(\overline{\mathbf{x}}) - \mathbf{v}^T(\mathbf{x} - \overline{\mathbf{x}})}{\|\mathbf{x} - \overline{\mathbf{x}}\|} \geq 0$$

.

b) $\mathbf{v}$ is a *general subgradient* of J at $\overline{\mathbf{x}}$ (i.e., $\mathbf{v} \in \partial J(\overline{\mathbf{x}})$) if there exists subsequences $\mathbf{x}^n \xrightarrow{J} \overline{\mathbf{x}}$ and $\mathbf{v}^n \in \hat{\partial} J(\mathbf{x}^n)$ such that $\mathbf{v}^n \to \mathbf{v}$.

Let $\overline{\mathbf{x}}$ be such that $J(\overline{\mathbf{x}}) < \infty$. It can be shown that $\partial J(\overline{\mathbf{x}}) = \limsup_{\mathbf{x} \xrightarrow{J} \overline{\mathbf{x}}} \hat{\partial} J(\mathbf{x})$, $\hat{\partial} J(\overline{\mathbf{x}}) \subset \partial J(\overline{\mathbf{x}})$ and both sets are closed.

Define the set of critical points $C_J := \{\mathbf{x} : 0 \in \partial J(\mathbf{x})\} = C_{J,min} \cup C_{J,saddle} \cup C_{J,max}$, where $C_{J,min}$ contains all the local minima, $C_{J,saddle}$ contains all the saddle points and $C_{J,max}$ contains all the local maxima.

Definition II.16. Let $A \subseteq \mathbb{R}^n$. Define the distance from a point $\mathbf{x} \in \mathbb{R}^n$ to the set A as $d(\mathbf{x}, A) := \inf_{\mathbf{a}\in A} \|\mathbf{x} - \mathbf{a}\|_2$.

2.11.4 Properties of objective function J_λ

The following set of properties will be used in Lemmas II.18, II.19 and Theorem II.6.

Property II.17. *1. $J_0 : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable (i.e., $f_0 \in C^1$)*

2. $\nabla J_0 : \mathbb{R}^d \to \mathbb{R}^d$ is uniformly continuous on bounded subsets $B \subset \mathbb{R}^d$

3. $J_i : \mathbb{R}^{d_i} \to \mathbb{R} \cup \{+\infty\}$ is proper [7] and lower semicontinuous (LSC), for $i = 1, \ldots, k$

4. $\eta_i : \mathbb{R}^d \to \mathbb{R}_+$ is uniformly continuous and bounded on bounded subsets $B \subset \mathbb{R}^d$, for $i = 1, \ldots, k$

5. J_λ is bounded below-i.e. $J^*_\lambda > -\infty$

6. J_λ is strictly convex in at least one block (for all the rest of the blocks held fixed)

where $J^*_\lambda = \inf_{\mathbf{X}_i \in S^{d_i}_{++}} J_\lambda(\mathbf{X}_1, \ldots, \mathbf{X}_k)$ is the optimal primal value.

2.11.5 Lemma II.18

Lemma II.18. *Given the notation established in Definition II.15 and J_λ given by (2.14), we have:*

$$\begin{aligned} \partial J_\lambda(\mathbf{x}_1, \ldots, \mathbf{x}_k) &= \times_{i=1}^k \{\nabla_{\mathbf{x}_i} J_0(\mathbf{x}_1, \ldots, \mathbf{x}_k) + \partial J_i(\mathbf{x}_i) + \bar{\lambda}_i \partial \eta_i(\mathbf{x}_i)\} \\ &= \times_{i=1}^k \{\partial_{\mathbf{x}_i} J_\lambda(\mathbf{x}_1, \ldots, \mathbf{x}_k)\} \end{aligned} \tag{2.22}$$

where $\partial_{\mathbf{x}_i} J_\lambda(\mathbf{x}_1, \ldots, \mathbf{x}_k)$ is the partial differential operator while all $\{\mathbf{x}_j : j \neq i\}$ are held fixed.

[7] A function $J : \mathbb{X} \to \mathbb{R} \cup \{\pm\infty\}$ is proper if $dom(J) = \{x \in \mathbb{X} : J(x) < \infty\} \neq \emptyset$ and $J(x) > -\infty, \forall x \in \mathbb{X}$.

Proof. First note that we have:

$$\partial J_\lambda(\mathbf{x}_1,\ldots,\mathbf{x}_k) = \nabla J_0(\mathbf{x}_1,\ldots,\mathbf{x}_k) + \partial\{\sum_{i=1}^{k} J_i(\mathbf{x}_i) + \bar{\lambda}_i\eta_i(\mathbf{x}_i)\} \tag{2.23}$$

$$= \nabla J_0(\mathbf{x}_1,\ldots,\mathbf{x}_k) + \partial\{\sum_{i=1}^{k} J_i(\mathbf{x}_i)\} + \partial\{\sum_{i=1}^{k} \bar{\lambda}_i\eta_i(\mathbf{x}_i)\} \tag{2.24}$$

$$= \nabla J_0(\mathbf{x}_1,\ldots,\mathbf{x}_k) + \times_{i=1}^{k}\{\partial J_i(\mathbf{x}_i)\} + \times_{i=1}^{k}\{\bar{\lambda}_i\partial\eta_i(\mathbf{x}_i)\} \tag{2.25}$$

$$= \times_{i=1}^{k}\{\nabla_{\mathbf{x}_k} J_0(\mathbf{x}_1,\ldots,\mathbf{x}_k) + \partial J_i(\mathbf{x}_i) + \bar{\lambda}_i\partial\eta_i(\mathbf{x}_i)\} \tag{2.26}$$

where (2.23) follows from Property II.17 and Exercise 8.8(c) in [101], (2.24) follows from Corollary 10.9 in [101], (2.25) follows from Proposition 10.5 and Equation 10(6) p.438 in [101] since $\lambda_i > 0$, and finally (2.26) follows from Minkowski sum properties.

□

2.11.6 Lemma II.19

Lemma II.19. *Let m denote the iteration index. For $m \in \mathbb{N}$, define:*

$$\begin{aligned}
(\mathbf{x}_1^m)^\circ &:= \nabla_{\mathbf{x}_1} J_0(\mathbf{x}_1^m, \mathbf{x}_2^m \ldots, \mathbf{x}_k^m) - \nabla_{\mathbf{x}_1} J_0(\mathbf{x}_1^m, \mathbf{x}_2^{m-1} \ldots, \mathbf{x}_k^{m-1}) \\
(\mathbf{x}_2^m)^\circ &:= \nabla_{\mathbf{x}_2} J_0(\mathbf{x}_1^m, \mathbf{x}_2^m \ldots, \mathbf{x}_k^m) - \nabla_{\mathbf{x}_2} J_0(\mathbf{x}_1^m, \mathbf{x}_2^m, \mathbf{x}_3^{m-1} \ldots, \mathbf{x}_k^{m-1}) \\
&\vdots \\
(\mathbf{x}_j^m)^\circ &:= \nabla_{\mathbf{x}_j} J_0(\mathbf{x}_1^m, \mathbf{x}_2^m \ldots, \mathbf{x}_k^m) - \nabla_{\mathbf{x}_j} J_0(\mathbf{x}_1^m, \ldots, \mathbf{x}_j^m, \mathbf{x}_{j+1}^{m-1} \ldots, \mathbf{x}_k^{m-1}) \\
&\vdots \\
(\mathbf{x}_k^m)^\circ &:= 0
\end{aligned}$$

Then, $((\mathbf{x}_1^m)^\circ,\ldots,(\mathbf{x}_k^m)^\circ) \in \partial J_\lambda(\mathbf{x}_1^m,\ldots,\mathbf{x}_k^m)$. Also, for all convergent subsequences $(\mathbf{x}^{m_j})_j$ of the sequence $(\mathbf{x}^m)_m$, we have

$$d(0, \partial J_\lambda(\mathbf{x}_1^{m_j},\ldots,\mathbf{x}_k^{m_j})) \to 0 \text{ as } j \to \infty$$

Proof. From Algorithm 2, we have:

$$\begin{aligned}
\mathbf{x}_1^m &\in \arg\min_{\mathbf{x}_1} J_\lambda(\mathbf{x}_1, \mathbf{x}_2^{m-1}, \ldots, \mathbf{x}_k^{m-1}) \\
\mathbf{x}_2^m &\in \arg\min_{\mathbf{x}_2} J_\lambda(\mathbf{x}_1^m, \mathbf{x}_2, \mathbf{x}_3^{m-1}, \ldots, \mathbf{x}_k^{m-1}) \\
&\vdots \\
\mathbf{x}_k^m &\in \arg\min_{\mathbf{x}_k} J_\lambda(\mathbf{x}_1^m, \ldots, \mathbf{x}_{k-1}^m, \mathbf{x}_k)
\end{aligned}$$

The first subiteration step of the algorithm implies that $0 \in \partial_{\mathbf{x}_1} J_\lambda(\mathbf{x}_1^m, \mathbf{x}_2^{m-1}, \ldots, \mathbf{x}_k^{m-1})$, the second subiteration step implies $0 \in \partial_{\mathbf{x}_2} J_\lambda(\mathbf{x}_1^m, \mathbf{x}_2^m, \mathbf{x}_3^{m-1}, \ldots, \mathbf{x}_k^{m-1})$, etc. Rewriting these using Lemma II.18, we have:

$$\begin{aligned}
0 &\in \nabla_{\mathbf{x}_1} J_0(\mathbf{x}_1^m, \mathbf{x}_2^{m-1}, \ldots, \mathbf{x}_k^{m-1}) + \partial J_1(\mathbf{x}_1^m) + \bar{\lambda}_1 \partial \eta_1(\mathbf{x}_1^m) \\
0 &\in \nabla_{\mathbf{x}_2} J_0(\mathbf{x}_1^m, \mathbf{x}_2^m, \mathbf{x}_3^{m-1}, \ldots, \mathbf{x}_k^{m-1}) + \partial J_2(\mathbf{x}_2^m) + \bar{\lambda}_2 \eta_2(\mathbf{x}_2^m) \\
&\vdots \\
0 &\in \nabla_{\mathbf{x}_k} J_0(\mathbf{x}_1^m, \mathbf{x}_2^m, \ldots, \mathbf{x}_k^m) + \partial J_k(\mathbf{x}_k^m) + \bar{\lambda}_k \partial \eta_k(\mathbf{x}_k^m)
\end{aligned}$$

This implies that for $i = 1, \ldots, k$:

$$(\mathbf{x}_i^m)^\circ \in \nabla_{\mathbf{x}_i} J_0(\mathbf{x}_1^m, \mathbf{x}_2^m, \ldots, \mathbf{x}_k^m) + \partial J_i(\mathbf{x}_i^m) + \bar{\lambda}_i \partial \eta_i(\mathbf{x}_i^m)$$

It is important to note that $\partial \eta_i(\mathbf{x}) \neq \emptyset, \forall \mathbf{x} \in \mathbb{R}^{d_i}$, for $i = 1, \ldots, k$, as a result of property II.17.4. To see why, apply Corollary 8.10 in [101] since η_i is finite and locally LSC at every point in its domain. This in turn implies $((\mathbf{x}_1^m)^\circ, \ldots, (\mathbf{x}_k^m)^\circ) \in \partial J_\lambda(\mathbf{x}_1^m, \ldots, \mathbf{x}_k^m)$ by Lemma II.18.

Now, take an arbitrary convergent subsequence $(\mathbf{x}_1^{m_j}, \ldots, \mathbf{x}_k^{m_j})_j$ of $(\mathbf{x}_1^m, \ldots, \mathbf{x}_k^m)_m$. The convergence of $(\mathbf{x}_1^{m_j}, \ldots, \mathbf{x}_k^{m_j})_j$ implies the convergence of $(\mathbf{x}_1^{m_j}, \mathbf{x}_2^{m_j-1}, \ldots, \mathbf{x}_k^{m_j-1})_j$, and $(\mathbf{x}_1^{m_j}, \ldots, \mathbf{x}_i^{m_j}, \mathbf{x}_{i+1}^{m_j-1}, \ldots, \mathbf{x}_k^{m_j-1})_j$ for $i = 2, \ldots, k-1$. Taking $j \to \infty$ and using properties II.17.2, we conclude

$$\lim_{j \to \infty} d(0, \partial J_\lambda(\mathbf{x}_1^{m_j}, \ldots, \mathbf{x}_k^{m_j})) = 0$$

since $\lim_{j\to\infty}((\mathbf{x}_1^{m_j})^\circ, \ldots, (\mathbf{x}_k^{m_j})^\circ) = (0, \ldots, 0)$.

□

2.11.7 Proof of Theorem II.6

Proof. 1. Let $L(\mathbf{x}^0) = L(\mathbf{x}_1^0, \ldots, \mathbf{x}_k^0)$ be the set of all limit points of $(\mathbf{x}^m)_{m\geq 0}$ starting from $\mathbf{x}^0$. The block-coordinate descent algorithm, Algorithm 2, implies

$$\begin{aligned} &J_0(\mathbf{x}_1^m, \mathbf{x}_2^{m-1}, \ldots, \mathbf{x}_k^{m-1}) + J_1(\mathbf{x}_1^m) + \bar{\lambda}_1\eta_1(\mathbf{x}_1^m) \\ &\quad \leq J_0(\alpha_1, \mathbf{x}_2^{m-1}, \ldots, \mathbf{x}_k^{m-1}) + J_1(\alpha_1) + \bar{\lambda}_1\eta_1(\alpha_1) \end{aligned}$$

for any $\alpha_1 \in \mathbb{R}^{d_1^2}$. Now, assume there exists a subsequence $(\mathbf{x}^{m_j})_j$ of $(\mathbf{x}^m)_m$ that converges to $\mathbf{x}^*$, where $\mathbf{x}^*$ is a limit point. This implies $(\mathbf{x}_1^{m_j}, \mathbf{x}_2^{m_j-1}, \ldots, \mathbf{x}_k^{m_j-1}) \to \mathbf{x}^*$ as $j \to \infty$. The above inequality combined with properties II.17.1 and II.17.4 (i.e. the continuity J_0 and η_i) then implies that

$$\begin{aligned} \limsup_{j\to\infty} J_1(\mathbf{x}_1^{m_j}) + J_0(\mathbf{x}_1^*, \ldots, \mathbf{x}_k^*) &\leq J_1(\alpha_1) \\ &\quad + J_0(\alpha_1, \mathbf{x}_2^*, \ldots, \mathbf{x}_k^*) + \bar{\lambda}_1(\eta_1(\alpha_1) - \eta_1(\mathbf{x}_1^*)) \end{aligned}$$

for all $\alpha_1 \in \mathbb{R}^{d_1^2}$. Taking $\alpha_1 = \mathbf{x}_1^*$ then yields $\limsup_{j\to\infty} J_1(\mathbf{x}_1^{m_j}) \leq J_1(\mathbf{x}_1^*)$. Using the lower semicontinuity property of J_1 (property II.17.3), we have

$$\liminf_{j\to\infty} J_1(\mathbf{x}_1^{m_j}) \geq J_1(\mathbf{x}_1^*)$$

. Thus, $\lim_{j\to\infty} J_1(\mathbf{x}_1^{m_j}) = J_1(\mathbf{x}_1^*)$.

By a similar line of reasoning, it can be shown that $J_i(\mathbf{x}_i^{m_j}) \to J_i(\mathbf{x}_i^*)$ as $j \to \infty$, for $i = 1, \ldots, k$. As a result, $\sum_{i=1}^k J_i(\mathbf{x}_i^{m_j}) \to \sum_{i=1}^k J_i(\mathbf{x}_i^*)$ as $j \to \infty$. Since $J_0(\cdot)$ is jointly continuous, $J_0(\mathbf{x}_1^{m_j}, \ldots, \mathbf{x}_k^{m_j}) \to J_0(\mathbf{x}_1^*, \ldots, \mathbf{x}_k^*)$. By continuity of $\eta_i(\cdot)$, $\sum_{i=1}^k \bar{\lambda}_i\eta_i(\mathbf{x}_i^{m_j}) \to \sum_{i=1}^k \bar{\lambda}_i\eta_i(\mathbf{x}_i^*)$. Thus, $J_\lambda(\mathbf{x}^{m_j}) \to J_\lambda(\mathbf{x}^*)$ as $j \to \infty$.

Now, Lemma II.19 implies that $((\mathbf{x}^{m_j})^\circ) \in \partial J_\lambda(\mathbf{x}^{m^j})$. Since the subsequence $(\mathbf{x}^{m_j})_j$ is convergent, by Lemma II.19, we have $(\mathbf{x}^{m_j})^\circ \to 0$ as $j \to \infty$. As a result, since $\partial J_\lambda(\mathbf{x}^{m_j})$ is closed (see Theorem 8.6 in [101]) for all j, we conclude that $\mathbf{x}^* \in C_J$. Thus, $L(\mathbf{x}^0) \subseteq C_J$.

We have thus proved that limit points are critical points of the objective function.

We can rule out convergence to local maxima thanks to property II.17.6. Let us show this rigorously. Assume there exists a local maximum at $\mathbf{x}' = (\mathbf{x}'_1, \ldots, \mathbf{x}'_k)$. Then, there exists $r > 0$ such that $J_\lambda(\mathbf{x}) \leq J_\lambda(\mathbf{x}')$ for all $\mathbf{x}$ such that $\|\mathbf{x} - \mathbf{x}'\|_2 < r$. Fix $\mathbf{x}_i = \mathbf{x}'_i$ for all $i \neq 1$. Without loss of generality, assume J_λ is strictly convex in the first block. Since strict convexity is maintained through linear transformation, without loss of generality, assume $d_1 = 1$. Let $\epsilon < r$. Define $x_{1,\epsilon} = x'_1 - \epsilon$ and $x_{2,\epsilon} = x'_1 + \epsilon$. Define $x_\theta = \theta x_{1,\epsilon} + (1-\theta)x_{2,\epsilon}$, where $\theta \in (0,1)$. Since $\|[x_\theta; \mathbf{x}_{\neq 1}] - \mathbf{x}'\|_2 = |x_\theta - x'_1| = \epsilon(1-2\theta) < r$, by the local maximum definition, there exists $\epsilon \in (0, r)$ small enough such that

$$\theta J_\lambda(x_{1,\epsilon}, \mathbf{x}'_{\neq 1}) + (1-\theta) J_\lambda(x_{2,\epsilon}, \mathbf{x}'_{\neq 1}) \leq J_\lambda(x_\theta, \mathbf{x}'_{\neq 1})$$

for some $\theta \in (0,1)$. Since $\epsilon > 0$, we have $x_{1,\epsilon} \neq x_{2,\epsilon}$, and this contradicts strict convexity. Thus, there are no local maxima. [8]

Next, we use the non-existence of local maxima and continuity of J_λ to rule out convergence to saddle points. Assume there exists a saddlepoint at $\mathbf{x}_s$. Then, by definition, $0 \in J_\lambda(\mathbf{x}_s)$ and $\mathbf{x}_s$ is not a local maximum or a local minimum. Since $\mathbf{x}_s$ is not a local minimum, for all $\epsilon > 0$, there exists a point $\mathbf{x}'$ such that $\|\mathbf{x}' - \mathbf{x}_s\|_2 < \epsilon$ and $J_\lambda(\mathbf{x}_s) > J_\lambda(\mathbf{x}')$. By continuity, it follows that there exists

[8] An alternative way to get a contradiction is to assume there exists a strict local maximum and use only convexity, instead of strict convexity.

$\delta > 0$ such that for all $\mathbf{x}$ satisfying $\|\mathbf{x} - \mathbf{x}'\|_2 < \delta$, we have $J_\lambda(\mathbf{x}_s) > J_\lambda(\mathbf{x})$, which implies that $\mathbf{x}_s$ is a local maximum. This is a contradiction and thus, $\mathbf{x}_s$ is a local minimum. So, no saddle points exist.

Theorem II.3 implies that $L(\mathbf{x}^0)$ is nonempty and singleton.

2. We show that if we do not start at a local minimum, strict descent follows. Let $\mu(\cdot)$ denote the point-to-point mapping during one iteration step, i.e., $\mathbf{x}^{m+1} = \mu(\mathbf{x}^m)$. We show that if $\mathbf{x}^0 \notin C_J$, then $L(\mathbf{x}^0) \subseteq C_{J,min}$. The result then follows by using the proof of the first part [9]. To this end, let $\mathbf{x}^{'}$ be a fixed point under μ, i.e., $\mu(\mathbf{x}^{'}) = \mathbf{x}^{'}$. Then, the subiteration steps of the algorithm yield $0 \in \partial_{\mathbf{x}_i} J_\lambda(\mathbf{x}_1^{'}, \ldots, \mathbf{x}_k^{'})$ for $i = 1, \ldots, k$, which implies $0 \in \partial J_\lambda(\mathbf{x}^{'})$, i.e., $\mathbf{x}^{'} \in C_J$. The contrapositive implies that if $\mathbf{x} \notin C_J$, then $J_\lambda(\mu(\mathbf{x})) < J_\lambda(\mathbf{x})$ (strict descent). A simple induction on the number of iterations then concludes the proof.

□

2.11.8 Proof of Lemma II.7

Proof. This proof is based on a large-deviation theory argument. Fix $(k, l) \in \{1, \ldots, f\}^2$. Note that $\mathbb{E}[\mathbf{T}(\mathbf{X})] = \mathbf{B}_*$. First we bound the upper tail probability on the difference $\mathbf{T}(\mathbf{X}) - \mathbf{B}_*$ and then we turn to the lower tail probability. Bounding the upper tail

[9] The first part of the proof showed $C_J = C_{J,min}$.

by using Markov's inequality, we have

$$\begin{aligned}
&\Pr\left([\mathbf{T}(\mathbf{X})]_{k,l} - [\mathbf{B}_*]_{k,l} > \epsilon\right) \\
&\quad = \Pr\left(\frac{1}{p}\sum_{i,j=1}^{p} \mathbf{X}_{i,j}[\hat{\mathbf{S}}_n(j,i)]_{k,l} - \frac{\mathrm{tr}(\mathbf{X}\mathbf{A}_0)}{p}[\mathbf{B}_0]_{k,l} > \epsilon\right) \\
&\quad = \Pr\left(\exp\left\{t\sum_{m=1}^{n}\sum_{i,j=1}^{p} \mathbf{X}_{i,j}\Big([\mathbf{z}_m]_{(i-1)f+k}[\mathbf{z}_m]_{(j-1)f+l} - [\mathbf{A}_0]_{i,j}[\mathbf{B}_0]_{k,l}\Big)\right\} > e^{tnp\epsilon}\right) \\
&\quad \leq e^{-tnp\epsilon}\Big(\mathbb{E}\Big[\exp\left\{t\tilde{Y}^{(k,l)}\right\}\Big]\Big)^n \qquad (2.27)
\end{aligned}$$

where we used the i.i.d. property of the data in (2.27) and

$$\tilde{Y}^{(k,l)} := \sum_{i,j=1}^{p} \mathbf{X}_{i,j}([\mathbf{z}]_{(i-1)f+k}[\mathbf{z}]_{(j-1)f+l} - [\mathbf{A}_0]_{i,j}[\mathbf{B}_0]_{k,l}).$$

Define $p^2 \times 1$ random vector $\mathbf{z}^{(k,l)}$ as

$$[\mathbf{z}^{(k,l)}]_{(i-1)p+j} := [\mathbf{z}]_{(i-1)f+k}[\mathbf{z}]_{(j-1)f+l} - [\mathbf{A}_0]_{i,j}[\mathbf{B}_0]_{k,l}$$

for $1 \leq i,j \leq p$. Clearly, this random vector is zero mean. The expectation term inside the parentheses in (2.27) is the MGF of the random variable $\tilde{Y}^{(k,l)} = \mathrm{vec}(\mathbf{X})^T\mathbf{z}^{(k,l)}$. For notational simplicity, let $\tilde{\phi}_Y(t) = \mathbb{E}[e^{tY}]$ denote the MGF of a random vector Y.

Performing a second order Taylor expansion on $\tilde{\phi}_{\tilde{Y}^{(k,l)}}$ about the origin, we obtain:

$$\tilde{\phi}_{\tilde{Y}^{(k,l)}}(t) = \tilde{\phi}_{\tilde{Y}^{(k,l)}}(0) + \frac{d\tilde{\phi}_{\tilde{Y}^{(k,l)}}(0)}{dt}t + \frac{1}{2}\frac{d^2\tilde{\phi}_{\tilde{Y}^{(k,l)}}(\delta t)}{dt^2}t^2$$

for some $\delta \in [0,1]$. Trivially, $\tilde{\phi}_{\tilde{Y}^{(k,l)}}(0) = 1$ and $\frac{d\tilde{\phi}_{\tilde{Y}^{(k,l)}}(0)}{dt} = \mathbb{E}[\mathrm{vec}(\mathbf{X})^T\mathbf{z}^{(k,l)}] = 0$. Using the linearity of the expectation operator, we have:

$$\begin{aligned}
\frac{d^2\tilde{\phi}_{\tilde{Y}^{(k,l)}}(\delta t)}{dt^2} &= \mathbb{E}[(\tilde{Y}^{(k,l)})^2 e^{t\delta\tilde{Y}^{(k,l)}}] \\
&= \sum_{m=0}^{\infty}\frac{(\delta t)^m}{m!}\mathbb{E}[(\mathrm{vec}(\mathbf{X})^T\mathbf{z}^{(k,l)})^{m+2}]
\end{aligned}$$

Using the elementary inequality $1 + y \leq e^y$ for $y > -1$, and after some algebra, we have:

$$n \ln(\tilde{\phi}_{\tilde{Y}^{(k,l)}}(t)) \leq \frac{n}{2} t^2 \sum_{m=0}^{\infty} T_m(t) \tag{2.28}$$

where $T_m(t) := \frac{(t\delta)^m}{m!} \mathbb{E}[(\text{vec}(\mathbf{X})^T \mathbf{z}^{(k,l)})^{m+2}]$. Note that

$$\begin{aligned} t^2 T_m(t) &\leq \frac{t^{m+2}}{m!} \mathbb{E}\left[\left(\sum_{i,j=1}^{p} \mathbf{X}_{i,j} ([\mathbf{z}]_{(i-1)f+k} [\mathbf{z}]_{(j-1)f+l} - [\mathbf{A}_0]_{i,j} [\mathbf{B}_0]_{k,l}) \right)^{m+2} \right] \\ &= \frac{t^{m+2}}{m!} \sum_{i_1,j_1=1}^{p} \cdots \sum_{i_{m+2},j_{m+2}=1}^{p} \mathbf{X}_{i_1,j_1} \cdots \mathbf{X}_{i_{m+2},j_{m+2}} \\ &\quad \times \mathbb{E}\Big[\prod_{\alpha=1}^{m+2} \Big([\mathbf{z}]_{(i_\alpha-1)f+k} [\mathbf{z}]_{(j_\alpha-1)f+l} - [\mathbf{A}_0]_{i_\alpha,j_\alpha} [\mathbf{B}_0]_{k,l} \Big) \Big] \\ &\leq \frac{t^{m+2}}{m!} (2m+2)!! \cdot p \left(\max_k [\mathbf{B}_0]_{k,k} \|\mathbf{X}\|_2 \|\mathbf{A}_0\|_2 \right)^{m+2} \qquad (2.29) \\ &= \frac{(2m+2)!!}{m!} (t\overline{k})^{m+2} p \end{aligned}$$

where (2.29) follows from Isserlis' formula [114]. Also, we defined the absolute constant $\overline{k} = \max_k [\mathbf{B}_0]_{k,k} \|\mathbf{X}\|_2 \|\mathbf{A}_0\|_2$. Summing the result over m, and letting $u := t\overline{k} > 0$, $a_m(u) := \frac{(2m+2)!!}{m!} u^m$, $\psi(u) = \sum_{m=0}^{\infty} a_m(u)$, we obtain:

$$t^2 \sum_{m=0}^{\infty} T_m(t) \leq p u^2 \psi(u) \Big|_{u=t\overline{k}} \tag{2.30}$$

By the ratio test [7], the infinite series $\sum_{m=0}^{\infty} a_m(u)$ converges if $u < 1/2$. Using (2.30) in (2.28), and the result in (2.27), we obtain the exponential bound:

$$\Pr([\mathbf{T}(\mathbf{X})]_{k,l} - [\mathbf{B}_*]_{k,l} > \epsilon) \leq \exp\Big\{ -tnp\epsilon + \frac{np(t\overline{k})^2}{2} \psi(t\overline{k}) \Big\}$$

Let $t < \frac{1}{(2+\tau)\overline{k}}$ and $\epsilon < \frac{1}{2+\tau} \psi(\frac{1}{2+\tau}) \overline{k} < \infty$. By the monotonicity of $\psi(\cdot)$, we have:

$$\Pr([\mathbf{T}(\mathbf{X})]_{k,l} - [\mathbf{B}_*]_{k,l} > \epsilon) \leq \exp\Big\{ -tnp\epsilon + \frac{npt^2\overline{k}^2}{2} \psi(\frac{1}{2+\tau}) \Big\} \tag{2.31}$$

Optimizing (2.31) over t, we obtain $t^* = \frac{\epsilon}{\overline{k}^2\psi(\frac{1}{2+\tau})}$. Clearly, $t^* < \frac{1}{(2+\tau)\overline{k}}$. Plugging this into (2.31) and letting $C := \frac{1}{2\overline{k}^2\psi(\frac{1}{2+\tau})}$ [10], we obtain for all $\epsilon < \frac{1}{2+\tau}\psi(\frac{1}{2+\tau})\overline{k}$:

$$\Pr([\mathbf{T}(\mathbf{X})]_{k,l} - [\mathbf{B}_*]_{k,l} > \epsilon) \leq e^{-np\epsilon^2 C} \tag{2.32}$$

From (2.32) and a similar lower tail bound, we conclude that for all $\epsilon < \frac{1}{2+\tau}\psi(\frac{1}{2+\tau})\overline{k}$:

$$\Pr(|[\mathbf{T}(\mathbf{X})]_{k,l} - \mathbb{E}[[\mathbf{T}(\mathbf{X})]_{k,l}]| > \epsilon) \leq 2e^{-np\epsilon^2 C}$$

The union bound over $(k,l) \in \{1,\ldots,f\}^2$ completes the proof. This bound can be re-expressed as in the statement of Lemma II.7 (see [114] for more details). □

2.11.9 Proposition II.20

Proposition II.20. *Let $\mathbf{S}_{p,f,n}$ be a $d' \times d'$ (where $d' = p$ or $d' = f$) random matrix such that with probability $1 - \frac{2}{n^2}$, $|\mathbf{S}_{p,f,n} - \mathbf{\Sigma}_*|_\infty \leq r_{p,f,n}$. Assume $\mathbf{\Sigma}_* \in S_{++}^{d'}$ has uniformly bounded spectrum as $p, f \to \infty$ (analog to Assumption 1). Choose $\lambda_{p,f,n} = c \cdot r_{p,f,n}$ for some absolute constant $c > 0$. Consider the Glasso operator $\mathbf{G}(\cdot,\cdot)$ defined in (2.4). Let $s = s_{\mathbf{\Theta}_*}$ be the sparsity parameter associated with $\mathbf{\Theta}_* := \mathbf{\Sigma}_*^{-1}$. Assume $\sqrt{d'+s} \cdot r_{p,f,n} = o(1)$. Then, with probability $1 - \frac{2}{n^2}$,*

$$\|\mathbf{G}(\mathbf{S}_{p,f,n}, \lambda_{p,f,n}) - \mathbf{\Theta}_*\|_F \leq \frac{2\sqrt{2}(1+c)}{\lambda_{min}(\mathbf{\Sigma}_*)^2}\sqrt{d'+s} \cdot r_{p,f,n}$$

as $p, f, n \to \infty$.

Proof. The proof follows from a slight modification of Thm. 1 in [103], or Thm. 3 in [139]. This modification is due to the different $r_{p,f,n}$. □

2.11.10 Proof of Theorem II.11

Proof. As in the proof of Thm. 1 in [130], let $\mathbf{B}_* = \frac{\mathrm{tr}(\mathbf{A}_0\mathbf{A}_{init}^{-1})}{p}\mathbf{B}_0$ and $\mathbf{A}_* = (\frac{\mathrm{tr}(\mathbf{A}_0\mathbf{A}_{init}^{-1})}{p})^{-1}\mathbf{A}_0$. Note that Assumption 1 implies that $\|\mathbf{B}_*\|_2 = \Theta(1)$ and $\|\mathbf{A}_*\|_2 =$

[10] Since $\psi(\frac{1}{2+\tau})$ is finite, $C > 0$ is finite.

$\Theta(1)$ as $p, f \to \infty$. For conciseness, the statement "with probability $1 - cn^{-2}$ (where $c > 0$ is a constant independent of p, f, n)" will be abbreviated as "w.h.p."-i.e., with high probability.

For concreteness, we first present the result for $k = 2$ iterations. Then, we generalize the analysis to all finite flip-flop iterations by induction. The growth assumptions in the theorem imply

$$\max\left\{p, f, \frac{f^2}{p}, \left(\frac{\sqrt{pf} + f\sqrt{\frac{f}{p}} + p\sqrt{\frac{p}{f}}}{p+f}\right)^2\right\} \log M \leq C'n \tag{2.33}$$

for some constant $C' > 0$ large enough [11]. In fact, the growth assumption in the theorem statement can be relaxed to (2.33).

As in the proof of Thm. 1 in [130], we vectorize the operations (2.8) and (2.9):

$$\text{vec}(\hat{\mathbf{A}}(\mathbf{B})) = \frac{1}{f}\hat{\mathbf{R}}_A \text{vec}(\mathbf{B}^{-1})$$
$$\text{vec}(\hat{\mathbf{B}}(\mathbf{A})) = \frac{1}{p}\hat{\mathbf{R}}_B \text{vec}(\mathbf{A}^{-1})$$

where $\hat{\mathbf{R}}_A$ and $\hat{\mathbf{R}}_B$ are permuted versions of the sample covariance matrix [130].

Define intermediate error matrices:

$$\tilde{\mathbf{B}}^0 = \hat{\mathbf{B}}(\mathbf{A}_{init}) - \mathbf{B}_*$$
$$\tilde{\mathbf{A}}^1 = \hat{\mathbf{A}}(\hat{\mathbf{B}}(\mathbf{A}_{init})) - \mathbf{A}_*$$

Define $\mathbf{Y}_* = \mathbf{B}_*^{-1}$ and $\mathbf{X}_* = \mathbf{A}_*^{-1}$. Also, define:

$$\mathbf{Y}_1 = \hat{\mathbf{B}}(\mathbf{A}_{init})^{-1}$$
$$\mathbf{X}_2 = \hat{\mathbf{A}}(\hat{\mathbf{B}}(\mathbf{A}_{init}))^{-1}$$

These inverses exist if $n \geq \max(\frac{p}{f}, \frac{f}{p}) + 1$ (see [87]). Define the error $\tilde{\mathbf{\Sigma}}_{FF}(k) = \mathbf{\Sigma}_{FF}(k) - \mathbf{\Sigma}_0$ for $k \geq 2$. For notational simplicity, let $\mathbf{B}_0^{max} := \max_k [\mathbf{B}_0]_{k,k}$ and $\mathbf{A}_0^{max} := \max_i [\mathbf{A}_0]_{i,i}$, $\psi_\tau := \psi(\frac{1}{2+\tau})$, where $\psi(\cdot)$ is defined in Lemma II.7.

[11] This constant is independent of p, f, n, but may depend on the constants in Assumption II.10.

Lemma II.7 implies that for

$$n > \frac{8(2+\tau)^2}{\psi_\tau} \log M \tag{2.34}$$

then with probability $1 - 2n^{-2}$, we have:

$$\|\tilde{\mathbf{B}}^0\|_F \leq C_0 f p^{-1/2} \sqrt{\frac{\log M}{n}} \tag{2.35}$$

where $C_0 = 2\sqrt{2\psi_\tau}\mathbf{B}_0^{max}\|\mathbf{A}_{init}^{-1}\mathbf{A}_0\|_2$.

Let $\epsilon' > 1$. Note that from (2.35), for

$$n \geq (\epsilon' C_0)^2 f^2 p^{-1} \log M \tag{2.36}$$

with probability $1 - 2n^{-2}$,

$$\begin{aligned}\lambda_{min}(\hat{\mathbf{B}}(\mathbf{A}_{init})) &= \lambda_{min}(\tilde{\mathbf{B}}^0 + \mathbf{B}_*) \geq \lambda_{min}(\mathbf{B}_*) - \|\tilde{\mathbf{B}}^0\|_2 \\ &\geq \lambda_{min}(\mathbf{B}_*) - \|\tilde{\mathbf{B}}^0\|_F \geq \left(1 - \frac{1}{\epsilon'}\right)\lambda_{min}(\mathbf{B}_*) > 0\end{aligned}$$

Thus, letting $\boldsymbol{\Delta}_Y^1 = \mathbf{Y}_1 - \mathbf{Y}_*$, w.h.p.,

$$\begin{aligned}\|\boldsymbol{\Delta}_Y^1\|_F &= \|\mathbf{Y}_1(\hat{\mathbf{B}}(\mathbf{A}_{init}) - \mathbf{B}_*)\mathbf{Y}_*\|_F \\ &\leq \|\mathbf{Y}_1\|_2\|\mathbf{Y}_*\|_2\|\tilde{\mathbf{B}}^0\|_F = \frac{\|\tilde{\mathbf{B}}^0\|_F}{\lambda_{min}(\mathbf{B}_*)\lambda_{min}(\hat{\mathbf{B}}(\mathbf{A}_{init}))} \\ &\leq C_0\left(1 - \frac{1}{\epsilon'}\right)^{-1}\|\mathbf{Y}_*\|_2^2 f p^{-1/2}\sqrt{\frac{\log M}{n}} \qquad (2.37)\end{aligned}$$

Expanding $\tilde{\mathbf{A}}^1$:

$$\begin{aligned}\mathrm{vec}(\tilde{\mathbf{A}}^1) &= \frac{1}{f}\hat{\mathbf{R}}_A \mathrm{vec}(\mathbf{Y}_1) - \mathrm{vec}(\mathbf{A}_*) \\ &= \frac{\mathrm{tr}(\mathbf{B}_0\boldsymbol{\Delta}_Y^1)}{f}\mathrm{vec}(\mathbf{A}_0) + \mathrm{vec}(\hat{\mathbf{A}}(\mathbf{B}_*) - \mathbf{A}_*) \\ &\quad + \frac{1}{f}\tilde{\mathbf{R}}_A \mathrm{vec}(\boldsymbol{\Delta}_Y^1) \qquad (2.38)\end{aligned}$$

where we used $\mathbf{R}_A = \text{vec}(\mathbf{A}_0)\text{vec}(\mathbf{B}_0^T)^T$ (see Eq. (91) from [130]). Using the triangle inequality in (2.38), the Cauchy-Schwarz inequality, and standard matrix norm bounds:

$$\|\tilde{\mathbf{A}}^1\|_F \leq \underbrace{\sqrt{\frac{p}{f}}\|\boldsymbol{\Sigma}_0\|_2\|\boldsymbol{\Delta}_Y^1\|_F}_{T_1} + \underbrace{p|\hat{\mathbf{A}}(\mathbf{B}_*) - \mathbf{A}_*|_\infty}_{T_2}$$
$$+ \underbrace{\frac{p}{f}\|\tilde{\mathbf{R}}_A\text{vec}(\boldsymbol{\Delta}_Y^1)\|_\infty}_{T_3}$$

We note upon expanding:

$$\frac{1}{f}\|\tilde{\mathbf{R}}_A\text{vec}(\boldsymbol{\Delta}_Y^1)\|_\infty = \left|\frac{1}{f}\sum_{k,l=1}^{f}[\boldsymbol{\Delta}_Y^1]_{k,l}\bar{\hat{\mathbf{S}}}_n(k,l) - \frac{\text{tr}(\mathbf{B}_0\boldsymbol{\Delta}_Y^1)}{f}\mathbf{A}_0\right|_\infty$$

From (2.37), there exists $c > 0$ such that:

$$\mathbb{P}\left(T_1 \geq C_1 f^{1/2}\sqrt{\frac{\log M}{n}}\right) \leq cn^{-2}$$

where $C_1 = \|\boldsymbol{\Sigma}_0\|_2 C_0(1-1/\epsilon')^{-1}\|\mathbf{Y}_*\|_2^2$ is an absolute constant. Lemma II.7 implies:

$$\mathbb{P}\left(T_2 \geq C_2 f^{-1/2}\sqrt{\frac{\log M}{n}}\right) \leq 2n^{-2}$$

where $C_2 = 2\sqrt{2\psi_\tau}A_0^{max}\|\mathbf{Y}_*\mathbf{B}_0\|_2$ is an absolute constant. To bound T_3, we define the following events:

$$E_0 = \left\{\|\boldsymbol{\Delta}_Y^1\|_F \leq \frac{C_1}{\|\boldsymbol{\Sigma}_0\|_2} f p^{-1/2}\sqrt{\frac{\log M}{n}}\right\}$$
$$E_1 = \left\{\left|\frac{1}{f}\sum_{k,l=1}^{f}[\boldsymbol{\Delta}_Y^1]_{k,l}\bar{\hat{\mathbf{S}}}_n(k,l) - \frac{\text{tr}(\mathbf{B}_0\boldsymbol{\Delta}_Y^1)}{f}\mathbf{A}_0\right|_\infty \leq 2\sqrt{2\psi_\tau}A_0^{max}\|\boldsymbol{\Delta}_Y^1\|_F\|\mathbf{B}_0\|_2\sqrt{\frac{\log M}{nf}}\right\}$$
$$E_2 = \left\{T_3 \leq C_3\sqrt{pf}\sqrt{\frac{\log M}{n}}\right\}$$

where $C_3 = 2\sqrt{2\psi_\tau}A_0^{max}\|\mathbf{B}_0\|_2 C_0(1-1/\epsilon')^{-1}\|\mathbf{Y}_*\|_2^2$ is an absolute constant. From (2.37), it follows that $\mathbb{P}(E_0) \geq 1 - cn^{-2}$ and from Lemma (II.7), it follows that

$\mathbb{P}\left(E_1|E_0\right) \geq 1 - 2n^{-2}$. As a result, we have $\mathbb{P}(E_2) \geq \mathbb{P}(E_1 \cap E_0) = \mathbb{P}(E_1|E_0)\mathbb{P}(E_0) \geq 1 - (c+2)n^{-2}$. Putting it together with the union bound, we have:

$$\begin{aligned}
&\mathbb{P}\left(\|\tilde{\mathbf{A}}^1\|_F \geq (C_1 f^{1/2} + C_2 p f^{-1/2})\sqrt{\frac{\log M}{n}} + C_3\sqrt{pf}\frac{\log M}{n} \right) \\
&\quad \leq \mathbb{P}\left(T_1 \geq \frac{C_1}{3} f^{1/2}\sqrt{\frac{\log M}{n}} \right) + \mathbb{P}\left(T_2 \geq \frac{C_2}{3} p f^{-1/2}\sqrt{\frac{\log M}{n}} \right) \\
&\qquad + \mathbb{P}\left(T_3 \geq \frac{C_3}{3}\sqrt{pf}\frac{\log M}{n} \right) \\
&\quad \leq c' n^{-2} \qquad (2.39)
\end{aligned}$$

for some $c' > 0$ absolute constant.

Let $c_1 > 0$. For

$$n \geq \left(\frac{C_3}{c_1 \max(C_1, C_2)} \right)^2 \frac{pf}{(f^{1/2} + pf^{-1/2})^2} \log M \qquad (2.40)$$

then, from (2.39), we have w.h.p.,

$$\|\tilde{\mathbf{A}}^1\|_F \leq \max(C_1, C_2)(1 + c_1)(\sqrt{f} + pf^{-1/2})\sqrt{\frac{\log M}{n}} \qquad (2.41)$$

Using properties of the Kronecker product:

$$\begin{aligned}
\tilde{\mathbf{\Sigma}}_{FF}(2) &= \tilde{\mathbf{A}}^1 \otimes \mathbf{B}_* + \mathbf{A}_* \otimes \tilde{\mathbf{B}}^0 \\
&\quad + \tilde{\mathbf{A}}^1 \otimes \tilde{\mathbf{B}}^0 \qquad (2.42)
\end{aligned}$$

From (2.35),(2.41), (2.42), under conditions (2.34),(2.36), and (2.40), w.h.p.,

$$\begin{aligned}
\|\tilde{\mathbf{\Sigma}}_{FF}(2)\|_F &\leq \|\tilde{\mathbf{A}}_1\|_F \|\mathbf{B}_*\|_F \\
&\quad + \|\mathbf{A}_*\|_F \|\tilde{\mathbf{B}}^0\|_F + \|\tilde{\mathbf{A}}^1\|_F \|\tilde{\mathbf{B}}^0\|_F \\
&\leq \tilde{C}_3(p + 2f)\sqrt{\frac{\log M}{n}} + \tilde{C}_4(f\sqrt{f/p} + \sqrt{pf})\frac{\log M}{n} \qquad (2.43)
\end{aligned}$$

where $\tilde{C}_3 = \max(\|\mathbf{B}_*\|_2 \max(C_1, C_2)(1 + c_1), C_0\|\mathbf{A}_*\|_2)$ and $\tilde{C}_4 = C_0 \max(C_1, C_2)(1 + c_1)$ are constants [114].

Let $c_2 > 0$. For

$$n \geq (\frac{\tilde{C}_4}{\tilde{C}_3 c_2})^2 \frac{(f\sqrt{f/p} + \sqrt{pf})^2}{(p+2f)^2} \log M$$

then, from (2.43) w.h.p.,

$$\|\tilde{\mathbf{\Sigma}}_{FF}(2)\|_F \leq \tilde{C}_3(1+c_2)(p+2f)\sqrt{\frac{\log M}{n}}$$

The proof for $k = 2$ iterations is complete. Using a simple induction, it follows that the rate (2.16) holds for all k finite.

Next, we show that the convergence rate in the precision matrix Frobenius error is on the same order as the covariance matrix error. Let $\mathbf{\Theta}_{FF}(2) := \mathbf{\Sigma}_{FF}(2)^{-1}$. From (2.41), for

$$n > (\epsilon' \|\mathbf{X}_*\|_2 \max(C_1, C_2)(1+c_1))^2 (\sqrt{f} + pf^{-1/2})^2 \log M$$

then, letting $\mathbf{\Delta}_X^2 = \mathbf{X}_2 - \mathbf{X}_*$, we have w.h.p.,

$$\begin{aligned} \|\mathbf{\Delta}_X^2\|_F \leq \left(1 - \frac{1}{\epsilon'}\right)^{-1} \|\mathbf{X}_*\|_2^2 \tilde{C}_1(1+c_1) \\ \times (\sqrt{f} + pf^{-1/2})\sqrt{\frac{\log M}{n}} \end{aligned} \tag{2.44}$$

Using (2.37) and (2.44), we have w.h.p.,

$$\begin{aligned} \|\mathbf{\Theta}_{FF}(2) - \mathbf{\Theta}_0\|_F &\leq \|\mathbf{\Delta}_X^2\|_F \|\mathbf{Y}_*\|_F \\ &+ \|\mathbf{\Delta}_Y^1\|_F \|\mathbf{X}_*\|_F + \|\mathbf{\Delta}_X^2\|_F \|\mathbf{\Delta}_Y^1\|_F \\ &\leq \tilde{D}_1(2f+p)\sqrt{\frac{\log M}{n}} + \tilde{D}_2(f\sqrt{\frac{f}{p}} + \sqrt{pf})\frac{\log M}{n} \end{aligned} \tag{2.45}$$

where $\tilde{D}_1$ and $\tilde{D}_2$ are constants.

For

$$n > (\frac{\tilde{D}_2}{\tilde{D}_1 d'})^2 (\frac{f\sqrt{f/p} + \sqrt{pf}}{2f+p})^2 \log M$$

the bound (2.45) becomes w.h.p.,

$$\|\mathbf{\Theta}_{FF}(2) - \mathbf{\Theta}_0\|_F \leq \tilde{D}_1(1+d')(2f+p)\sqrt{\frac{\log M}{n}}$$

Thus, the same rate $O_P\left(\sqrt{\frac{(p^2+f^2)\log M}{n}}\right)$ holds for the precision matrix Frobenius error.

□

2.11.11 Proof of Theorem II.13

Proof. We show that the first iteration of the KGL algorithm yields a fast statistical convergence rate of $O_P\left(\sqrt{\frac{(p+f)\log M}{n}}\right)$ by appropriately adjusting the regularization parameters. A simple induction finishes the proof. Adopt the notation from the proof of Thm. II.11.

Lemma II.7 implies that for

$$n \geq \frac{8(2+\tau)^2}{\psi_\tau}\log M \tag{2.46}$$

then with probability $1 - 2n^{-2}$,

$$|\tilde{\mathbf{B}}^0|_\infty \leq C_0 p^{-1/2}\sqrt{\frac{\log M}{n}} \tag{2.47}$$

where $\tilde{\mathbf{B}}^0 = \hat{\mathbf{B}}(\mathbf{A}_{init}) - \mathbf{B}_*$. From Proposition II.20 and (2.47), we obtain w.h.p.,

$$\begin{aligned}\|\mathbf{Y}_1 - \mathbf{Y}_*\|_F &\leq 2\sqrt{2}(1+c_y)\sqrt{1+c_{Y_0}}\|\mathbf{Y}_*\|_2^2 \\ &\times C_0\sqrt{\frac{f\log M}{np}}\end{aligned} \tag{2.48}$$

where we also used $s_{Y_0} \leq c_{Y_0} f$ and $\mathbf{Y}_1 := \mathbf{G}(\hat{\mathbf{B}}(\mathbf{A}_{init}), \lambda_Y^{(1)}) = \mathbf{B}_1^{-1}$. Note that $fp^{-1}\log M = o(n)$ was used here. Let $\mathbf{\Delta}_Y^1 = \mathbf{Y}_1 - \mathbf{Y}_*$.

Let $\grave{\mathbf{A}}^1 := \hat{\mathbf{A}}(\mathbf{B}_1) - \mathbf{A}_*$. Then, we have

$$\begin{aligned}\mathrm{vec}(\grave{\mathbf{A}}^1) &= \frac{1}{f}\hat{\mathbf{R}}_A \mathrm{vec}(\mathbf{Y}_1) - \mathrm{vec}(\mathbf{A}_*) \\ &= \frac{\mathrm{tr}(\mathbf{B}_0 \boldsymbol{\Delta}_Y^1)}{f}\mathrm{vec}(\mathbf{A}_0) + \mathrm{vec}(\hat{\mathbf{A}}(\mathbf{B}_*) - \mathbf{A}_*) \\ &\quad + \frac{1}{f}\tilde{\mathbf{R}}_A \mathrm{vec}(\boldsymbol{\Delta}_Y^1) \qquad (2.49)\end{aligned}$$

where we used $\mathbf{R}_A = \mathrm{vec}(\mathbf{A}_0)\mathrm{vec}(\mathbf{B}_0^T)^T$ (see Eq. (91) in [130]).

From (2.49), applying the triangle inequality and using the Cauchy-Schwarz inequality:

$$\begin{aligned}|\grave{\mathbf{A}}^1|_\infty \leq \underbrace{\frac{\sqrt{f}\|\mathbf{B}_0\|_2 \|\boldsymbol{\Delta}_Y^1\|_F}{f}|\mathbf{A}_0|_\infty}_{T_1} + \underbrace{|\hat{\mathbf{A}}(\mathbf{B}_*) - \mathbf{A}_*|_\infty}_{T_2} \\ + \underbrace{\frac{1}{f}\|\tilde{\mathbf{R}}_A \mathrm{vec}(\boldsymbol{\Delta}_Y^1)\|_\infty}_{T_3} \qquad (2.50)\end{aligned}$$

(2.51)

Let $\tilde{C}_0 = C_0 2\sqrt{2}(1+c_y)\sqrt{1+c_{Y_0}}\|\mathbf{Y}_*\|_2^2$ and $\bar{C}_1 = \tilde{C}_0|\mathbf{A}_0|_\infty\|\mathbf{B}_0\|_2$. The bound (2.48) implies

$$\mathbb{P}\left(T_1 \geq \bar{C}_1\sqrt{\frac{\log M}{np}}\right) \leq cn^{-2}$$

for some $c > 0$. Let $\bar{C}_2 = 2\sqrt{2\psi_\tau}A_0^{max}\|\mathbf{Y}_*\mathbf{B}_0\|_2$. Lemma II.7 implies

$$\mathbb{P}\left(T_2 \geq \bar{C}_2\sqrt{\frac{\log M}{nf}}\right) \leq 2n^{-2}$$

Let $\bar{C}_3 = \tilde{C}_0 2\sqrt{2\psi_\tau}A_0^{max}\|\mathbf{B}_0\|_2$. To bound T_3, we use the same technique as in the proof of Thm. II.11. Define the events:

$$\begin{aligned}E_0 &= \left\{\|\boldsymbol{\Delta}_Y^1\|_F \leq \tilde{C}_0\sqrt{\frac{f\log M}{np}}\right\} \\ E_1 &= \left\{\frac{1}{f}\|\tilde{\mathbf{R}}_A\mathrm{vec}(\boldsymbol{\Delta}_Y^1)\|_\infty \leq 2\sqrt{2\psi_\tau}A_0^{max}\|\mathbf{B}_0\|_2\|\boldsymbol{\Delta}_Y^1\|_F\sqrt{\frac{\log M}{nf}}\right\} \\ E_2 &= \left\{T_3 \leq \bar{C}_3\frac{1}{\sqrt{p}}\frac{\log M}{n}\right\}\end{aligned}$$

From (2.48), we have $\mathbb{P}(E_0) \geq 1 - cn^{-2}$ and from Lemma II.7 we have $\mathbb{P}(E_1|E_0) \geq 1 - 2n^{-2}$. Thus, $\mathbb{P}(E_2) \geq \mathbb{P}(E_1|E_0)\mathbb{P}(E_0) \geq 1 - c'n^{-2}$.

Using (2.50) and the union bound:

$$\begin{aligned}
&\mathbb{P}\left(|\grave{\mathbf{A}}^1|_\infty \geq (\frac{\bar{C}_1}{\sqrt{p}} + \frac{\bar{C}_2}{\sqrt{f}})\sqrt{\frac{\log M}{n}} + \frac{\bar{C}_3}{\sqrt{p}}\frac{\log M}{n}\right) \\
&\leq \mathbb{P}\left(T_1 \geq \frac{\bar{C}_1}{3\sqrt{p}}\sqrt{\frac{\log M}{n}}\right) + \mathbb{P}\left(T_2 \geq \frac{\bar{C}_2}{3\sqrt{f}}\sqrt{\frac{\log M}{n}}\right) \\
&\quad + \mathbb{P}\left(T_3 \geq \frac{\bar{C}_3}{3\sqrt{p}}\frac{\log M}{n}\right) \\
&\leq c'' n^{-2}
\end{aligned}$$

for some $c'' > 0$. Thus, for $n \geq (\frac{\bar{C}_3}{\bar{C}_1 c_1})^2 \log M$, $c_1 > 0$, we have w.h.p.,

$$|\grave{\mathbf{A}}^1|_\infty \leq \max(\bar{C}_1, \bar{C}_2)(1 + c_1)\left(\frac{1}{\sqrt{p}} + \frac{1}{\sqrt{f}}\right)\sqrt{\frac{\log M}{n}} \tag{2.52}$$

Let $\mathbf{\Delta}_X^1 = \mathbf{X}_1 - \mathbf{X}_*$. From Proposition II.20 and (2.52), we obtain w.h.p.:

$$\begin{aligned}
\|\mathbf{\Delta}_X^1\|_F &\leq 2\sqrt{2}(1 + c_x)\sqrt{1 + c_{X_0}}\|\mathbf{X}_*\|_2^2 \max(\bar{C}_1, \bar{C}_2)(1 + c_1) \\
&\quad \times \left(1 + \sqrt{\frac{p}{f}}\right)\sqrt{\frac{\log M}{n}}
\end{aligned} \tag{2.53}$$

where we used $s_{X_0} \leq c_{X_0} p$ and $\mathbf{X}_1 := \mathbf{G}(\hat{\mathbf{A}}(\mathbf{B}_1), \lambda_X^{(1)})$, $\mathbf{X}_* := \mathbf{A}_*^{-1}$. Note that $(1 + \sqrt{p/f})^2 \log M = o(n)$ was used here.

Finally, using (2.48) and (2.53), we obtain w.h.p.:

$$\begin{aligned}
&\|\mathbf{\Theta}_{KGL}(2) - \mathbf{\Theta}_0\|_F = \|\mathbf{X}_1 \otimes \mathbf{Y}_1 - \mathbf{X}_* \otimes \mathbf{Y}_*\|_F \\
&\leq \|\mathbf{\Delta}_Y^1\|_F \sqrt{p}\|\mathbf{X}_*\|_2 + \|\mathbf{\Delta}_X^1\|_F \sqrt{f}\|\mathbf{Y}_*\|_2 \\
&\quad + \|\mathbf{\Delta}_Y^1\|_F \|\mathbf{\Delta}_X^1\|_F \\
&\leq \bar{C}'_3(2\sqrt{f} + \sqrt{p})\sqrt{\frac{\log M}{n}} + \bar{C}'_4(1 + \sqrt{\frac{f}{p}})\frac{\log M}{n}
\end{aligned} \tag{2.54}$$

where $\bar{C}'_3$ and $\bar{C}'_4$ are constants [114]. For

$$n > (\frac{\bar{C}'_4}{\bar{C}'_3\bar{c}})^2 \left(\frac{1+\sqrt{f/p}}{2\sqrt{f}+\sqrt{p}}\right)^2 \log M$$

the bound (2.54) further becomes:

$$\|\mathbf{\Theta}_{KGL}(2) - \mathbf{\Theta}_0\|_F \leq \bar{C}'_3(1+\bar{c})(2\sqrt{f}+\sqrt{p})\sqrt{\frac{\log M}{n}}$$

Note that $\|\mathbf{\Theta}_{KGL}(2)-\mathbf{\Theta}_0\|_F^2 = O_P\left(\frac{(p+f+\sqrt{pf})\log M}{n}\right) = O_P\left(\frac{(p+f)\log M}{n}\right)$ as $p, f, n \to \infty$. This concludes the first part of the proof. The rest of the proof follows by similar bounding arguments coupled with induction. The rate remains the same as the number of iterations increases, but the constant on front may change.

Next, we show that the convergence rate in the covariance matrix Frobenius error is on the same order as the inverse. From (2.48), for

$$n > (\epsilon'\tilde{C}_0\|\mathbf{Y}_*\|_2)^2 f p^{-1} \log M$$

we have w.h.p. $\lambda_{min}(\mathbf{Y}_1) \geq \lambda_{min}(\mathbf{Y}_*) - \|\mathbf{Y}_1 - \mathbf{Y}_*\|_F \geq (1-\frac{1}{\epsilon'})\lambda_{min}(\mathbf{Y}_*)$, which in turn implies w.h.p.,

$$\|\mathbf{\Delta}_B^1\|_F = \|\mathbf{B}_1 - \mathbf{B}_*\|_F \leq \underbrace{(1-1/\epsilon')^{-1}\tilde{C}_0\|\mathbf{B}_*\|_2^2}_{\bar{C}_B^1}$$

$$\times \sqrt{\frac{f}{p}}\sqrt{\frac{\log M}{n}} \tag{2.55}$$

[12] Using a similar argument, from (2.53), for $n \geq C'(1+\sqrt{\frac{p}{f}})^2 \log M$ (for some constant C') we have w.h.p.,

$$\|\mathbf{\Delta}_A^1\|_F = \|\mathbf{A}_1 - \mathbf{A}_*\|_F \leq \underbrace{(1-1/\epsilon')^{-1}\|\mathbf{A}_*\|_2^2\bar{C}_X^1}_{\bar{C}_A^1}$$

$$\times \left(1+\sqrt{\frac{p}{f}}\right)\sqrt{\frac{\log M}{n}} \tag{2.56}$$

[12] Here, $\mathbf{B}_1 = \mathbf{Y}_1^{-1}$ exists since $\mathbf{Y}_1$ is positive definite (see (2.4)).

where $\mathbf{A}_1 = \mathbf{X}_1^{-1}$.

Let $\mathbf{\Sigma}_{KGL}(2) := \mathbf{\Theta}_{KGL}(2)^{-1} = \mathbf{A}_1 \otimes \mathbf{B}_1$. Then, w.h.p.,

$$\begin{aligned} \|\mathbf{\Sigma}_{KGL}(2) - \mathbf{\Sigma}_0\|_F &\leq \|\mathbf{\Delta}_A^1\|_F \|\mathbf{B}_*\|_F \\ &\quad + \|\mathbf{\Delta}_B^1\|_F \|\mathbf{A}_*\|_F + \|\mathbf{\Delta}_A^1\|_F \|\mathbf{\Delta}_B^1\|_F \\ &\leq \bar{D}_1(2\sqrt{f} + \sqrt{p})\sqrt{\frac{\log M}{n}} + \bar{D}_2(1 + \sqrt{\frac{f}{p}})\frac{\log M}{n} \end{aligned} \tag{2.57}$$

where $\bar{D}_1$ and $\bar{D}_2$ are constants [114]. For

$$n > (\frac{\bar{D}_2}{\bar{D}_1 d})^2 \left(\frac{1 + \sqrt{\frac{f}{p}}}{2\sqrt{f} + \sqrt{p}} \right)^2 \log M$$

then (2.57) implies w.h.p.,

$$\|\mathbf{\Sigma}_{KGL}(2) - \mathbf{\Sigma}_0\|_F \leq \bar{D}_1(1 + d)(2\sqrt{f} + \sqrt{p})\sqrt{\frac{\log M}{n}}$$

Thus, the same rate $O_P\left(\sqrt{\frac{(p+f)\log M}{n}}\right)$ holds for the error in the covariance matrix.

□

CHAPTER III

Kronecker PCA: A Series Decomposition of Covariance Matrices using Permuted Rank-Penalized Least Squares

This chapter presents a new method for estimating high dimensional covariance matrices. The method, permuted rank-penalized least-squares (PRLS), is based on a Kronecker product series expansion of the true covariance matrix. Assuming an i.i.d. Gaussian random sample, we establish high dimensional rates of convergence to the true covariance as both the number of samples and the number of variables go to infinity. For covariance matrices of low separation rank, our results establish that PRLS has significantly faster convergence than the standard sample covariance matrix (SCM) estimator. The convergence rate captures a fundamental tradeoff between estimation error and approximation error, thus providing a scalable covariance estimation framework in terms of separation rank, similar to low rank approximation of covariance matrices [85]. The MSE convergence rates generalize the high dimensional rates recently obtained for the ML Flip-flop algorithm [116, 114] for Kronecker product covariance estimation. We show that a class of block Toeplitz covariance matrices is approximatable by low separation rank and give bounds on the minimal separation rank r that ensures a given level of bias. Simulations are presented to validate the theoretical bounds. As a real world application, we illustrate the utility of the proposed Kronecker covariance estimator for spatio-temporal linear least

squares prediction of multivariate wind speed measurements.

3.1 Introduction

Covariance estimation is a fundamental problem in multivariate statistical analysis. It has received attention in diverse fields including economics and financial time series analysis (e.g., portfolio selection, risk management and asset pricing [4]), bioinformatics (e.g. gene microarray data [133, 63], functional MRI [44]) and machine learning (e.g., face recognition [137], recommendation systems [2]). In many modern applications, data sets are very large with both large number of samples n and large dimension d, often with $d \gg n$, leading to a number of covariance parameters that greatly exceeds the number of observations. The search for good low-dimensional representations of these data sets has led to much progress in their analysis. Recent examples include sparse covariance estimation [136, 5, 99, 103], low rank covariance estimation [50, 51, 74, 85], and Kronecker product esimation [49, 130, 38, 116, 114].

Kronecker product (KP) structure is a different covariance constraint from sparse or low rank constraints. KP represents a $pq \times pq$ covariance matrix $\mathbf{\Sigma}_0$ as the Kronecker product of two lower dimensional covariance matrices. When the variables are multivariate Gaussian with covariance following the KP model, the variables are said to follow a matrix normal distribution [38, 49, 60]. This model has applications in channel modeling for MIMO wireless communications [131], geostatistics [36], genomics [134], multi-task learning [16], face recognition [137], recommendation systems [2] and collaborative filtering [135]. The main difficulty in maximum likelihood estimation of structured covariances is the nonconvex optimization problem that arises. Thus, an alternating optimization approach is usually adopted. In the case where there is no missing data, an extension of the alternating optimization

algorithm of Werner *et al* [130], that the authors called the flip flop (FF) algorithm, can be applied to estimate the parameters of the Kronecker product model, called KGlasso in [116].

In this chapter, we assume that the covariance can be represented as a sum of Kronecker products of two lower dimensional factor matrices, where the number of terms in the summation may depend on the factor dimensions. More concretely, we assume that there are $d = pq$ variables whose covariance $\mathbf{\Sigma}_0$ has Kronecker product representation:

$$\mathbf{\Sigma}_0 = \sum_{\gamma=1}^{r} \mathbf{A}_{0,\gamma} \otimes \mathbf{B}_{0,\gamma} \tag{3.1}$$

where $\{\mathbf{A}_{0,\gamma}\}$ are $p \times p$ linearly independent matrices and $\{\mathbf{B}_{0,\gamma}\}$ are $q \times q$ linearly independent matrices [1]. We assume that the factor dimensions p, q are known. We note $1 \leq r \leq r_0 = \min(p^2, q^2)$ and refer to r as the *separation rank*. The model (3.1) is analogous to separable approximation of continuous functions [12]. It is evocative of a type of low rank principal component decomposition where the components are Kronecker products. However, the components in (3.1) are neither orthogonal nor normalized. The model (3.1) with separation rank 1 is relevant to channel modeling for MIMO wireless communications, where $\mathbf{A}_0$ is a transmit covariance matrix and $\mathbf{B}_0$ is a receive covariance matrix [131]. The rank 1 model is also relevant to other transposable models arising in recommendation systems like NetFlix and in gene expression analysis [2]. The model (3.1) with $r \geq 1$ has applications in spatiotemporal MEG/EEG covariance modeling [41, 40, 15, 75], SAR data analysis [105] and other multimodal data. Due to the spatiotemporal character of certain data sets, one expects the covariance matrix to be better represented by a low separation rank model

[1] Linear independence is with respect to the trace inner product defined in the space of symmetric matrices. We note that the matrices $\{\mathbf{A}_{0,\gamma}\}, \{\mathbf{B}_{0,\gamma}\}$ need not be positive semi-definite (psd), but the sum (3.1) must be as it is a covariance matrix.

of the form (3.1) than a low algebraic rank model (i.e., a PCA decomposition). We finally note that Van Loan and Pitsianis [84] have shown that any $pq \times pq$ matrix $\mathbf{\Sigma}_0$ can be written as an orthogonal expansion of Kronecker products of the form (3.1), thus allowing any covariance matrix to be approximated by a bilinear decomposition of this form. The Kronecker product can also be represented as a multi-way tensor.

The main contribution of this chapter is a convex optimization approach to estimating covariance matrices with KP structure of the form (3.1) and the derivation of tight high-dimensional MSE convergence rates as n, p and q go to infinity. We call our method the Permuted Rank-penalized Least Squares (PRLS) estimator. Similarly to other studies of high dimensional covariance estimation [23, 116, 103, 14, 124], we analyze the estimator convergence rate in Frobenius norm of PRLS, providing specific convergence rates holding with certain high probability. In other words, our anlaysis provides high probability guarantees up to absolute constants in all sample sizes and dimensions.

For estimating separation rank r covariance matrices of the form (3.1), we establish that PRLS achieves high dimensional consistency with a convergence rate of $O_P\left(\frac{r(p^2+q^2+\log\max(p,q,n))}{n}\right)$. This can be significantly faster than the convergence rate $O_P\left(\frac{p^2q^2}{n}\right)$ of the standard sample covariance matrix (SCM). For separation rank $r = 1$ this rate is identical to that of the FF algorithm, which fits the sample covariance matrix to a single Kronecker factor.

The PRLS method for estimating the Kronecker product expansion (3.1) generalizes previously proposed Kronecker product covariance models [49, 38] to the case of $r > 1$. This is a fundamentally different generalization than the $r = 1$ sparse KP models proposed in [2, 116, 114, 83]. Independently in [116, 114] and [83], it was established that the high dimensional convergence rate for these sparse KP models

is of order $O_P\left(\frac{(p+q)\log\max(p,q,n)}{n}\right)$. While we do not pursue the the additional constraint of sparsity in this chapter, we speculate that sparsity can be combined with the Kronecker sum model (3.1), achieving even better convergence.

Advantages of the proposed PRLS covariance estimator is illustrated on both simulated and real data. The application of PRLS to the NCEP wind dataset shows that a low order Kronecker sum provides a remarkably good fit to the spatio-temporal sample covariance matrix: over 86% of all the energy is contained in the first Kronecker component of the Kronecker expansion as compared to only 41% in the principal component of the standard PCA eigen-expansion. Furthermore, by replacing the SCM in the standard linear predictor by our Kronecker sum estimator we demonstrate a 1.9 dB RMSE advantage for predicting next-day wind speeds from NCEP network past measurements.

3.2 Notation

For a square matrix $\mathbf{M}$, define $|\mathbf{M}|_1 = \|\mathrm{vec}(\mathbf{M})\|_1$ and $|\mathbf{M}|_\infty = \|\mathrm{vec}(\mathbf{M})\|_\infty$, where $\mathrm{vec}(\mathbf{M})$ denotes the vectorized form of $\mathbf{M}$ (concatenation of columns into a vector). $\|\mathbf{M}\|_2$ is the spectral norm of $\mathbf{M}$. $\mathbf{M}_{i,j}$ and $[\mathbf{M}]_{i,j}$ are the (i,j)th element of $\mathbf{M}$. Let the inverse transformation (from a vector to a matrix) be defined as: $\mathrm{vec}^{-1}(\mathbf{x}) = \mathbf{X}$, where $\mathbf{x} = \mathrm{vec}(\mathbf{X})$. Define the $pq \times pq$ permutation operator $\mathbf{K}_{p,q}$ such that $\mathbf{K}_{p,q}\mathrm{vec}(\mathbf{N}) = \mathrm{vec}(\mathbf{N}^T)$ for any $p \times q$ matrix $\mathbf{N}$. For a symmetric matrix $\mathbf{M}$, $\lambda(\mathbf{M})$ will denote the vector of real eigenvalues of $\mathbf{M}$ and define $\lambda_{max}(\mathbf{M}) = \|\mathbf{M}\|_2 = \max\lambda_i(\mathbf{M})$ for p.d. symmetric matrix, and $\lambda_{min}(\mathbf{M}) = \min\lambda_i(\mathbf{M})$. For any matrix $\mathbf{M}$, define the nuclear norm $\|\mathbf{M}\|_* = \sum_{l=1}^{r_M}|\sigma_l(\mathbf{M})|$, where $r_M = \mathrm{rank}(\mathbf{M})$ and $\sigma_l(\mathbf{M})$ is the lth singular value of $\mathbf{M}$.

For a matrix $\mathbf{M}$ of size $pq \times pq$, let $\{\mathbf{M}(i,j)\}_{i,j=1}^{p}$ denote its $q \times q$ block sub-

matrices, where each block submatrix is $\mathbf{M}(i,j) = [\mathbf{M}]_{(i-1)q+1:iq,(j-1)q+1:jq}$. Also let $\{\overline{\mathbf{M}}(k,l)\}_{k,l=1}^{q}$ denote the $p \times p$ block submatrices of the permuted matrix $\overline{\mathbf{M}} = \mathbf{K}_{p,q}^T \mathbf{M} \mathbf{K}_{p,q}$. Define the permutation operator $\mathcal{R} : \mathbb{R}^{pq \times pq} \to \mathbb{R}^{p^2 \times q^2}$ by setting the $(i-1)p+j$ row of $\mathcal{R}(\mathbf{M})$ equal to $\text{vec}(\mathbf{M}(i,j))^T$. An illustration of this permutation operator is shown in Fig. 3.1.

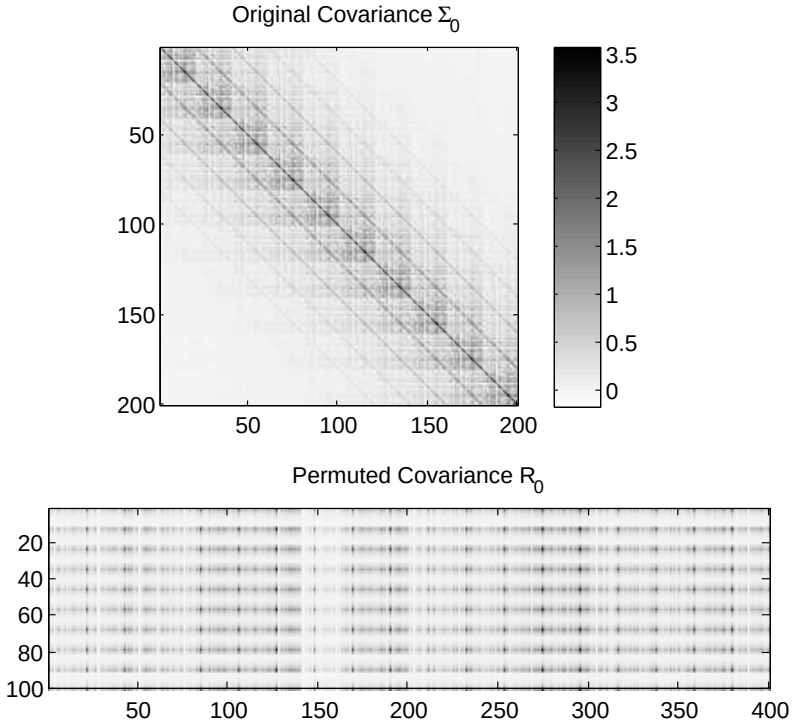


Figure 3.1: Original (top) and permuted covariance (bottom) matrix. The original covariance is $\mathbf{\Sigma}_0 = \mathbf{A}_0 \otimes \mathbf{B}_0$, where $\mathbf{A}_0$ is a 10×10 Toeplitz matrix and $\mathbf{B}_0$ is a 20×20 unstructured p.d. matrix. Note that the permutation operator $\mathcal{R}$ maps a symmetric p.s.d. matrix $\mathbf{\Sigma}_0$ to a non-symmetric rank 1 matrix $\mathbf{R}_0 = \mathcal{R}(\mathbf{\Sigma}_0)$.

Define the set of symmetric matrices $S^p = \{\mathbf{A} \in \mathbb{R}^{p \times p} : \mathbf{A} = \mathbf{A}^T\}$, the set of symmetric positive semidefinite (psd) matrices S_+^p, and the set of symmetric positive definite (pd) matrices S_{++}^p. $\mathbf{I}_d$ is a $d \times d$ identity matrix. It can be shown that S_{++}^p is a convex set, but is not closed [18]. Note that S_{++}^p is simply the interior of the closed convex cone S_+^p.

For a subspace U, define $\mathbf{P}_U$ and $\mathbf{P}_U^{\perp}$ as the orthogonal projection onto U and

$U^{\perp}$, respectively. The unit Euclidean sphere in $\mathbb{R}^{d'}$ is denoted by $\mathcal{S}^{d'-1} = \{\mathbf{x} \in \mathbb{R}^{d'} : \|\mathbf{x}\|_2 = 1\}$. Let $(x)_+ = \max(x, 0)$.

Statistical convergence rates will be denoted by the $O_P(\cdot)$ notation, which is defined as follows. Consider a sequence of real random variables $\{X_n\}_{n\in\mathbb{N}}$ defined on a probability space $(\Omega, \mathcal{F}, P)$ and a deterministic (positive) sequence of reals $\{b_n\}_{n\in\mathbb{N}}$. By $X_n = O_P(1)$ is meant: $\sup_{n\in\mathbb{N}} \Pr(|X_n| > K) \to 0$ as $K \to \infty$, where X_n is a sequence indexed by n, for fixed p, q. The notation $X_n = O_P(b_n)$ is equivalent to $\frac{X_n}{b_n} = O_P(1)$. By $X_n = o_p(1)$ is meant $\Pr(|X_n| > \epsilon) \to 0$ as $n \to \infty$ for any $\epsilon > 0$. By $\lambda_n \asymp b_n$ is meant $c_1 \leq \frac{\lambda_n}{b_n} \leq c_2$ for all n, where $c_1, c_2 > 0$ are absolute constants.

3.3 Permuted Rank-penalized Least-squares

Available are n i.i.d. multivariate Gaussian observations $\{\mathbf{z}_t\}_{t=1}^n$, $\mathbf{z}_t \in \mathbb{R}^{pq}$, having zero-mean and covariance equal to (3.1). A sufficient statistic for estimating the covariance is the well-known sample covariance matrix (SCM):

$$\hat{\mathbf{S}}_n = \frac{1}{n}\sum_{t=1}^{n} \mathbf{z}_t \mathbf{z}_t^T \tag{3.2}$$

The SCM is an unbiased estimator of the true covariance matrix. However, when the number of samples n is smaller than the number of variables $d = pq$ the SCM suffers from high variance and a low rank approximation to the SCM is commonly used. The most common low rank approximation is to perform the eigendecomposition of $\hat{\mathbf{S}}_n$ and retain only the top r principal components resulting in an estimator, called the PCA estimator, of the form:

$$\hat{\mathbf{S}}_n^{PCA} = \sum_{i=1}^{r} \sigma_i^2 \nu_i \nu_i^T, \tag{3.3}$$

where $r < d$ is selected according to some heuristic. It is now well known [82, 97] that this PCA estimator suffers from high bias when n is smaller than $d = pq$.

An alternative approach to low rank covariance estimation was proposed in [85] specifying a low rank covariance estimator as the solution of the penalized least squares problem [2]:

$$\hat{\mathbf{\Sigma}}_n^{\lambda} \in \arg\min_{\mathbf{S}\in S_{++}^d} \|\hat{\mathbf{S}}_n - \mathbf{S}\|_F^2 + \lambda \mathrm{tr}(\mathbf{S}) \tag{3.4}$$

where $\lambda > 0$ is a regularization parameter.

The estimator (3.4) has several useful interpretations. First, it can be interpreted as a convex relaxation of the non-convex rank constrained Frobenius norm minimization problem

$$\arg\min_{\mathbf{S}\in S_{++}^d, \mathrm{rank}(\mathbf{S})\leq r} \| \hat{\mathbf{S}}_n - \mathbf{S} \|_F^2,$$

whose solution, by the Eckhart-Young theorem, is the PCA estimator (3.3). Second, it can be interpreted as a covariance version of the lasso regression problem, i.e., finding a low rank psd ℓ_2 approximation to the sample covariance matrix. The term $\mathrm{tr}(\mathbf{S})$ in 3.4 is equivalent to the ℓ_1 norm on the eigenvalues of the psd matrix $\mathbf{S}$. As shown in [85] the solution to the convex minimization in (3.4) converges to the ensemble covariance $\mathbf{\Sigma}_0 = \mathbb{E}[\mathbf{z}_t \mathbf{z}_t^T]$ at the minimax optimal rate. Corollary 1 in [85] establishes that, for $\lambda = C' \|\mathbf{\Sigma}_0\|_2 \sqrt{\frac{r(\mathbf{\Sigma}_0)\log(2d)}{n}}$, $n \geq c r(\mathbf{\Sigma}_0)\log^2(\max(2d, n))$ and $C', c > 0$ sufficiently large, establishes a tight Frobenius norm error bound, which states that with probability $1 - \frac{1}{2d}$:

$$\|\hat{\mathbf{\Sigma}}_n^{\lambda} - \mathbf{\Sigma}_0\|_F^2 \leq \inf_{\mathbf{S}\succ 0} \|\mathbf{\Sigma}_0 - \mathbf{S}\|_F^2 + C\|\mathbf{\Sigma}_0\|_2^2 \mathrm{rank}(\mathbf{S}) \frac{r(\mathbf{\Sigma}_0)\log(2d)}{n}$$

where $r(\mathbf{\Sigma}_0) = \frac{\mathrm{tr}(\mathbf{\Sigma}_0)}{\|\mathbf{\Sigma}_0\|_2} \leq \mathrm{rank}(\mathbf{\Sigma}_0)$ is the effective rank [85]. The absolute constant C is given by $\frac{(1+\sqrt{2})^2}{8}(C')^2$.

Here we propose a similar nuclear norm penalization approach to estimate low separation-rank covariance matrices of form (3.1). Motivated by Van Loan and

[2] The estimator (3.4) was developed in [85] for the more general problem where there could be missing data.

Pitsianis's work [84], we propose:

$$\hat{\mathbf{R}}_n^{\lambda} \in \arg \min_{\mathbf{R} \in \mathbb{R}^{p^2 \times q^2}} \|\hat{\mathbf{R}}_n - \mathbf{R}\|_F^2 + \lambda \|\mathbf{R}\|_* \tag{3.5}$$

where $\hat{\mathbf{R}}_n = \mathcal{R}(\hat{\mathbf{S}}_n)$ is the permuted SCM of size $p^2 \times q^2$ (see Notation section). The minimum-norm problem considered in [84] is:

$$\min_{\mathbf{R} \in \mathbb{R}^{p^2 \times q^2} : \text{rank}(\mathbf{R}) \leq r} \|\hat{\mathbf{R}}_n - \mathbf{R}\|_F^2 \tag{3.6}$$

Specifically, let $\mathbf{S} = \sum_{i=1}^{r} \mathbf{A}_i \otimes \mathbf{B}_i$ where for all i the dimensions of the matrices $\mathbf{A}_i$ and $\mathbf{B}_i$ are fixed. Then, as the Frobenius norm of a matrix is invariant to permutation of its elements, it follows that $\| \mathbf{S}_n - \mathbf{S} \|_F = \| \mathbf{R}_n - \mathbf{R} \|_F$ where $\mathbf{R}_n = \mathcal{R}(\mathbf{S}_n)$ and $\mathbf{R} = \mathcal{R}(\mathbf{S})$ (which is a matrix of algebraic rank r).

We note that (3.5) is a convex relaxation of (3.6) and is more amenable to numerical optimization. Furthermore, we show a tradeoff between approximation error (i.e., the error induced by model mismatch between the true covariance and the model) and estimation error (i.e., the error due to finite sample size) by analyzing the solution of (3.5). We also note that (3.5) is a strictly convex problem, so there exists a unique solution that can be efficiently found using well established numerical methods [18].

The solution of (3.5) is closed form and is given by a thresholded singular value decomposition:

$$\hat{\mathbf{R}}_n^{\lambda} = \sum_{j=1}^{\min(p^2, q^2)} \left(\sigma_j(\hat{\mathbf{R}}_n) - \frac{\lambda}{2} \right)_+ \mathbf{u}_j \mathbf{v}_j^T \tag{3.7}$$

where $\mathbf{u}_j$ and $\mathbf{v}_j$ are the left and right singular vectors of $\hat{\mathbf{R}}_n$. This is converted back to a square $pq \times pq$ matrix $\hat{\mathbf{\Sigma}}_n^{\lambda}$ by applying the inverse permutation operator $\mathcal{R}^{-1}$ to $\hat{\mathbf{R}}_n$ (see Notation section).

Efficient methods for numerically evaluating penalized objectives like (3.5) have been recently proposed [21, 22] and do not require computing the full SVD. Although empirically observed to be fast, the computational complexity of the algorithms presented in [21] and [22] is unknown. The rank-r SVD can be computed with $O(p^2q^2r)$ floating point operations. There exist faster randomized methods for truncated SVD requiring only $O(p^2q^2\log(r))$ floating point operations [61]. Thus, the computational complexity of solving (3.5) scales well with respect to the desired separation rank r.

The next theorem shows that the de-permuted version of (3.7) is symmetric and positive definite.

Theorem III.1. *Consider the de-permuted solution* $\hat{\mathbf{\Sigma}}_n^\lambda = \mathcal{R}^{-1}(\hat{\mathbf{R}}_n^\lambda)$. *The following are true:*

1. *The solution* $\hat{\mathbf{\Sigma}}_n^\lambda$ *is symmetric with probability 1.*

2. *If* $n \geq pq$, *then the solution* $\hat{\mathbf{\Sigma}}_n^\lambda$ *is positive definite with probability 1.*

Proof. See Appendix. □

We believe that the PRLS estimate $\hat{\mathbf{\Sigma}}_n^\lambda$ is positive definite even if $n < pq$ for appropriately selected $\lambda > 0$. In our simulations, we always found $\hat{\mathbf{\Sigma}}_n^\lambda$ to be positive definite. We have also found that the condition number of the PRLS estimate is orders of magnitude smaller than that of the SCM.

3.4 High Dimensional Consistency of PRLS

In this section, we show that RPLS achieves the MSE statistical convergence rate of $O_P\left(\frac{r(p^2+q^2+\log M)}{n}\right)$. This result is clearly superior to the statistical convergence rate of the naive SCM estimator [124],

$$\|\hat{\mathbf{S}}_n - \mathbf{\Sigma}_0\|_F^2 = O_P\left(\frac{p^2q^2}{n}\right), \tag{3.8}$$

particularly when $p, q \to \infty$.

The next result provides a relation between the spectral norm of $\hat{\mathbf{R}}_n - \mathbf{R}_0$, the Frobenius norm of $\mathbf{R} - \mathbf{R}_0$ and the Frobenius norm of the the estimation error $\hat{\mathbf{R}}_n^\lambda - \mathbf{R}_0$.

Theorem III.2. *Consider the convex optimization problem (3.5). When* $\lambda \geq 2\|\hat{\mathbf{R}}_n - \mathbf{R}_0\|_2$, *the following holds:*

$$\|\hat{\mathbf{R}}_n^\lambda - \mathbf{R}_0\|_F^2 \leq \inf_{\mathbf{R}} \left\{ \|\mathbf{R} - \mathbf{R}_0\|_F^2 + \frac{(1+\sqrt{2})^2}{4} \lambda^2 rank(\mathbf{R}) \right\} \tag{3.9}$$

Proof. See Appendix. □

3.4.1 High Dimensional Operator Norm Bound for the Permuted Sample Covariance Matrix

In this subsection, we establish a tight bound on the spectral norm of the error matrix

$$\boldsymbol{\Delta}_n = \hat{\mathbf{R}}_n - \mathbf{R}_0 = \mathcal{R}(\hat{\mathbf{S}}_n - \boldsymbol{\Sigma}_0). \tag{3.10}$$

The standard strong law of large numbers implies that for fixed dimensions p, q, we have $\boldsymbol{\Delta}_n \to 0$ almost surely as $n \to \infty$. The next result will characterize the finite sample fluctuations of this convergence (in probability) measured by the spectral norm as a function of the sample size n and Kronecker factor dimensions p, q. This result will be useful for establishing a tight bound on the Frobenius norm convergence rate of PRLS and can guide the selection of the regularization paramater in (3.5).

Theorem III.3. *(Operator Norm Bound on Permuted SCM) Assume* $\|\boldsymbol{\Sigma}_0\|_2 < \infty$ *for all* p, q *and define* $M = \max(p, q, n)$. *Fix the constant* $\epsilon' < \frac{1}{2}$. *Assume* $t \geq \max(\sqrt{4C_1 \ln(1 + \frac{2}{\epsilon'})}, 4C_2 \ln(1 + \frac{2}{\epsilon'}))$ *and* $C = \max(C_1, C_2) > 0$ [3]. *Then, with proba-*

[3] The constants C_1, C_2 are defined in Lemma III.7 in Appendix B.

bility at least $1 - 2M^{-\frac{t}{4C}}$,

$$\|\mathbf{\Delta}_n\|_2 \leq \frac{C_0 t}{1-2\epsilon'} \max\left\{\frac{p^2+q^2+\log M}{n}, \sqrt{\frac{p^2+q^2+\log M}{n}}\right\} \tag{3.11}$$

where $C_0 = \|\mathbf{\Sigma}_0\|_2 > 0$ [4].

Proof. See Appendix. □

The proof technique is based on a large deviation inequality, derived in Lemma III.7 in Appendix C. This inequality characterizes the tail behavior of the quadratic form $\mathbf{x}^T \mathbf{\Delta}_n \mathbf{y}$ over the spheres $\mathbf{x} \in \mathcal{S}_{p^2-1}$ and $\mathbf{y} \in \mathcal{S}_{q^2-1}$. Using Lemma III.7 and a sphere covering argument, the result of Theorem III.3 follows (see Appendix). Fig. 3.2 empirically validates the tightness of the bound (3.11) under the trivial separation rank 1 covariance $\mathbf{\Sigma}_0 = \mathbf{I}_p \otimes \mathbf{I}_q$.

3.4.2 High Dimensional MSE Convergence Rate for PRLS

Using the result in Thm. III.3 and the bound in Thm. III.2, we next provide a tight bound on the MSE estimation error.

Theorem III.4. *Define the variable* $M = \max(p, q, n)$. *Set the regularization parameter* $\lambda = \lambda_n = \frac{2C_0 t}{1-2\epsilon'} \max\left\{\frac{p^2+q^2+\log M}{n}, \sqrt{\frac{p^2+q^2+\log M}{n}}\right\}$ *with* t *satisfying the conditions of Thm. III.3. Then, with probability at least* $1 - 2M^{-\frac{t}{4C}}$*:*

$$\begin{aligned}\|\hat{\mathbf{\Sigma}}_n^{\lambda} - \mathbf{\Sigma}_0\|_F^2 &\leq \inf_{\mathbf{R}: rank(\mathbf{R}) \leq r} \|\mathbf{R} - \mathbf{R}_0\|_F^2 \\ &+ C' r \max\left\{\left(\frac{p^2+q^2+\log M}{n}\right)^2, \frac{p^2+q^2+\log M}{n}\right\}\end{aligned} \tag{3.12}$$

where $C' = \left(C_0 t \frac{1+\sqrt{2}}{1-2\epsilon'}\right)^2 = \left(3(1+\sqrt{2})C_0 t\right)^2 > 0$.

Proof. See Appendix. □

[4] The constant $\frac{C_0 t}{1-2\epsilon'}$ in front of the rate can be optimized by minimizing it as a function of ϵ' over the interval $(0, 1/2)$.

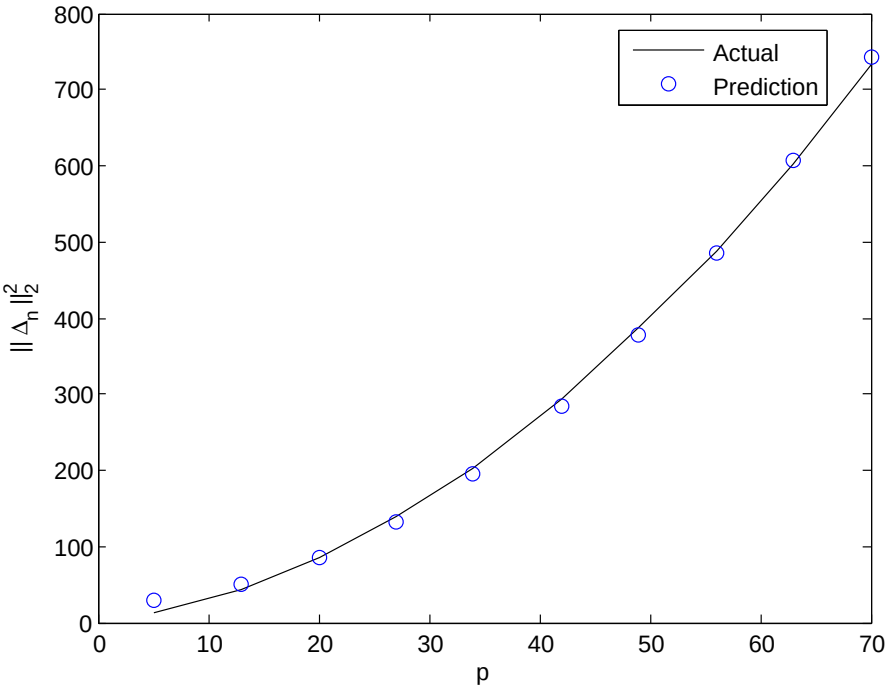


Figure 3.2: Monte Carlo simulation for growth of spectral norm $\|\mathbf{\Delta}_n\|_2^2$ as a function of p for fixed $n = 10$ and $q = 5$. The predicted curve is a least-square fit of a quadratic model $y = ax^2 + b$ to the empirical curve, and is a great fit. This example shows the tightness of the probabilistic bound (3.11).

When $\mathbf{\Sigma}_0$ is truly a sum of r Kronecker products with factor dimensions p and q, there is no model mismatch and the approximation error $\inf_{\{\mathbf{R}:\text{rank}(\mathbf{R})\leq r\}} \|\mathbf{R} - \mathbf{R}_0\|_F^2$ is zero. In this case, in the large-p, q, n asymptotic regime where $p^2 + q^2 + \log M = o(n)$, it follows that $\|\hat{\mathbf{\Sigma}}_n^\lambda - \mathbf{\Sigma}_0\|_F = O_P(\sqrt{\frac{r(p^2+q^2+\log M)}{n}}) = o_p(1)$. This asymptotic MSE convergence rate of the estimated covariance to the true covariance reflects the number of degrees of freedom of the model, which is on the order of the total number $r(p^2 + q^2)$ of unknown parameters. This result extends the recent high-dimensional results obtained in [116, 114, 115] for the single Kronecker product model (i.e., $r = 1$).

Recall that $r \leq r_0 = \min(p^2, q^2)$. For the case when $p \sim q$, and $r \sim r_0$, we have a fully saturated Kronecker product model and the number of model parameters are of the order $p^4 \sim d^2$, and the SCM convergence rate (3.8) coincides with the rate

obtained in Thm. III.4.

For covariance models of low separation rank-i.e., $r \ll r_0$, Thm. III.4 asserts that the high dimensional MSE convergence rate of PRLS can be much lower than the naive SCM convergence rate. Thus PRLS is an attractive alternative to rank-based series expansions like principal component analysis (PCA). We note that each term in the expansion $\mathbf{A}_{0,\gamma} \otimes \mathbf{B}_{0,\gamma}$ can be full-rank, while each term in the standard PCA expansion is rank 1.

Finally, we observe that Thm. III.4 captures the tradeoff between estimation error and approximation error. In other words, choosing a smaller r than the true separation rank would incur a larger approximation error $\inf_{\{\mathbf{R}:\mathrm{rank}(\mathbf{R})\leq r\}} \|\mathbf{R}-\mathbf{R}_0\|_F^2 > 0$, but smaller estimation error on the order of $O_P(\frac{r(p^2+q^2+\log M)}{n})$.

3.4.3 Approximation Error

It is well known from least-squares approximation theory that the residual error can be rewritten as:

$$\inf_{\mathbf{R}:\mathrm{rank}(\mathbf{R})\leq r} \|\mathbf{R}-\mathbf{R}_0\|_F^2 = \sum_{k=r+1}^{r_0} \sigma_k^2(\mathbf{R}_0), \tag{3.13}$$

where $\{\sigma_k(\mathbf{R}_0)\}$ are the singular values of $\mathbf{R}_0$. In the high dimensional setting, the sample size n grows with the dimensions p, q so that the maximum separation rank r_0 also grows to infinity, and the approximation error (3.13) may not be finite. In this case the bound in Theorem III.4 will not be finite. Hence, an additional condition will be needed to ensure that the sum (3.13) remains finite as $p, q \to \infty$: the singular values of $\mathbf{R}_0$ need to decay faster than $O(1/k)$.

We show next that the class of block-Toeplitz covariance matrices have bounded approximation error if the separation rank scales like $\log(\max(p, q))$. To show this, we first provide a tight variational bound on the singular value spectrum of any $p^2 \times q^2$

matrix $\mathbf{R}$. Note that the work on high dimensional Toeplitz covariance estimation under operator and Frobenius norms [23, 14] are not applicable to the block-Toeplitz case. To establish Thm. III.6 on block Toeplitz matrices we first need the following Lemma.

Lemma III.5. *(Variational Bound on Singular Value Spectrum) Let* $\mathbf{R}$ *be an arbitrary matrix of size* $p^2 \times q^2$. *Let* $\mathbf{P}_k$ *be an orthogonal projection of* $\mathbb{R}^{q^2}$ *onto* $\mathbb{R}^k$. *Then, for* $k = 1, \ldots, r_0 - 1$ *we have:*

$$\sigma_{k+1}^2(\mathbf{R}) \leq \|(\mathbf{I}_{q^2} - \mathbf{P}_k)\mathbf{R}^T\|_2^2 \tag{3.14}$$

with equality iff $\mathbf{P}_k = \mathbf{V}_k\mathbf{V}_k^T$. *Also,* $\mathbf{V}_k = [\mathbf{v}_1, \ldots, \mathbf{v}_k]$, *where* $\mathbf{v}_i$ *is the* i*th column of* $\mathbf{V}$ *and* $\mathbf{R} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ *is the singular value decomposition.*

Proof. See Appendix. □

Using this fundamental lemma, we can characterize the approximation error for estimating block-Toeplitz matrices with exponentially decaying off-diagonal norms. Such matrices arise, for example, as covariance matrices of multivariate stationary random processes of dimension m (see (3.17)) and take the block Toeplitz form:

$$\underbrace{\mathbf{\Sigma}_0}_{(N+1)m\times(N+1)m} = \begin{bmatrix} \mathbf{\Sigma}(0) & \mathbf{\Sigma}(1) & \ldots & \mathbf{\Sigma}(N) \\ \mathbf{\Sigma}(-1) & \mathbf{\Sigma}(0) & \ldots & \mathbf{\Sigma}(N-1) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{\Sigma}(-N) & \mathbf{\Sigma}(-N+1) & \ldots & \mathbf{\Sigma}(0) \end{bmatrix} \tag{3.15}$$

where each submatrix is of size $m \times m$. For a zero-mean vector process $\mathbf{y} = \{\mathbf{y}(0), \ldots, \mathbf{y}(N)\}$, the submatrices are given by $\mathbf{\Sigma}(\tau) = \mathbb{E}[\mathbf{y}(0)\mathbf{y}(\tau)^T]$.

Theorem III.6. *Consider a block-Toeplitz p.d. matrix* $\mathbf{\Sigma}_0$ *of size* $(N+1)m \times (N+1)m$, *with* $\|\mathbf{\Sigma}(\tau)\|_F^2 \leq C'u^{2|\tau|}q$ *for all* $\tau = -N, \ldots, N$ *and constant* $u \in (0, 1)$. *Let* $\hat{\mathbf{\Sigma}}_n^\lambda$

be the de-permuted matrix $\mathcal{R}^{-1}(\hat{\mathbf{R}}_n^\lambda)$*, where* $\hat{\mathbf{R}}_n^\lambda$ *is given in (3.7). Using the minimal separation rank* r*:*

$$r \geq \frac{\log(pq/\epsilon)}{\log(1/u)}.$$

Then, the PRLS algorithm estimates $\boldsymbol{\Sigma}_0$ *up to an absolute tolerance* $\epsilon \in (0,1)$ *with convergence rate guarantee:*

$$\|\hat{\boldsymbol{\Sigma}}_n^\lambda - \boldsymbol{\Sigma}_0\|_F^2 \leq \epsilon + C' r \frac{p^2 + q^2 + \log M}{n} \tag{3.16}$$

holding with probability at least $1 - \max(p,q,n)^{-t/4C}$ *for* λ *chosen as perscribed in Thm. III.4. Here,* $t > 1$ *is constant and* $C, C' > 0$ *are constants specified in Thm. III.4.*

Proof. See Appendix. □

The exponential norm decay condition of Thm. III.6 is satisfied by a first-order vector autoregressive process:

$$\mathbf{Z}_t = \Phi \mathbf{Z}_{t-1} + \mathcal{E}_t \tag{3.17}$$

with $u = \|\Phi\|_2 \in (0,1)$, where $\mathbf{Z}_t \in \mathbb{R}^m$. For $\mathcal{E}_t \sim N(0, \boldsymbol{\Sigma}_\epsilon)$, this is a multivariate Gaussian process. Collecting data over a time horizon of size $N+1$, we concatenate these observations into a large random vector $\mathbf{z}$ of dimension $(N+1)m$, where m is the process dimension. The resulting covariance matrix has the block-Toeplitz form assumed in Thm. III.6. Figure 3.3 shows bounds constructed using the Frobenius upper bound on the spectral norm in (3.14) and using the projection matrix $\mathbf{P}_k$ as discussed in the proof of Thm. III.6. The bound given in the proof of Thm. III.6 (in black) is shown to be linear in log-scale, thus justifying the exponential decay of the Kronecker spectrum.

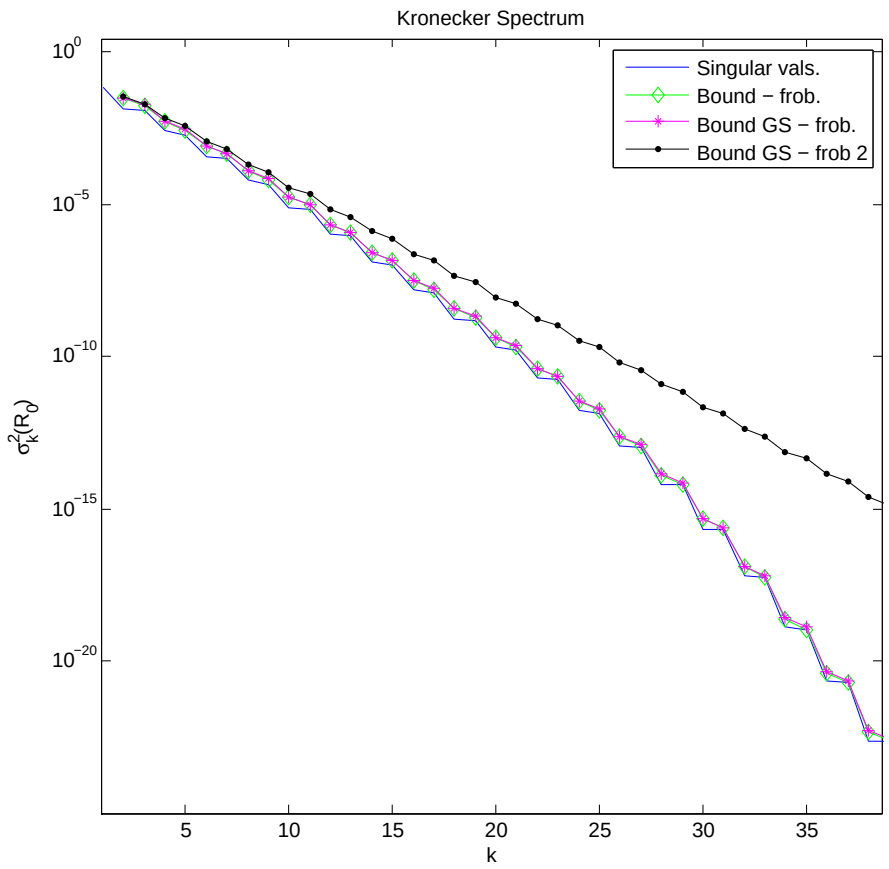


Figure 3.3: Kronecker spectrum and bounds based on Lemma III.5. The upper bound 'Bound - frob' (in green) is obtained using the bound (3.14) using the basis associated with the minimum ℓ_2 approximation error (i.e., the optimal basis computed by SVD as outlined in the equality condition of Lemma III.5). The upper bound 'Bound GS - frob' (in magenta) is constructed using the variational bound (3.14) with projection matrix $\mathbf{P}_k$ having columns drawn from the orthonormal basis constructed in the proof of Thm. III.6. The upper bound 'Bound GS - frob 2' (in black) is constructed from the bound (3.47) in the proof of Thm. III.6.

3.5 Simulation Results

We consider dense positive definite matrices $\mathbf{\Sigma}_0$ of dimension $d = 625$. Taking $p = q = 25$, we note that the number of free parameters that describe each Kronecker product is of the order $p^2 + q^2 \sim p^2$, which is essentially of the same order as the number of unknown parameters required to specify each eigenvector of $\mathbf{\Sigma}_0$, i.e., $pq \sim p^2$.

3.5.1 Sum of Kronecker Product Covariance

The covariance matrix shown in Fig. 3.4 was constructed using (3.1) with $r = 3$, with each p.d. factor chosen as $\mathbf{C}\mathbf{C}^T$, where $\mathbf{C}$ is a square Gaussian random matrix. Fig. 3.5 shows the empirical performance of covariance matching (CM) (i.e., solution of (3.6) with $r = 3$), PRLS and SVT (i.e., solution of (3.4)). We note that the Kronecker spectrum contains only three nonzero terms while the true covariance is full rank. The PRLS spectrum is more concentrated than the eigenspectrum and, from Fig. 3.5, we observe PRLS outperforms covariance matching (CM), SVT and SCM across all n.

3.5.2 Block Toeplitz Covariance

The covariance matrix shown in Fig. 3.6 was constructed by first generating a Gaussian random square matrix Φ of spectral norm $0.95 < 1$, and then simulating the block Toeplitz covariance for the process shown in (3.17). Fig. 3.7 compares the empirical performance of PRLS and SVT (i.e., the solution of (3.4) with appropriate scaling for the regularization parameter). We observe that the Kronecker product estimator performs much better than both SVT (i.e., the solution of (3.4)) and naive SCM estimator. This is most likely due to the fact that the repetitive block structure of Kronecker products better summarizes the covariance structure. We observe from Fig. 3.6 that for this block Toeplitz covariance, the Kronecker spectrum decays more rapidly (exponentially) than the eigenspectrum.

3.6 Application to Wind Speed Prediction

In this section, we demonstrate the performance of PRLS in a real world application: wind speed prediction. We apply our methods to the Irish wind speed dataset and the NCEP dataset.

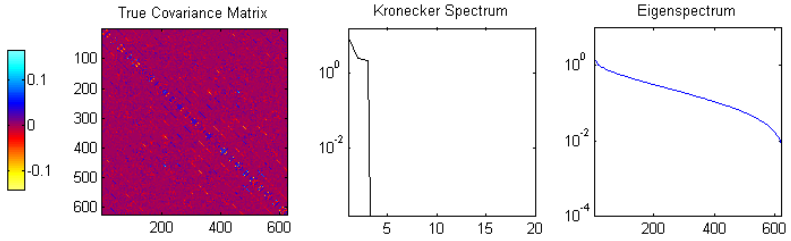


Figure 3.4: Simulation A. True dense covariance is constructed using the sum of KP model (3.1), with $r = 3$. Left panel: True positive definite covariance matrix $\mathbf{\Sigma}_0$. Middle panel: Kronecker spectrum (eigenspectrum of $\mathbf{\Sigma}_0$ in permuted domain). Right panel: Eigenspectrum (Eigenvalues of $\mathbf{\Sigma}_0$). Note that the Kronecker spectrum is much more concentrated than the eigenspectrum.

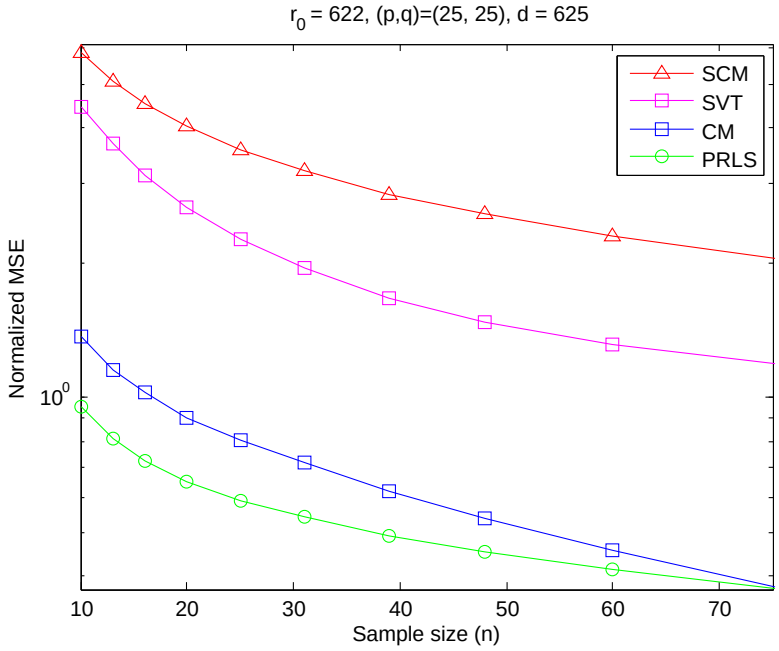


Figure 3.5: Simulation A. Normalized MSE performance for true covariance matrix in Fig. 3.4 as a function of sample size n. PRLS outperforms CM, SVT (i.e., solution of (3.4)) and the standard SCM estimator. Here, $p = q = 25$ and $N_{MC} = 80$. For $n = 20$, PRLS achieves a 7.91 dB MSE reduction over SCM and SVT achieves a 1.80 dB MSE reduction over SCM.

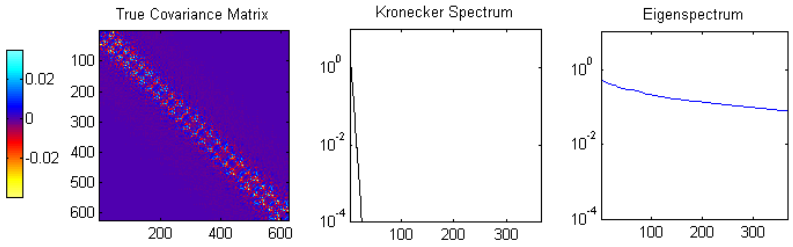


Figure 3.6: Simulation B. True dense block-Toeplitz covariance matrix. Left panel: True positive definite covariance matrix $\mathbf{\Sigma}_0$. Middle panel: Kronecker spectrum (eigenspectrum of $\mathbf{\Sigma}_0$ in permuted domain). Right panel: Eigenspectrum (Eigenvalues of $\mathbf{\Sigma}_0$). Note that the Kronecker spectrum is much more concentrated than the eigenspectrum.

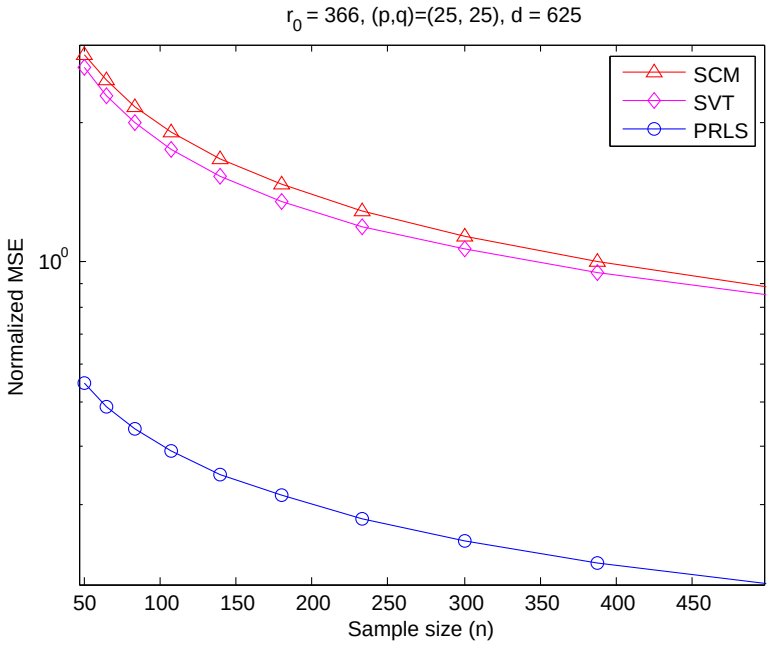


Figure 3.7: Simulation B. Normalized MSE performance for covariance matrix in Fig. 3.6 as a function of sample size n. PRLS outperforms SVT (i.e., solution of (3.4)) and the standard SCM estimator. Here, $p = q = 25$ and $N_{MC} = 80$. For $n = 108$, PRLS achieves a 6.88 dB MSE reduction over SCM and SVT achieves a 0.37 dB MSE reduction over SCM. Note again that the Kronecker spectrum is much more concentrated than the eigenspectrum.

3.6.1 Irish Wind Speed Data

We use data consisting of time series consisting of daily average wind speed recordings during the period $1961 - 1978$ at $q = 11$ meteorological stations. This data set has many temporal coordinates, spanning a total of $n_{total} = 365 \cdot 8 = 2920$ daily average recordings of wind speed at each station. More details on this data set can be found in [62, 57, 39, 109] and it can be downloaded from Statlib http://lib.stat.cmu.edu/datasets. We used the same square root transformation, estimated seasonal effect offset and station-specific mean offset as in [62], yielding the multiple (11) velocity measures. We used the data from years $1969-1970$ for training and the data from $1971 - 1978$ for testing.

The task is to predict the average velocity for the next day using the average wind velocity in each of the $p - 1$ previous days. The full dimension of each observation vector is $d = pq$, and each d-dimensional observation vector is formed by concatenating the p time-consecutive q-dimensional vectors (each entry containing the velocity measure for each station) without overlapping the time segments. The SCM was estimated using data from the training period consisting of years $1969 - 1970$. Linear predictors over the time series were constructing by using these estimated covariance matrices in an ordinary least squares predictor. Specifically, we constructed the SCM linear predictor of all stations' wind velocity from the $p - 1$ previous samples of the $q = 11$ stations' time series:

$$\hat{\mathbf{v}}_t = \mathbf{\Sigma}_{2,1}\mathbf{\Sigma}_{1,1}^{-1}\mathbf{v}_{t-1:t-(p-1)} \tag{3.18}$$

where $\mathbf{v}_{t-1:t-(p-1)} \in \mathbb{R}^{(p-1)q}$ is the stacked wind velocities from the previous $p - 1$ time instants and $\mathbf{\Sigma}_{2,1} \in \mathbb{R}^{q \times q(p-1)}$ and $\mathbf{\Sigma}_{1,1} \in \mathbb{R}^{q(p-1) \times q(p-1)}$ are submatrices of the

$qp \times qp$ standard SCM:

$$\hat{\mathbf{S}}_n = \begin{bmatrix} \mathbf{\Sigma}_{1,1} & \mathbf{\Sigma}_{1,2} \\ \mathbf{\Sigma}_{2,1} & \mathbf{\Sigma}_{2,2} \end{bmatrix}$$

The PRLS predictor was similarly constructed using our proposed estimator of the $qp \times qp$ Kronecker sum covariance matrix instead of the SCM. The coefficients of each of these predictors, $\mathbf{\Sigma}_{2,1}\mathbf{\Sigma}_{1,1}^{-1}$, were subsequently applied to predict over the test set.

The predictors were tested on the data from years $1971 - 1978$, corresponding to $n_{test} = 365 \cdot 8 = 2920$ days, as the ground truth. Using non-overlapping samples and $p = 8$, we have a total of $n = \lceil \frac{365 \cdot 2}{p} \rceil = 91$ training samples of full dimension $d = 88$.

Fig. 3.8 shows the Kronecker product factors that make up the solution of Eq. (3.6) with $r = 1$ and the PRLS estimate. The PRLS estimate contains $r_{eff} = 6$ nonzero terms in the KP expansion. It is observed that the first order temporal factor gives a decay in correlations over time, and spatial correlations between weather stations are present. The second order temporal and spatial factors can potentially give insight into long range dependencies.

Fig. 3.10 shows the root mean squared error (RMSE) prediction performance over the testing period of 2920 days for the forecasts based on the standard SCM, PRLS estimator, Lounici's SVT estimator [85], and regularized Tyler [33]. The PRLS estimator was implemented using a regularization parameter $\lambda_n = C\|\hat{\mathbf{S}}_n\|_2 \sqrt{\frac{p^2+q^2+\log(\max(p,q,n))}{n}}$ with $C = 0.13$. The constant C was chosen by optimizing the prediction RMSE on the training set over a range of regularization parameters λ parameterized by C. The SVT estimator proposed by Lounici [85] was implemented using a regularization parameter $\lambda = C\sqrt{\mathrm{tr}(\hat{S}_n)\|\hat{S}_n\|_2}\sqrt{\frac{\log(2pq)}{n}}$ with constant $C = 1.9$ optimized in a similar manner. The regularized Tyler estimator was implemented using the data-dependent

shrinkage coefficient suggested in Eqn. (13) in [33]. Fig. 3.11 shows a sample period of 150 days. We observe that PRLS tracks the actual wind speed better than the SCM-based predictor does.

3.6.2 NCEP Wind Speed Data

We use data representative of the wind conditions in the lower troposphere (surface data at .995 sigma level) for the global grid (90°N - 90°S, 0°E - 357.5°E). We obtained the data from the National Centers for Environmental Prediction reanalysis project (Kalnay et al. [76]), which is available online at the NOAA website ftp://ftp.cdc.noaa.gov/Datasets/ncep.reanalysis.dailyavgs/surface. Daily averages of U (east-west) and V (north-south) wind components were collected using a station grid of size 144×73 (2.5 degree latitude $\times$ 2.5 degree longitude global grid) over the years $1948 - 2012$. The wind speed is computed by taking the magnitude of the wind vector.

Continental US Region

We considered a 10×10 grid of stations, corresponding to latitude range 25°N-47.5°N and longitude range 125°W-97.5°W. For this selection of variables, $q = 10 \cdot 10 = 100$ is the total number of stations and $p - 1 = 7$ is the prediction time lag. We preprocessed the raw data using the detrending procedure outlined in Haslett et al. [62]. More specifically, we first performed a square root transformation, then estimated and subtracted the station-specific means from the data and finally estimated and subtracted the seasonal effect (see Fig. 3.12). The resulting features/observations are called the velocity measures [62]. The SCM was estimated using data from the training period consisting of years $2003 - 2007$. Since the SCM is not full rank, the linear preictor (3.18) was implemented with the Moore-Penrose

pseudo-inverse of $\mathbf{\Sigma}_{1,1}$. The predictors were tested on the data from years $2008-2012$ as the ground truth. Using non-overlapping samples and $p = 8$, we have a total of $n = \lceil \frac{365 \cdot 5}{p} \rceil = 228$ training samples of full dimension $d = 800$.

Fig. 3.13 shows the Kronecker product factors that make up the solution of Eq. (3.6) with $r = 2$ and the PRLS covariance estimate. The PRLS estimate contains $r_{eff} = 6$ nonzero terms in the KP expansion. It is observed that the first order temporal factor gives a decay in correlations over time, and spatial correlations between weather stations are present. The second order temporal and spatial factors give some insight into longer range dependencies.

Fig. 3.15 shows the root mean squared error (RMSE) prediction performance over the testing period of 1825 days for the forecasts based on the standard SCM, PRLS, SVT [85] and regularized Tyler [33]. The PRLS estimator was implemented using a regularization parameter $\lambda_n = C\|\hat{\mathbf{S}}_n\|_2 \sqrt{\frac{p^2+q^2+\log(\max(p,q,n))}{n}}$ with $C = 0.036$. The constant C was chosen by optimizing the prediction RMSE on the training set over a range of regularization parameters λ parameterized by C (as in Irish wind speed data set). The SVT estimator proposed by Lounici [85] was implemented using a regularization parameter $\lambda = C\sqrt{\mathrm{tr}(\hat{S}_n)\|\hat{S}_n\|_2}\sqrt{\frac{\log(2pq)}{n}}$ with constant $C = 0.31$ optimized in a similar manner. Fig. 3.16 shows a sample period of 150 days. It is observed that SCM has unstable performance, while the Kronecker product estimator offers better tracking of the wind speeds.

Arctic Ocean Region

We considered a 10×10 grid of stations, corresponding to latitude range 90°N-67.5°N and longitude range 0°E-22.5°E. For this selection of variables, $q = 10 \cdot 10 = 100$ is the total number of stations and $p - 1 = 7$ is the prediction time lag. We pre-processed the raw data using the detrending procedure outlined in Haslett et al. [62].

More specifically, we first performed a square root transformation, then estimated and subtracted the station-specific means from the data and finally estimated and subtracted the seasonal effect (see Fig. 3.17). The resulting features/observations are called the velocity measures [62]. The SCM was estimated using data from the training period consisting of years $2003 - 2007$. Since the SCM is not full rank, the linear preictor (3.18) was implemented with the Moore-Penrose pseudo-inverse of $\mathbf{\Sigma}_{1,1}$. The predictors were tested on the data from years $2008 - 2012$ as the ground truth. Using non-overlapping samples and $p = 8$, we have a total of $n = \lceil \frac{365 \cdot 5}{p} \rceil = 228$ training samples of full dimension $d = 800$.

Fig. 3.18 shows the Kronecker product factors that make up the solution of Eq. (3.6) with $r = 2$ and the PRLS covariance estimate. The PRLS estimate contains $r_{eff} = 2$ nonzero terms in the KP expansion. It is observed that the first order temporal factor gives a decay in correlations over time, and spatial correlations between weather stations are present. The second order temporal and spatial factors give some insight into longer range dependencies.

Fig. 3.20 shows the root mean squared error (RMSE) prediction performance over the testing period of 1825 days for the forecasts based on the standard SCM, PRLS, and regularized Tyler [33]. The PRLS estimator was implemented using a regularization parameter $\lambda_n = C\|\hat{\mathbf{S}}_n\|_2 \sqrt{\frac{p^2+q^2+\log(\max(p,q,n))}{n}}$ with $C = 0.073$. The constant C was chosen by optimizing the prediction RMSE on the training set over a range of regularization parameters λ parameterized by C (as in Irish wind speed data set). The SVT estimator proposed by Lounici [85] was implemented using a regularization parameter $\lambda = C\sqrt{\mathrm{tr}(\hat{S}_n)\|\hat{S}_n\|_2}\sqrt{\frac{\log(2pq)}{n}}$ with constant $C = 0.47$ optimized in a similar manner. Fig. 3.21 shows a sample period of 150 days. It is observed that SCM has unstable performance, while the Kronecker product estimator

offers better tracking of the wind speeds.

3.7 Conclusion

We have introduced a framework for covariance estimation based on separation rank decompositions using a series of Kronecker product factors. We proposed a least-squares estimator in a permuted linear space with nuclear norm penalization, named PRLS. We established high dimensional consistency for PRLS with guaranteed rates of convergence. The analysis shows that for low separation rank covariance models, our proposed method outperforms the standard SCM estimator. For the class of block-Toeplitz matrices with exponentially decaying off-diagonal norms, we showed that the separation rank is small, and specialized our convergence bounds to this class. We also presented synthetic simulations that showed the benefits of our methods.

As a real world application we demonstrated the performance of the proposed Kronecker product-based estimator in wind speed prediction using an Irish wind speed dataset and a recent US NCEP dataset. Implementation of a standard covariance-based prediction scheme using our Kronecker product estimator achieved performance gains as compared to standard with respect to previously proposed covariance-based predictors.

There are several questions that remain open and are worthy of additional study. First, while the proposed penalized least squares Kronecker sum approximation yields a unique solution, the solution requires specification of the parameter λ, which specifies both the separation rank, and the amount of spectral shrinkage in the approximation. It would be worthwhile to investigate optimal or consistent methods of choosing this regularization parameter, e.g. using Stein's theory of unbiased risk

minimization. Second, while we have proven positive definiteness of the Kronecker sum approximation when the number of samples is greater than than the variable dimension in our experiments we have observed that positive definiteness is preserved more generally. Maximum likelihood estimation of Kronecker sum covariance and inverse covariance matrices is a worthwhile open problem. Finally, extensions of the low separation rank estimation method (PRLS) developed here to missing data follow naturally through the methodology of low rank covariance estimation studied in [85].

3.8 Appendix

3.8.1 Proof of Theorem III.1

Proof. 1) Symmetry

Recall the permuted version of the sample covariance $\hat{\mathbf{S}}_n$, i.e., $\hat{\mathbf{R}}_n = \mathcal{R}(\hat{\mathbf{S}}_n)$. The SVD of $\hat{\mathbf{R}}_n$ can be obtained as a solution to the minimum norm problem (Thm. 1 and Cor. 2 in [84], Sec. 3 in [123]):

$$\min_{\{\mathbf{A}_k, \mathbf{B}_k\}_k} \| \hat{\mathbf{S}}_n - \sum_{k=1}^{r} \mathbf{A}_k \otimes \mathbf{B}_k \|_F^2 \tag{3.19}$$

subject to the orthogonality constraints $\mathrm{tr}(\mathbf{A}_k^T \mathbf{A}_l) = \mathrm{tr}(\mathbf{B}_k^T \mathbf{B}_l) = 0$ for $k \neq l$. Since the Frobenius norm is invariant to permutations, we have the equivalent optimization problem:

$$\min_{\{\mathbf{u}_k, \mathbf{v}_k\}_k} \| \hat{\mathbf{R}}_n - \sum_{k=1}^{r} \sigma_k \mathbf{u}_k \mathbf{v}_k^T \|_F^2 \tag{3.20}$$

subject to the orthonormality conditions $\mathbf{u}_k^T \mathbf{u}_l = \mathbf{v}_k^T \mathbf{v}_l = 1$ for $k = l$ and 0 if $k \neq l$. The correspondence of (3.19) with (3.20) is given by the mapping $\mathbf{u}_k = \mathrm{vec}(\mathbf{A}_k)$ and $\mathbf{v}_k = \sigma_k \mathrm{vec}(\mathbf{B}_k)$. The SVD of $\hat{\mathbf{R}}_n$ can be written in matrix form as $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$.

We next show that the symmetry of $\hat{\mathbf{S}}_n$ implies that the PRLS solution is symmetric by showing that the reshaped singular vectors $\mathbf{u}_k$ and $\mathbf{v}_k$ correspond to symmetric

matrices. From the SVD definition [59], the right singular vectors $\mathbf{v}_k$ are eigenvectors of $\mathbf{M}_n = \hat{\mathbf{R}}_n^T \hat{\mathbf{R}}_n$ and thus satisfy the eigenrelation:

$$\mathbf{M}_n \mathbf{v}_k = \sigma_k^2 \mathbf{v}_k \tag{3.21}$$

where $\sigma_k = [\mathbf{\Sigma}]_{k,k}$. Expressing (3.21) in terms of the permutation operator $\mathcal{R}$, we obtain:

$$\sum_{i,j=1}^{p} \left\langle \mathbf{v}_k, \mathrm{vec}(\hat{\mathbf{S}}_n(i,j)) \right\rangle \mathrm{vec}(\hat{\mathbf{S}}_n(i,j)) = \sigma_k^2 \mathbf{v}_k \tag{3.22}$$

Define the $q \times q$ matrix $\mathbf{V}_k$ such that $\mathbf{v}_k = \mathrm{vec}(\mathbf{V}_k)$. Rewriting (3.22) by reshaping vectors into matrices, we have after some algebra:

$$\begin{aligned}
\sigma_k^2 \mathbf{V}_k &= \sum_{i,j=1}^{p} \mathrm{tr}(\mathbf{V}_k^T \hat{\mathbf{S}}_n(i,j)) \hat{\mathbf{S}}_n(i,j) \\
&= \underbrace{\sum_{i=1}^{p} \mathrm{tr}(\mathbf{V}_k^T \hat{\mathbf{S}}_n(i,i)) \hat{\mathbf{S}}_n(i,i)}_{\mathbf{K}_1} + \underbrace{\sum_{i<j} \mathrm{tr}(\mathbf{V}_k^T \hat{\mathbf{S}}_n(i,j))(\hat{\mathbf{S}}_n(i,j) + \hat{\mathbf{S}}_n(j,i))}_{\mathbf{K}_2} \\
&\quad + \underbrace{\sum_{i<j} \mathrm{tr}(\mathbf{V}_k^T (\hat{\mathbf{S}}_n(j,i) - \hat{\mathbf{S}}_n(i,j))) \hat{\mathbf{S}}_n(j,i)}_{\mathbf{E}}
\end{aligned} \tag{3.23}$$

Clearly, $\mathbf{K}_1$ is symmetric since all submatrices $\hat{\mathbf{S}}_n(i,i)$ are symmetric. Since $\hat{\mathbf{S}}_n(j,i) = \hat{\mathbf{S}}_n(i,j)^T$, it follows that $\mathbf{K}_2$ is also symmetric. To finish the proof, we show $\mathbf{E} = 0$. Define the set

$$\begin{aligned}
\mathcal{L} = \Big\{ &(i,j) : i < j, \hat{\mathbf{S}}_n(i,j) \neq 0, \hat{\mathbf{S}}_n(i,j) \neq \hat{\mathbf{S}}_n(j,i), \\
&\hat{\mathbf{S}}_n(i,j) \neq \hat{\mathbf{S}}_n(i',j') \forall i' \neq i, j' \neq j \Big\}
\end{aligned}$$

The set $\mathcal{L}$ is nonempty with probability 1 for any sample size. Let $l = \mathrm{card}(\mathcal{L})$. Then, we can rewrite:

$$\mathbf{E} = \sum_{(i,j) \in \mathcal{L}} \mathrm{tr}(\mathbf{V}_k^T (\hat{\mathbf{S}}_n(j,i) - \hat{\mathbf{S}}_n(i,j))) \hat{\mathbf{S}}_n(j,i) \tag{3.24}$$

Since $\hat{\mathbf{S}}_n(j,i) \neq 0$ with probability 1, $\mathbf{E} = 0$ iff $\mathrm{tr}(\mathbf{V}_k^T(\hat{\mathbf{S}}_n(j,i) - \hat{\mathbf{S}}_n(i,j))) = 0$ for all $i < j$. Using the properties of the trace operator, rewriting $\mathrm{tr}(\mathbf{V}_k^T(\hat{\mathbf{S}}_n(j,i) - \hat{\mathbf{S}}_n(i,j))) = \mathrm{tr}((\mathbf{V}_k^T - \mathbf{V}_k)\hat{\mathbf{S}}_n(j,i))$, we conclude from the decomposition $\sigma_k^2 \mathbf{V}_k = \mathbf{K}_1 + \mathbf{K}_2 + \mathbf{E}$ that $\mathbf{V}_k = \mathbf{V}_k^T$ if $\mathbf{E} = 0$. To finish the proof, we show that $\mathbf{E} = 0$ with probability 1. Taking the $\mathrm{vec}(\cdot)$ of (3.24), we conclude that $\mathbf{E} = 0$ is equivalent to

$$0 = \sum_{(i,j)\in\mathcal{L}} a_{i,j} \hat{\mathbf{S}}_n(j,i) \tag{3.25}$$

where $a_{i,j} = \mathrm{tr}((\mathbf{V}_k^T - \mathbf{V}_k)\hat{\mathbf{S}}_n(j,i))$. The equation (3.25) can be rewritten as the linear equations:

$$\mathbf{D}\mathbf{a} = \mathbf{0} \tag{3.26}$$

where $\mathbf{a} = \{a_{i,j}\}_{(i,j)\in\mathcal{L}} \in \mathbb{R}^l$ and the columns of the $q^2 \times l$ matrix $\mathbf{D}$ are given by $\mathbf{d}_{i,j} = \mathrm{vec}(\hat{\mathbf{S}}_n(j,i)) \in \mathbb{R}^{q^2}$. Solutions of (3.26) are given by $\mathbf{a} \in \mathrm{Nul}(\mathbf{D})$. Since the matrix $\mathbf{D}$ is full-rank, $\mathbf{a} = \mathbf{0}$ is the only solution of (3.25). This implies $\mathbf{E} = 0$, and therefore, $\mathbf{V}_k = \mathbf{V}_k^T$. Since k is arbitrary, all reshaped right singular vectors of $\hat{\mathbf{R}}_n$ are symmetric. A similar argument holds for all reshaped left singular vectors $\mathbf{u}_k$. The proof is complete.

2) Positive Definiteness

The sample covariance matrix $\hat{\mathbf{S}}_n$ is positive definite with probability 1 if $n \geq pq$. First, consider the minimum norm problem (3.19). The factors $\mathbf{A}_k$ and $\mathbf{B}_k$ are symmetric by part (1). If we show that a solution to (3.19) has p.d. Kronecker factors, then the weighted sum with positive scalars is also p.d. and as a result, the PRLS solution given by $\hat{\mathbf{\Sigma}}_n^{\lambda} = \sum_{k=1}^{r_0} \left(\sigma_k(\hat{\mathbf{R}}_n) - \frac{\lambda}{2}\right)_+ \mathbf{U}_k \otimes \mathbf{V}_k$ is positive definite (see (3.7)).

Fix $l \in \{1, \ldots, r_0\}$. We will show that in (3.19) $\mathbf{A}_k$ and $\mathbf{B}_k$ can be restricted to

be p.d. matrices. Define the eigendecompositions of $\mathbf{A}_l$ and $\mathbf{B}_l$:

$$\mathbf{A}_l = \mathbf{\Psi}_l \mathbf{D}_l \mathbf{\Psi}_l^T$$

$$\mathbf{B}_l = \mathbf{\Xi}_l \mathbf{\Lambda}_l \mathbf{\Xi}_l^T$$

where $\{\mathbf{\Psi}_l\}_l, \{\mathbf{\Xi}_l\}_l$ are sets of orthonormal matrices and $\mathbf{D}_l, \mathbf{\Lambda}_l$ are diagonal matrices. Let $\mathbf{D}_l = \text{diag}(d_l^1, \ldots, d_l^p)$ and $\mathbf{\Lambda}_l = \text{diag}(\lambda_l^1, \ldots, \lambda_l^q)$. Set $\mathbf{Q}_l = \mathbf{\Psi}_l \otimes \mathbf{\Xi}_l$. Define $\mathbf{F}_l = \mathbf{Q}_l^T \hat{\mathbf{S}}_n \mathbf{Q}_l$. The objective function (3.19) can be rewritten as:

$$\begin{aligned}
(3.27) \qquad & \|\hat{\mathbf{S}}_n - \sum_{k=1}^{r} \mathbf{A}_k \otimes \mathbf{B}_k\|_F^2 \\
&= \|\mathbf{Q}_l^T \left(\hat{\mathbf{S}}_n - \sum_{k=1}^{r} \mathbf{A}_k \otimes \mathbf{B}_k \right) \mathbf{Q}_l\|_F^2 \\
&= \|\mathbf{F}_l - \sum_{k=1}^{r} \mathbf{Q}_l^T (\mathbf{A}_k \otimes \mathbf{B}_k) \mathbf{Q}_l\|_F^2 \\
&= \| \underbrace{\mathbf{F}_l - \sum_{k \neq l} (\mathbf{\Psi}_l^T \mathbf{A}_k \mathbf{\Psi}_l) \otimes (\mathbf{\Xi}_l^T \mathbf{B}_k \mathbf{\Xi}_l)}_{\mathbf{M}_l} \\
&\quad - (\mathbf{\Psi}_l^T \mathbf{A}_l \mathbf{\Psi}_l) \otimes (\mathbf{\Xi}_l^T \mathbf{B}_l \mathbf{\Xi}_l)\|_F^2 \\
&= \|\mathbf{M}_l - \mathbf{D}_l \otimes \mathbf{\Lambda}_l\|_F^2 \\
&= \|\mathbf{M}_l\|_F^2 + \|\mathbf{D}_l \otimes \mathbf{\Lambda}_l\|_F^2 - 2\text{tr}\left(\mathbf{F}_l (\mathbf{D}_l \otimes \mathbf{\Lambda}_l)\right) \\
&\quad + 2 \sum_{k \neq l} \text{tr}((\mathbf{\Psi}_l^T \mathbf{A}_k \mathbf{\Psi}_l \otimes \mathbf{\Xi}_l^T \mathbf{B}_k \mathbf{\Xi}_l)(\mathbf{D}_l \otimes \mathbf{\Lambda}_l)) \\
&= \|\mathbf{M}_l\|_F^2 - \|\mathbf{F}_l\|_F^2 + \|\mathbf{F}_l - \mathbf{D}_l \otimes \mathbf{\Lambda}_l\|_F^2 \\
&\quad + 2 \sum_{k \neq l} \text{tr}(\mathbf{B}_k \mathbf{B}_l) \text{tr}(\mathbf{A}_k \mathbf{A}_l) \\
&= \|\mathbf{M}_l\|_F^2 - \|\mathbf{F}_l\|_F^2 \\
(3.28) \qquad & \quad + \|\mathbf{F}_l - \text{diag}(\mathbf{F}_l) + \text{diag}(\mathbf{F}_l) - \mathbf{D}_l \otimes \mathbf{\Lambda}_l\|_F^2
\end{aligned}$$

$$
\begin{aligned}
&= \|\mathbf{M}_l\|_F^2 - \|\mathbf{F}_l\|_F^2 \\
&\quad + \|\mathbf{F}_l - \mathrm{diag}(\mathbf{F}_l)\|_F^2 + \|\mathrm{diag}(\mathbf{F}_l) - \mathbf{D}_l \otimes \mathbf{\Lambda}_l\|_F^2 \\
&\quad + 2\mathrm{tr}\left((\mathbf{F}_l - \mathrm{diag}(\mathbf{F}_l))(\mathrm{diag}(\mathbf{F}_l) - \mathbf{D}_l \otimes \mathbf{\Lambda}_l)\right) \\
&= \|\mathbf{M}_l\|_F^2 - \|\mathbf{F}_l\|_F^2 + \|\mathbf{F}_l - \mathrm{diag}(\mathbf{F}_l)\|_F^2 \\
&\quad + \|\mathrm{diag}(\mathbf{F}_l) - \mathbf{D}_l \otimes \mathbf{\Lambda}_l\|_F^2 \qquad (3.29)
\end{aligned}
$$

where in equality (3.28) we used the orthogonality of Kronecker factors in the SVD. In equality (3.29), we used the fact that the matrices $\mathbf{F}_l - \mathrm{diag}(\mathbf{F}_l)$ and $\mathrm{diag}(\mathbf{F}_l) - \mathbf{D}_l \otimes \mathbf{\Lambda}_l$ have disjoint support. We note that the term $\|\mathbf{M}_l\|_F^2 - \|\mathbf{F}_l\|_F^2 + \|\mathbf{F}_l - \mathrm{diag}(\mathbf{F}_l)\|_F^2$ is independent of $\mathbf{D}_l, \mathbf{\Lambda}_l$. The positive definiteness of $\hat{\mathbf{S}}_n$ implies that the diagonal elements of $\mathbf{F}_l$ are all positive. Let $\mathrm{diag}(\mathbf{F}_l) = \mathrm{diag}(\{f_{(i-1)q+j}\}_{i,j}) > 0$. Simple algebra yields:

$$
\begin{aligned}
&\|\mathrm{diag}(\mathbf{F}_l) - \mathbf{D}_l \otimes \mathbf{\Lambda}_l\|_F^2 \\
&\quad = \sum_{i=1}^{p}\sum_{j=1}^{q}(f_{(i-1)q+j} - d_l^i \lambda_l^j)^2 = a_l + b_l
\end{aligned}
$$

where

$$
\begin{aligned}
a_l &= \sum_{i=1}^{p}\sum_{j=1}^{q}(f_{(i-1)q+j} - |d_l^i||\lambda_l^j|)^2 \\
b_l &= 2\sum_{i=1}^{p}\sum_{j=1}^{q} f_{(i-1)q+j}(|d_l^i||\lambda_l^j| - d_l^i \lambda_l^j)
\end{aligned}
$$

We note that the term a_l is invariant to any sign changes of the eigenvalues $\{d_l^i, \lambda_l^j\}_{i,j}$ and the term b_l is non-negative and equals zero iff d_l^i, λ_l^j have the same sign for all i, j. By contradiction, it follows that the eigenvalues $\{d_l^i\}_{i=1}^p$ and $\{\lambda_l^j\}_{j=1}^q$ must all have the same sign (if not, then the minimum norm is not achieved by $(\mathbf{A}_l, \mathbf{B}_l)$). Without loss of generality (since $\mathbf{A}_l \otimes \mathbf{B}_l = (-\mathbf{A}_l) \otimes (-\mathbf{B}_l)$, the signs can be assumed to

be positive. We conclude that there exist p.d. matrices $(\mathbf{A}_l, \mathbf{B}_l)$ that achieve the minimum norm of (3.27). This holds for any l so the proof is complete.

□

3.8.2 Proof of Theorem III.2

Proof. The proof generalizes Thm. 1 in [85] to nonsquare matrices. A necessary and sufficient condition for the minimizer of (3.5) is that there exists a $\hat{\mathbf{V}} \in \partial\|\hat{\mathbf{R}}^\lambda\|_*$ such that:

$$\left\langle 2(\hat{\mathbf{R}}^\lambda - \hat{\mathbf{R}}_n) + \lambda\hat{\mathbf{V}}, \hat{\mathbf{R}}^\lambda - \mathbf{R} \right\rangle \leq 0 \tag{3.30}$$

for all $\mathbf{R}$. From (3.30), we obtain for any $\mathbf{V} \in \partial\|\mathbf{R}\|_1$:

$$\begin{aligned} &2\left\langle \hat{\mathbf{R}}^\lambda - \mathbf{R}_0, \hat{\mathbf{R}}^\lambda - \mathbf{R} \right\rangle + \lambda\left\langle \hat{\mathbf{V}} - \mathbf{V}, \hat{\mathbf{R}}^\lambda - \mathbf{R} \right\rangle \\ &\leq -\lambda\left\langle \mathbf{V}, \hat{\mathbf{R}}^\lambda - \mathbf{R} \right\rangle + 2\left\langle \hat{\mathbf{R}}_n - \mathbf{R}_0, \hat{\mathbf{R}}^\lambda - \mathbf{R} \right\rangle \end{aligned} \tag{3.31}$$

The monotonicity of subdifferentials of convex functions implies:

$$\left\langle \hat{\mathbf{V}} - \mathbf{V}, \hat{\mathbf{R}}^\lambda - \mathbf{R} \right\rangle \geq 0 \tag{3.32}$$

From Example 2 in [129], we have the characterization of the subdifferential of a nuclear norm of a nonsquare matrix:

$$\partial\|\mathbf{R}\|_* = \left\{ \sum_{j=1}^{r} \mathbf{u}_j(\mathbf{R})\mathbf{v}_j(\mathbf{R})^T + \mathbf{P}_U^\perp \mathbf{W} \mathbf{P}_V^\perp : \|\mathbf{W}\|_2 \leq 1 \right\}$$

where $r = \text{rank}(\mathbf{R})$, $U = \text{span}\{\mathbf{u}_j\}$ and $V = \text{span}\{\mathbf{v}_j\}$. Thus, for $\mathbf{R} = \sum_{j=1}^{r} \sigma_j(\mathbf{R})\mathbf{u}_j\mathbf{v}_j^T$, $r = \text{rank}(\mathbf{R})$, we can write:

$$\mathbf{V} = \sum_{j=1}^{r} \mathbf{u}_j\mathbf{v}_j^T + \mathbf{P}_U^\perp \mathbf{W} \mathbf{P}_V^\perp \tag{3.33}$$

where $\mathbf{W}$ can be chosen such that $\|\mathbf{W}\|_2 \leq 1$ and

$$\left\langle \mathbf{P}_U^\perp \mathbf{W} \mathbf{P}_V^\perp, \hat{\mathbf{R}}^\lambda - \mathbf{R} \right\rangle = \|\mathbf{P}_U^\perp \hat{\mathbf{R}}^\lambda \mathbf{P}_V^\perp\|_* \tag{3.34}$$

Next, note the equality:

$$\begin{aligned} &\|\hat{\mathbf{R}}^\lambda - \mathbf{R}_0\|_F^2 + \|\hat{\mathbf{R}}^\lambda - \mathbf{R}\|_F^2 - \|\mathbf{R} - \mathbf{R}_0\|_F^2 \\ &\qquad = 2\left\langle \hat{\mathbf{R}}^\lambda - \mathbf{R}_0, \hat{\mathbf{R}}^\lambda - \mathbf{R} \right\rangle \end{aligned} \tag{3.35}$$

Using (3.32), (3.34) and (3.35) in (3.31), we obtain:

$$\begin{aligned} &\|\hat{\mathbf{R}}^\lambda - \mathbf{R}_0\|_F^2 + \|\hat{\mathbf{R}}^\lambda - \mathbf{R}\|_F^2 + \lambda\|\mathbf{P}_U^\perp \hat{\mathbf{R}}^\lambda \mathbf{P}_V^\perp\|_* \\ &\leq \|\mathbf{R} - \mathbf{R}_0\|_F^2 + \lambda\left\langle \sum_{j=1}^r \mathbf{u}_j \mathbf{v}_j^T, -(\hat{\mathbf{R}}^\lambda - \mathbf{R}) \right\rangle \\ &\quad + 2\left\langle \hat{\mathbf{R}}_n - \mathbf{R}_0, \hat{\mathbf{R}}^\lambda - \mathbf{R} \right\rangle \end{aligned} \tag{3.36}$$

From trace duality, we have:

$$\begin{aligned} &\left\langle \sum_{j=1}^r \mathbf{u}_j \mathbf{v}_j^T, -(\hat{\mathbf{R}}^\lambda - \mathbf{R}) \right\rangle \\ &= \left\langle \mathbf{P}_U \sum_{j=1}^r \mathbf{u}_j \mathbf{v}_j^T \mathbf{P}_V, -(\hat{\mathbf{R}}^\lambda - \mathbf{R}) \right\rangle \\ &\leq \|\sum_{j=1}^r \mathbf{u}_j \mathbf{v}_j^T\|_2 \|\mathbf{P}_U^T(\hat{\mathbf{R}}^\lambda - \mathbf{R})\mathbf{P}_V^T\|_* \\ &= \|\mathbf{P}_U(\hat{\mathbf{R}}^\lambda - \mathbf{R})\mathbf{P}_V\|_* \end{aligned}$$

where we used the symmetry of projection matrices. Using this bound in (3.36), we obtain:

$$\begin{aligned} &\|\hat{\mathbf{R}}^\lambda - \mathbf{R}_0\|_F^2 + \|\hat{\mathbf{R}}^\lambda - \mathbf{R}\|_F^2 + \lambda\|\mathbf{P}_U^\perp \hat{\mathbf{R}}^\lambda \mathbf{P}_V^\perp\|_* \\ &\leq \|\mathbf{R} - \mathbf{R}_0\|_F^2 + \lambda\|\mathbf{P}_U(\hat{\mathbf{R}}^\lambda - \mathbf{R})\mathbf{P}_V\|_* \\ &\quad + 2\left\langle \mathbf{\Delta}_n, \hat{\mathbf{R}}^\lambda - \mathbf{R} \right\rangle \end{aligned} \tag{3.37}$$

where $\boldsymbol{\Delta}_n = \hat{\mathbf{R}}_n - \mathbf{R}_0$. Define the orthogonal projection of $\mathbf{R}$ onto the outer product span of U and V as $\mathcal{P}_{U,V}(\mathbf{R}) = \mathbf{R} - \mathbf{P}_U^{\perp}\mathbf{R}\mathbf{P}_V^{\perp}$. Then, we decompose:

$$\begin{aligned}\left\langle \boldsymbol{\Delta}_n, \hat{\mathbf{R}}^{\lambda} - \mathbf{R} \right\rangle &= \left\langle \boldsymbol{\Delta}_n, \mathcal{P}_{U,V}\left(\hat{\mathbf{R}}^{\lambda} - \mathbf{R}\right) \right\rangle \\ &\quad + \left\langle \boldsymbol{\Delta}_n, \mathbf{P}_U^{\perp}(\hat{\mathbf{R}}^{\lambda} - \mathbf{R})\mathbf{P}_V^{\perp} \right\rangle\end{aligned}$$

By the Cauchy-Schwarz inequality and trace-duality:

$$\begin{aligned}\|\mathbf{P}_U(\hat{\mathbf{R}}^{\lambda} - \mathbf{R})\mathbf{P}_V\|_* &\leq \sqrt{\text{rank}(\mathbf{R})}\|\hat{\mathbf{R}}^{\lambda} - \mathbf{R}\|_F \\ |\left\langle \boldsymbol{\Delta}_n, \mathcal{P}_{U,V}(\hat{\mathbf{R}}^{\lambda} - \mathbf{R}) \right\rangle| &\leq \|\boldsymbol{\Delta}_n\|_2 \|\mathcal{P}_{U,V}(\hat{\mathbf{R}}^{\lambda} - \mathbf{R})\|_* \\ &\leq \|\boldsymbol{\Delta}_n\|_2 \sqrt{2\text{rank}(\mathbf{R})}\|\hat{\mathbf{R}}^{\lambda} - \mathbf{R}\|_F \\ |\left\langle \boldsymbol{\Delta}_n, \mathbf{P}_U^{\perp}(\hat{\mathbf{R}}^{\lambda} - \mathbf{R})\mathbf{P}_V^{\perp} \right\rangle| &\leq \|\boldsymbol{\Delta}_n\|_2 \|\mathbf{P}_U^{\perp}\hat{\mathbf{R}}^{\lambda}\mathbf{P}_V^{\perp}\|_*\end{aligned}$$

where we used $\mathbf{P}_U^{\perp}\mathbf{R}\mathbf{P}_V^{\perp} = 0$. Using these bounds in (3.37), we further obtain:

$$\begin{aligned}&\|\hat{\mathbf{R}}^{\lambda} - \mathbf{R}_0\|_F^2 + \|\hat{\mathbf{R}}^{\lambda} - \mathbf{R}\|_F^2 + (\lambda - 2\|\boldsymbol{\Delta}_n\|_2)\|\mathbf{P}_U^{\perp}\hat{\mathbf{R}}^{\lambda}\mathbf{P}_V^{\perp}\|_* \\ &\leq \|\mathbf{R} - \mathbf{R}_0\|_F^2 + ((2\sqrt{2}\|\boldsymbol{\Delta}_n\|_2 + \lambda)\sqrt{r})(\sqrt{\|\hat{\mathbf{R}}^{\lambda} - \mathbf{R}\|_F^2}) \qquad (3.38)\end{aligned}$$

Using the arithmetic-mean geometric-mean inequality in the RHS of (3.38) and the assumption $\lambda \geq 2\|\boldsymbol{\Delta}_n\|_2$, we obtain:

$$\|\hat{\mathbf{R}}^{\lambda} - \mathbf{R}_0\|_F^2 \leq \|\mathbf{R} - \mathbf{R}_0\|_F^2 + \frac{\lambda^2(1+\sqrt{2})^2}{4}r$$

This concludes the proof. □

3.8.3 Lemma III.7

Lemma III.7. *(Concentration of Measure for Coupled Gaussian Chaos) Let* $\mathbf{x} = [x_1, \ldots, x_{p^2}]^T \in \mathcal{S}_{p^2-1}$ *and* $\mathbf{y} = [y_1, \ldots, y_{q^2}]^T \in \mathcal{S}_{q^2-1}$. *In the SCM (3.2) assume that* $\{\mathbf{z}_t\}$ *are i.i.d. multivariate normal* $\mathbf{z}_t \sim N(0, \boldsymbol{\Sigma}_0)$. *Recall* $\boldsymbol{\Delta}_n$ *in (3.10). For all* $\tau \geq 0$*:*

$$\mathbb{P}(|\mathbf{x}^T\boldsymbol{\Delta}_n\mathbf{y}| \geq \tau) \leq 2\exp\left(\frac{-n\tau^2/2}{C_1\|\boldsymbol{\Sigma}_0\|_2^2 + C_2\|\boldsymbol{\Sigma}_0\|_2\tau}\right) \qquad (3.39)$$

where $C_1 = \frac{4e}{\sqrt{6\pi}} \approx 2.5044$ *and* $C_2 = e\sqrt{2} \approx 3.8442$ *are absolute constants.*

Proof. This proof is based on concentration of measure for Gaussian matrices and is similar to proof techniques used in compressed sensing (see Appendix A in [98]). Note that by the definition of the reshaping permutation operator $\mathcal{R}(\cdot)$, we have:

$$\boldsymbol{\Delta}_n = \frac{1}{n}\sum_{t=1}^{n} \begin{bmatrix} \text{vec}(\mathbf{z}_t(1)\mathbf{z}_t(1)^T)^T - \mathbb{E}[\text{vec}(\mathbf{z}_t(1)\mathbf{z}_t(1)^T)^T] \\ \vdots \\ \text{vec}(\mathbf{z}_t(p)\mathbf{z}_t(p)^T)^T - \mathbb{E}[\text{vec}(\mathbf{z}_t(p)\mathbf{z}_t(p)^T)^T] \end{bmatrix}$$

where $\mathbf{z}_t(i) = [\mathbf{z}_t]_{(i-1)q+1:iq}$ is the ith subvector of the tth observation $\mathbf{z}_t$. Thus, we can write:

$$\mathbf{x}^T\boldsymbol{\Delta}_n\mathbf{y} = \frac{1}{n}\sum_{t=1}^{n}\psi_t$$

where

$$\psi_t = \sum_{i,j=1}^{p}\sum_{k,l=1}^{q}\mathbf{X}_{i,j}\mathbf{Y}_{k,l}$$

$$\times\left([\mathbf{z}_t]_{(i-1)q+k}[\mathbf{z}_t]_{(j-1)q+l} - \mathbb{E}[[\mathbf{z}_t]_{(i-1)q+k}[\mathbf{z}_t]_{(j-1)q+l}]\right) \tag{3.40}$$

and $\mathbf{X} \in \mathbb{R}^{p\times p}$ and $\mathbf{Y} \in \mathbb{R}^{q\times q}$ are reshaped versions of $\mathbf{x}$ and $\mathbf{y}$. Defining $\mathbf{M} = \mathbf{X}\otimes\mathbf{Y}$, we can write (3.40) as:

$$\psi_t = \mathbf{z}_t^T\mathbf{M}\mathbf{z}_t - \mathbb{E}[\mathbf{z}_t^T\mathbf{M}\mathbf{z}_t]$$

The statistic (3.40) has the form of Gaussian chaos of order 2 [81]. Many of the random variables involved in the summation (3.40) are correlated, which makes the analysis difficult. To simplify the concentration of measure derivation, using the joint Gaussian property of the data, we note that a stochastic equivalent of $\mathbf{z}_t^T\mathbf{M}\mathbf{z}_t$ is $\boldsymbol{\beta}_t^T\tilde{\mathbf{M}}\boldsymbol{\beta}_t$, where $\tilde{\mathbf{M}} = \boldsymbol{\Sigma}_0^{1/2}\mathbf{M}\boldsymbol{\Sigma}_0^{1/2}$, and $\boldsymbol{\beta}_t \sim N(\mathbf{0}, \mathbf{I}_{pq})$ is a random vector with i.i.d.

standard normal components. With this decoupling, we have:

$$\begin{aligned}
\mathbb{E}|\psi_t|^2 &= \mathbb{E}\left|\boldsymbol{\beta}_t^T\tilde{\mathbf{M}}\boldsymbol{\beta}_t - \mathbb{E}[\boldsymbol{\beta}_t^T\tilde{\mathbf{M}}\boldsymbol{\beta}_t]\right|^2 \\
&= \mathbb{E}\left|\sum_{i_1\neq i_2}[\boldsymbol{\beta}_t]_{i_1}[\boldsymbol{\beta}_t]_{i_2}\tilde{\mathbf{M}}_{i_1,i_2} + \sum_{i_1=1}^{d}([\boldsymbol{\beta}_t]_{i_1}^2-1)\tilde{\mathbf{M}}_{i_1,i_1}\right|^2 \\
&= \sum_{i_1\neq i_2}\sum_{i_1'\neq i_2'}\mathbb{E}[[\boldsymbol{\beta}_t]_{i_1}[\boldsymbol{\beta}_t]_{i_2}[\boldsymbol{\beta}_t]_{i_1'}[\boldsymbol{\beta}_t]_{i_2'}]\tilde{\mathbf{M}}_{i_1,i_2}\tilde{\mathbf{M}}_{i_1',i_2'} \\
&\quad + \sum_{i_1}\sum_{i_1'}\mathbb{E}[([\boldsymbol{\beta}_t]_{i_1}^2-1)([\boldsymbol{\beta}_t]_{i_1'}^2-1)]\tilde{\mathbf{M}}_{i_1,i_1}\tilde{\mathbf{M}}_{i_1',i_1'} \\
&= \sum_{i_1\neq i_2}\tilde{\mathbf{M}}_{i_1,i_2}^2 + 2\sum_{i_1}\tilde{\mathbf{M}}_{i_1,i_1}^2 \\
&= \|\tilde{\mathbf{M}}\|_F^2 + \|\mathrm{diag}(\tilde{\mathbf{M}})\|_F^2 \\
&\leq 2\|\tilde{\mathbf{M}}\|_F^2 \leq 2\|\boldsymbol{\Sigma}_0\|_2^2\|\mathbf{M}\|_F^2 = 2\|\boldsymbol{\Sigma}_0\|_2^2
\end{aligned}$$

where in the last step we used $\|\mathbf{M}\|_F = \|\mathbf{X}\|_F\|\mathbf{Y}\|_F = 1$.

Using a well known moment bound on Gaussian chaos (see p. 65 in [81]) and Stirling's formula, it can be shown (see, for example, Appendix A in [98]) that for all $m \geq 3$:

$$\mathbb{E}|\psi_t|^m \leq m!W^{m-2}v_t/2 \tag{3.41}$$

where

$$\begin{aligned}
W &= e\sqrt{\mathbb{E}|\psi_t|^2} \leq e\sqrt{2}\|\boldsymbol{\Sigma}_0\|_2 \\
v_t &= \frac{2e}{\sqrt{6\pi}}\mathbb{E}|\psi_t|^2 \leq \frac{4e}{\sqrt{6\pi}}\|\boldsymbol{\Sigma}_0\|_2^2
\end{aligned}$$

From Bernstein's inequality (see Thm. 1.1 in [98]), we obtain:

$$\begin{aligned}
\mathbb{P}\left(\left|\frac{1}{n}\sum_{t=1}^{n}\psi_t\right| \geq \tau\right) &\leq 2\exp\left(\frac{-n^2\tau^2/2}{nv_1 + Wn\tau}\right) \\
&\leq 2\exp\left(\frac{-n\tau^2/2}{C_1\|\boldsymbol{\Sigma}_0\|_2^2 + C_2\|\boldsymbol{\Sigma}_0\|_2\tau}\right)
\end{aligned}$$

This concludes the proof.

□

3.8.4 Proof of Theorem III.3

Proof. Let $\mathcal{N}(\mathcal{S}_{d'-1}, \epsilon')$ denote an ϵ'-net on the d'-dimensional sphere $\mathcal{S}_{d'-1}$. Let $\mathbf{x}_1 \in \mathcal{S}_{p^2-1}$ and $\mathbf{y}_1 \in \mathcal{S}_{q^2-1}$ be such that $|\mathbf{x}_1^T \boldsymbol{\Delta}_n \mathbf{y}_1| = \|\boldsymbol{\Delta}_n\|_2$. By the definition of ϵ'-net, there exists $\mathbf{x}_2 \in \mathcal{N}(\mathcal{S}_{p^2-1}, \epsilon')$ and $\mathbf{y}_2 \in \mathcal{N}(\mathcal{S}_{q^2-1}, \epsilon')$ such that $\|\mathbf{x}_1 - \mathbf{x}_2\|_2 \leq \epsilon'$ and $\|\mathbf{y}_1 - \mathbf{y}_2\|_2 \leq \epsilon'$. Then, by the Cauchy-Schwarz inequality:

$$\begin{aligned}
&|\mathbf{x}_1^T \boldsymbol{\Delta}_n \mathbf{y}_1| - |\mathbf{x}_2^T \boldsymbol{\Delta}_n \mathbf{y}_2| \leq |\mathbf{x}_1^T \boldsymbol{\Delta}_n \mathbf{y}_1 - \mathbf{x}_2^T \boldsymbol{\Delta}_n \mathbf{y}_2| \\
&= |\mathbf{x}_1^T \boldsymbol{\Delta}_n (\mathbf{y}_1 - \mathbf{y}_2) > + (\mathbf{x}_1 - \mathbf{x}_2)^T \boldsymbol{\Delta}_n \mathbf{y}_2 > | \\
&\leq 2\epsilon' \|\boldsymbol{\Delta}_n\|_2
\end{aligned}$$

Since $\|\boldsymbol{\Delta}_n\|_2 = |\mathbf{x}_1^T \boldsymbol{\Delta}_n \mathbf{y}_1|$, this implies:

$$\begin{aligned}
&\|\boldsymbol{\Delta}_n\|_2 (1 - 2\epsilon') \\
&\leq \max \Big\{ |\mathbf{x}_2^T \boldsymbol{\Delta}_n \mathbf{y}_2| : \mathbf{x}_2 \in \mathcal{N}(\mathcal{S}_{p^2-1}, \epsilon'), \mathbf{y}_2 \in \mathcal{N}(\mathcal{S}_{q^2-1}, \epsilon'), \\
&\qquad \|\mathbf{x}_1 - \mathbf{x}_2\|_2 \leq \epsilon', \|\mathbf{y}_1 - \mathbf{y}_2\|_2 \leq \epsilon' \Big\} \\
&\leq \max \Big\{ |\mathbf{x}^T \boldsymbol{\Delta}_n \mathbf{y}| : \mathbf{x} \in \mathcal{N}(S^{p^2-1}, \epsilon'), \mathbf{y} \in \mathcal{N}(S^{q^2-1}, \epsilon') \Big\}
\end{aligned}$$

As a result,

$$\|\boldsymbol{\Delta}_n\|_2 \leq (1 - 2\epsilon')^{-1} \max_{\mathbf{x} \in \mathcal{N}(\mathcal{S}_{p^2-1}, \epsilon'), \mathbf{y} \in \mathcal{N}(\mathcal{S}_{q^2-1}, \epsilon')} |\mathbf{x}^T \boldsymbol{\Delta}_n \mathbf{y}| \tag{3.42}$$

From Lemma 5.2 in [124], we have the bound on the cardinality of the ϵ'-net:

$$\text{card}(\mathcal{N}(\mathcal{S}_{d'-1}, \epsilon')) \leq \left(1 + \frac{2}{\epsilon'}\right)^{d'}. \tag{3.43}$$

From (3.42), (3.43) and the union bound:

$$\begin{aligned}
&\mathbb{P}(\|\mathbf{\Delta}_n\|_2 \geq \epsilon) \\
&\quad \leq \mathbb{P}\left(\max_{\mathbf{x}\in\mathcal{N}(\mathcal{S}_{p^2-1},\epsilon'),\mathbf{y}\in\mathcal{N}(\mathcal{S}_{q^2-1},\epsilon')} |\mathbf{x}^T\mathbf{\Delta}_n\mathbf{y}| \geq \epsilon(1-2\epsilon')\right) \\
&\quad \leq \mathbb{P}\left(\bigcup_{\mathbf{x}\in\mathcal{N}(\mathcal{S}_{p^2-1},\epsilon'),\mathbf{y}\in\mathcal{N}(\mathcal{S}_{q^2-1},\epsilon')} |\mathbf{x}^T\mathbf{\Delta}_n\mathbf{y}| \geq \epsilon(1-2\epsilon')\right) \\
&\quad \leq \mathrm{card}(\mathcal{N}(\mathcal{S}_{p^2-1},\epsilon'))\mathrm{card}(\mathcal{N}(\mathcal{S}_{q^2-1},\epsilon')) \\
&\qquad \times \max_{\mathbf{x}\in\mathcal{N}(\mathcal{S}_{p^2-1},\epsilon'),\mathbf{y}\in\mathcal{N}(\mathcal{S}_{q^2-1},\epsilon')} \mathbb{P}(|\mathbf{x}^T\mathbf{\Delta}_n\mathbf{y}| \geq \epsilon(1-2\epsilon')) \\
&\quad \leq \left(1+\frac{2}{\epsilon'}\right)^{p^2+q^2} \mathbb{P}\left(|\mathbf{x}^T\mathbf{\Delta}_n\mathbf{y}| \geq \epsilon(1-2\epsilon')\right)
\end{aligned}$$

Using Lemma III.7, we further obtain:

$$\begin{aligned}
&\mathbb{P}(\|\mathbf{\Delta}_n\|_2 \geq \epsilon) \\
&\quad \leq 2\left(1+\frac{2}{\epsilon'}\right)^{p^2+q^2} \exp\left(\frac{-n\epsilon^2(1-2\epsilon')^2/2}{C_1\|\mathbf{\Sigma}_0\|_2^2 + C_2\|\mathbf{\Sigma}_0\|_2\epsilon(1-2\epsilon')}\right) \qquad (3.44)
\end{aligned}$$

We finish the proof by considering the two separate regimes. First, let us consider the Gaussian tail regime which occurs when $\epsilon \leq \frac{C_1\|\mathbf{\Sigma}_0\|_2}{C_2(1-2\epsilon')}$. For this regime, the bound (3.44) can be relaxed to:

$$\begin{aligned}
&\mathbb{P}(\|\mathbf{\Delta}_n\|_2 \geq \epsilon) \\
&\quad \leq 2\left(1+\frac{2}{\epsilon'}\right)^{p^2+q^2} \exp\left(\frac{-n\epsilon^2(1-2\epsilon')^2/2}{2C_1\|\mathbf{\Sigma}_0\|_2^2}\right) \qquad (3.45)
\end{aligned}$$

Let us choose:

$$\epsilon = \frac{t\|\mathbf{\Sigma}_0\|_2}{1-2\epsilon'}\sqrt{\frac{p^2+q^2+\log M}{n}}$$

Then, from (3.45), we have:

$$\begin{aligned}
&\mathbb{P}\left(\|\boldsymbol{\Delta}_n\|_2 \geq \frac{t\|\boldsymbol{\Sigma}_0\|_2}{1-2\epsilon'}\sqrt{\frac{p^2+q^2+\log M}{n}}\right) \\
&\leq 2\left(1+\frac{2}{\epsilon'}\right)^{p^2+q^2} \exp\left(\frac{-t^2(p^2+q^2+\log M)}{4C_1}\right) \\
&\leq 2\left(\left(1+\frac{2}{\epsilon'}\right)e^{-t^2/(4C_1)}\right)^{p^2+q^2} M^{-t^2/(4C_1)} \\
&\leq 2M^{-t^2/(4C_1)}
\end{aligned}$$

This concludes the bound for the Gaussian tail regime. The exponential tail regime follows by similar arguments. Assuming $\epsilon \geq \frac{C_1\|\boldsymbol{\Sigma}_0\|_2}{C_2(1-2\epsilon')}$, and setting $\epsilon = \frac{t\|\boldsymbol{\Sigma}_0\|_2}{1-2\epsilon'}\frac{p^2+q^2+\log M}{n}$, we obtain from (3.44):

$$\begin{aligned}
&\mathbb{P}\left(\|\boldsymbol{\Delta}_n\|_2 \geq \frac{t\|\boldsymbol{\Sigma}_0\|_2}{1-2\epsilon'}\frac{p^2+q^2+\log M}{n}\right) \\
&\leq 2\left(1+\frac{2}{\epsilon'}\right)^{p^2+q^2} \exp\left(\frac{-t(p^2+q^2+\log M)}{4C_2}\right) \\
&\leq 2\left(\left(1+\frac{2}{\epsilon'}\right)e^{-t/(4C_2)}\right)^{p^2+q^2} M^{-t/(4C_2)} \\
&\leq 2M^{-t/(4C_2)}
\end{aligned}$$

where we used the assumption $t \geq 4C_2\ln(1+\frac{2}{\epsilon'})$. The proof is completed by combining both regimes and letting $C_0 = \|\boldsymbol{\Sigma}_0\|_2$ and noting that $t > 1$, along with $\frac{tC_2}{C_1} > 1$.

□

3.8.5 Proof of Theorem III.4

Proof. Define the event

$$\mathcal{E}_r = \left\{\|\hat{\mathbf{R}}_n^{\lambda} - \mathbf{R}_0\|_F^2 > \inf_{\mathbf{R}:\mathrm{rank}(\mathbf{R})\leq r} \|\mathbf{R}-\mathbf{R}_0\|_F^2 + \frac{(1+\sqrt{2})^2}{4}\lambda_n^2 r\right\}$$

where λ_n is chosen as in the statement of the theorem.

Theorem III.2 implies that on the event $\lambda_n \geq 2\|\mathbf{\Delta}_n\|_2$, with probability 1, we have for any $1 \leq r \leq r_0$:

$$\|\hat{\mathbf{R}}_n^{\lambda} - \mathbf{R}_0\|_F^2 \leq \inf_{\mathbf{R}:\mathrm{rank}(\mathbf{R})\leq r} \|\mathbf{R} - \mathbf{R}_0\|_F^2 + \frac{(1+\sqrt{2})^2}{4}\lambda_n^2 r$$

Using this and Theorem III.3, we obtain:

$$\begin{aligned}
\mathbb{P}\left(\mathcal{E}_r\right) &= \mathbb{P}\left(\mathcal{E}_r \cap \{\lambda_n \geq 2\|\mathbf{\Delta}_n\|_2\}\right) + \mathbb{P}\left(\mathcal{E}_r \cap \{\lambda_n < 2\|\mathbf{\Delta}_n\|_2\}\right) \\
&\leq \cancelto{0}{\mathbb{P}(\mathcal{E}_r|\lambda_n \geq 2\|\mathbf{\Delta}_n\|_2)}\mathbb{P}(\lambda_n \geq 2\|\mathbf{\Delta}_n\|_2) \\
&\quad + \mathbb{P}\left(\lambda_n < 2\|\mathbf{\Delta}_n\|_2\right) \\
&= \mathbb{P}\Big(\|\mathbf{\Delta}_n\|_2 > \frac{C_0 t}{1-2\epsilon'} \\
&\quad \times \max\left\{\frac{p^2+q^2+\log M}{n}, \sqrt{\frac{p^2+q^2+\log M}{n}}\right\}\Big) \\
&\leq 2M^{-t/4C}
\end{aligned}$$

This concludes the proof.

□

3.8.6 Proof of Lemma III.5

Proof. From the min-max theorem of Courant-Fischer-Weyl [64]:

$$\begin{aligned}
\sigma_{k+1}^2(\mathbf{R}) &= \lambda_{k+1}(\mathbf{R}\mathbf{R}^T) \\
&= \min_{\mathcal{V}:\dim(\mathcal{V}^\perp)\leq k} \max_{\|\mathbf{v}\|_2=1, \mathbf{v}\in\mathcal{V}} \left\langle \mathbf{R}\mathbf{R}^T\mathbf{v}, \mathbf{v}\right\rangle
\end{aligned}$$

Define the set

$$\mathcal{V}_k = \{\mathbf{v} \in \mathbb{R}^{p^2} : \|\mathbf{v}\|_2 = 1, \mathbf{v} \perp \mathrm{Col}(\mathbf{R}\mathbf{P}_k\mathbf{R}^T)\} \subset S^{p^2-1}.$$

Choosing $\mathcal{V} = \mathrm{Col}(\mathbf{R}\mathbf{P}_k\mathbf{R}^T)^\perp$, we have the upper bound:

$$\sigma_{k+1}^2(\mathbf{R}) \leq \max_{\mathbf{v}\in\mathcal{V}_k} \left\langle \mathbf{R}\mathbf{R}^T\mathbf{v}, \mathbf{v}\right\rangle$$

Using the definition of $\mathcal{V}_k$ and the orthogonality principle, we have:

$$\begin{aligned}\left\langle \mathbf{R}\mathbf{R}^T\mathbf{v}, \mathbf{v}\right\rangle &= \left\langle \mathbf{R}(\mathbf{I}-\mathbf{P}_k)\mathbf{R}^T\mathbf{v}, \mathbf{v}\right\rangle \\ &= \left\langle (\mathbf{I}-\mathbf{P}_k)\mathbf{R}^T\mathbf{v}, \mathbf{R}^T\mathbf{v}\right\rangle \\ &= \left\langle (\mathbf{I}-\mathbf{P}_k)\mathbf{R}^T\mathbf{v}, (\mathbf{I}-\mathbf{P}_k)\mathbf{R}^T\mathbf{v}\right\rangle \\ &= \|(\mathbf{I}-\mathbf{P}_k)\mathbf{R}^T\mathbf{v}\|_2^2\end{aligned}$$

Using this equality and the definition of the spectral norm [64]:

$$\begin{aligned}\sigma_{k+1}^2(\mathbf{R}) &\leq \max_{\mathbf{v}\in\mathcal{V}_k} \|(\mathbf{I}-\mathbf{P}_k)\mathbf{R}^T\mathbf{v}\|_2^2 \\ &\leq \max_{\mathbf{v}\in S_{p^2-1}} \|(\mathbf{I}-\mathbf{P}_k)\mathbf{R}^T\mathbf{v}\|_2^2 \\ &= \|(\mathbf{I}-\mathbf{P}_k)\mathbf{R}^T\|_2^2\end{aligned}$$

Equality follows when choosing $\mathbf{P}_k = \mathbf{V}_k\mathbf{V}_k^T$. This is seen by writing $\mathbf{I} = \mathbf{V}\mathbf{V}^T$ and using the definition of the spectral norm and the sorting of the singular values. The proof is complete. □

3.8.7 Proof of Theorem III.6

Proof. Note that $(\lambda, \mathbf{u})$ is an eigenvalue-eigenvector pair of the square symmetric matrix $\mathbf{R}_0^T\mathbf{R}_0$ if:

$$\sum_{i,j} \mathrm{vec}(\mathbf{\Sigma}_0(i,j))\left\langle \mathbf{u}, \mathrm{vec}(\mathbf{\Sigma}_0(i,j))\right\rangle = \lambda\mathbf{u} \tag{3.46}$$

So for $\lambda > 0$, the eigenvector $\mathbf{u}$ must lie in the span of the vectorized submatrices $\{\mathrm{vec}(\mathbf{\Sigma}_0(i,j))\}_{i,j}$. Motivated by this result, we use the Gram-Schmidt procedure to construct a basis that incrementally spans more and more of the subspace $\mathrm{span}(\{\mathrm{vec}(\mathbf{\Sigma}_0(i,j))\}_{i,j})$. For the special case of the block-Toeplitz matrix, we have:

$$\mathrm{span}(\{\mathrm{vec}(\mathbf{\Sigma}_0(i,j))\}_{i,j}) = \mathrm{span}(\{\mathrm{vec}(\mathbf{\Sigma}(\tau))\}_{\tau=-N}^{N})$$

where the mapping is given by $\mathbf{\Sigma}_0(i,j) = \mathbf{\Sigma}(j-i)$. Note that $\mathbf{\Sigma}(-\tau) = \mathbf{\Sigma}(\tau)^T$.

For simplicity, consider the case $k = 2k'+1$ for some $k' \geq 0$. From Lemma III.5, we are free to choose an orthonormal basis set $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ and form the projection matrix $\mathbf{P}_k = \mathbf{V}_k \mathbf{V}_k^T$, where the columns of $\mathbf{V}_k$ are the vectors $\{\mathbf{v}_j\}$. We form the orthonormal basis using the Gram-Schmidt procedure [64]:

$$\begin{aligned}
\tilde{\mathbf{v}}_0 &= \mathrm{vec}(\mathbf{\Sigma}(0)), \\
\mathbf{v}_0 &= \frac{\tilde{\mathbf{v}}_0}{\|\tilde{\mathbf{v}}_0\|_2} \\
\tilde{\mathbf{v}}_1 &= \mathrm{vec}(\mathbf{\Sigma}(1)) - \frac{\langle \mathrm{vec}(\mathbf{\Sigma}(1)), \tilde{\mathbf{v}}_0 \rangle}{\|\tilde{\mathbf{v}}_0\|_2^2} \tilde{\mathbf{v}}_0, \\
\mathbf{v}_1 &= \frac{\tilde{\mathbf{v}}_1}{\|\tilde{\mathbf{v}}_1\|_2} \\
\tilde{\mathbf{v}}_{-1} &= \mathrm{vec}(\mathbf{\Sigma}(-1)) - \frac{\langle \mathrm{vec}(\mathbf{\Sigma}(-1)), \tilde{\mathbf{v}}_0 \rangle}{\|\tilde{\mathbf{v}}_0\|_2^2} \tilde{\mathbf{v}}_0 \\
&\quad - \frac{\langle \mathrm{vec}(\mathbf{\Sigma}(-1)), \tilde{\mathbf{v}}_1 \rangle}{\|\tilde{\mathbf{v}}_1\|_2^2} \tilde{\mathbf{v}}_1, \\
\mathbf{v}_{-1} &= \frac{\tilde{\mathbf{v}}_{-1}}{\|\tilde{\mathbf{v}}_{-1}\|_2} \\
&\textit{etc.}
\end{aligned}$$

With this choice of orthonormal basis, it follows that for every $k = 2k'+1$, we have the orthogonal projector:

$$\mathbf{P}_k = \mathbf{v}_0 \mathbf{v}_0^T + \sum_{l=1}^{k'} (\mathbf{v}_l \mathbf{v}_l^T + \mathbf{v}_{-l} \mathbf{v}_{-l}^T)$$

This corresponds to a variant of a sequence of Householder transformations [64].

Using Lemma III.5:

$$\begin{aligned}
\sigma_{k+1}^2(\mathbf{R}_0) &\leq \|\mathbf{R}_0(\mathbf{I}-\mathbf{P}_k)\|_2^2 \\
&\leq \|\mathbf{R}_0 - \mathbf{R}_0\mathbf{P}_k\|_F^2 \\
&\leq p\sum_{l=k'+1}^{N} \|\mathbf{\Sigma}(l)\|_F^2 + \|\mathbf{\Sigma}(-l)\|_F^2 \qquad (3.47) \\
&\leq 2C'pq\sum_{l=k'+1}^{N} u^{2l} \\
&\leq 2C'pq\frac{u^{2k'+2}}{1-u^2} \\
&\leq 2C'pq\frac{u^{k}}{1-u^2}
\end{aligned}$$

where we used Lemma III.8 to obtain (3.47). To finish the proof, using the bound above and (3.13):

$$\begin{aligned}
\inf_{\mathbf{R}:\mathrm{rank}(\mathbf{R})\leq r} \|\mathbf{R}-\mathbf{R}_0\|_F^2 &= \sum_{k=r}^{r_0-1}\sigma_{k+1}^2(\mathbf{R}_0) \\
&\leq \frac{2C'pq}{1-u^2}\sum_{k=r}^{r_0-1} u^k \\
&\leq 2C'pq\frac{u^r}{(1-u)^2}
\end{aligned}$$

The proof is complete. □

3.8.8 Lemma III.8

Lemma III.8. *Consider the notation and setting of proof of Thm. III.6. Then, for the projection matrix* $\mathbf{P}_k$ *chosen, we have for* $k = 2k'+1, k' \geq 1$*:*

$$\sigma_{k+1}^2(\mathbf{R}_0) \leq \|\mathbf{R}_0 - \mathbf{R}_0\mathbf{P}_k\|_F^2 \leq p\sum_{l=k'+1}^{N} \|\mathbf{\Sigma}(l)\|_F^2 + \|\mathbf{\Sigma}(-l)\|_F^2$$

Proof. To illustrate the row-subtraction technique, we consider the simplified scenario $k' = 1$. The proof can be easily generalized to all $k' \geq 1$. Without loss of generality,

we write the permuted covariance

$$\mathbf{\Sigma}_0 = \begin{bmatrix} \mathbf{\Sigma}(0) & \mathbf{\Sigma}(1) \\ \mathbf{\Sigma}(-1) & \mathbf{\Sigma}(0) \end{bmatrix} \tag{3.48}$$

as:

$$\mathbf{R}_0 = \mathcal{R}(\mathbf{\Sigma}_0) = \begin{bmatrix} \text{vec}(\mathbf{\Sigma}(0))^T \\ \text{vec}(\mathbf{\Sigma}(1))^T \\ \text{vec}(\mathbf{\Sigma}(-1))^T \\ \text{vec}(\mathbf{\Sigma}(0))^T \end{bmatrix}$$

Using the Gram-Schmidt submatrix basis construction of the proof of Thm. III.6, the sequence of projection matrices can be written as:

$$\mathbf{P}_1 = \mathbf{v}_0\mathbf{v}_0^T$$

$$\mathbf{P}_2 = \mathbf{v}_0\mathbf{v}_0^T + \mathbf{v}_1\mathbf{v}_1^T$$

$$\mathbf{P}_3 = \mathbf{v}_0\mathbf{v}_0^T + \mathbf{v}_1\mathbf{v}_1^T + \mathbf{v}_{-1}\mathbf{v}_{-1}^T$$

where $\mathbf{v}_i$ is the orthonormal basis constructed in the proof of Thm. III.6. The singular value bound $\sigma_1^2(\mathbf{R}_0) \leq \|\mathbf{R}_0\|_F^2 = 2\|\mathbf{\Sigma}(0)\|_F^2 + \|\mathbf{\Sigma}(1)\|_F^2 + \|\mathbf{\Sigma}(-1)\|_F^2$ is trivial [64].

For the second singular value, we want to prove the bound:

$$\sigma_2^2(\mathbf{R}_0) \leq \|\mathbf{\Sigma}(1)\|_F^2 + \|\mathbf{\Sigma}(-1)\|_F^2 \tag{3.49}$$

To show this, we use the variational bound of Lemma III.5:

$$\begin{aligned}
\sigma_2^2(\mathbf{R}_0) &\leq \|\mathbf{R}_0 - \mathbf{R}_0\mathbf{P}_1\|_F^2 \\
&= \left\| \begin{bmatrix} \operatorname{vec}(\boldsymbol{\Sigma}(0))^T - \langle \operatorname{vec}(\boldsymbol{\Sigma}(0)), \mathbf{v}_0 \rangle \mathbf{v}_0^T \\ \operatorname{vec}(\boldsymbol{\Sigma}(1))^T - \langle \operatorname{vec}(\boldsymbol{\Sigma}(1)), \mathbf{v}_0 \rangle \mathbf{v}_0^T \\ \operatorname{vec}(\boldsymbol{\Sigma}(-1))^T - \langle \operatorname{vec}(\boldsymbol{\Sigma}(-1)), \mathbf{v}_0 \rangle \mathbf{v}_0^T \\ \operatorname{vec}(\boldsymbol{\Sigma}(0))^T - \langle \operatorname{vec}(\boldsymbol{\Sigma}(0)), \mathbf{v}_0 \rangle \mathbf{v}_0^T \end{bmatrix} \right\|_F^2 \\
&= \left\| \begin{bmatrix} \mathbf{0}^T \\ \operatorname{vec}(\boldsymbol{\Sigma}(1))^T - \langle \operatorname{vec}(\boldsymbol{\Sigma}(1)), \mathbf{v}_0 \rangle \mathbf{v}_0^T \\ \operatorname{vec}(\boldsymbol{\Sigma}(-1))^T - \langle \operatorname{vec}(\boldsymbol{\Sigma}(-1)), \mathbf{v}_0 \rangle \mathbf{v}_0^T \\ \mathbf{0}^T \end{bmatrix} \right\|_F^2 \\
&= \|\operatorname{vec}(\boldsymbol{\Sigma}(1)) - \langle \operatorname{vec}(\boldsymbol{\Sigma}(1)), \mathbf{v}_0 \rangle \mathbf{v}_0\|_2^2 \\
&\quad + \|\operatorname{vec}(\boldsymbol{\Sigma}(-1)) - \langle \operatorname{vec}(\boldsymbol{\Sigma}(-1)), \mathbf{v}_0 \rangle \mathbf{v}_0\|_2^2 \\
&\leq \|\boldsymbol{\Sigma}(1)\|_F^2 + \|\boldsymbol{\Sigma}(-1)\|_F^2
\end{aligned}$$

where in the last step, we used the Pythagorean principle from least-squares theory [94]-i.e.$\|\mathbf{A} - \frac{<\mathbf{A},\mathbf{B}>}{\|\mathbf{B}\|_F^2}\mathbf{B}\|_F^2 \leq \|\mathbf{A}\|_F^2$ for any matrices $\mathbf{A}, \mathbf{B}$ of the same order. Next, we want to show

$$\sigma_3^2(\mathbf{R}_0) \leq \|\boldsymbol{\Sigma}(-1)\|_F^2 \tag{3.50}$$

Define $\gamma(j) = \operatorname{vec}(\boldsymbol{\Sigma}(j)) - \langle \operatorname{vec}(\boldsymbol{\Sigma}(j)), \mathbf{v}_0 \rangle \mathbf{v}_0$. Using similar bounds and the above,

after some algebra:

$$\begin{aligned}
\sigma_3^2(\mathbf{R}_0) &\leq \|\mathbf{R}_0 - \mathbf{R}_0\mathbf{P}_2\|_F^2 \\
&= \left\| \begin{bmatrix} \mathbf{0}^T \\ \gamma(1)^T - \langle \mathrm{vec}(\mathbf{\Sigma}(1)), \mathbf{v}_1 \rangle \, \mathbf{v}_1^T \\ \gamma(-1)^T - \langle \mathrm{vec}(\mathbf{\Sigma}(-1)), \mathbf{v}_1 \rangle \, \mathbf{v}_1^T \\ \mathbf{0}^T \end{bmatrix} \right\|_F^2 \\
&= \|\mathrm{vec}(\mathbf{\Sigma}(-1))^T - \langle \mathrm{vec}(\mathbf{\Sigma}(-1)), \mathbf{v}_0 \rangle \, \mathbf{v}_0^T \\
&\quad - \langle \mathrm{vec}(\mathbf{\Sigma}(-1)), \mathbf{v}_1 \rangle \, \mathbf{v}_1^T \|_2^2 \\
&= \|\mathrm{vec}(\mathbf{\Sigma}(-1))^T\|_2^2 - |\, \langle \mathrm{vec}(\mathbf{\Sigma}(-1)), \mathbf{v}_0 \rangle \,|^2 \\
&\quad - |\, \langle \mathrm{vec}(\mathbf{\Sigma}(-1)), \mathbf{v}_1 \rangle \,|^2 \\
&\leq \|\mathbf{\Sigma}(-1)\|_F^2
\end{aligned}$$

where we observed that $\gamma(1) = \langle \mathrm{vec}(\mathbf{\Sigma}(1)), \mathbf{v}_1 \rangle \, \mathbf{v}_1$ and used the Pythagorean principle again.

Using $\mathbf{P}_3$ and similar bounds, it follows that $\sigma_4^2(\mathbf{R}_0) = 0$, which makes sense since the separation rank of (3.48) is at most 3. Generalizing to $k' \geq 1$ and noting that $\|\mathbf{\Sigma}_0\|_F^2 = p\|\mathbf{\Sigma}(0)\|_F^2 + \sum_{l=1}^{p-1}(p-l)\|\mathbf{\Sigma}(l)\|_F^2 + \|\mathbf{\Sigma}(l)\|_F^2 \leq p\|\mathbf{\Sigma}(0)\|_F^2 + p\sum_{l=1}^{p-1}\|\mathbf{\Sigma}(l)\|_F^2 + \|\mathbf{\Sigma}(-l)\|_F^2$, we conclude the proof. □

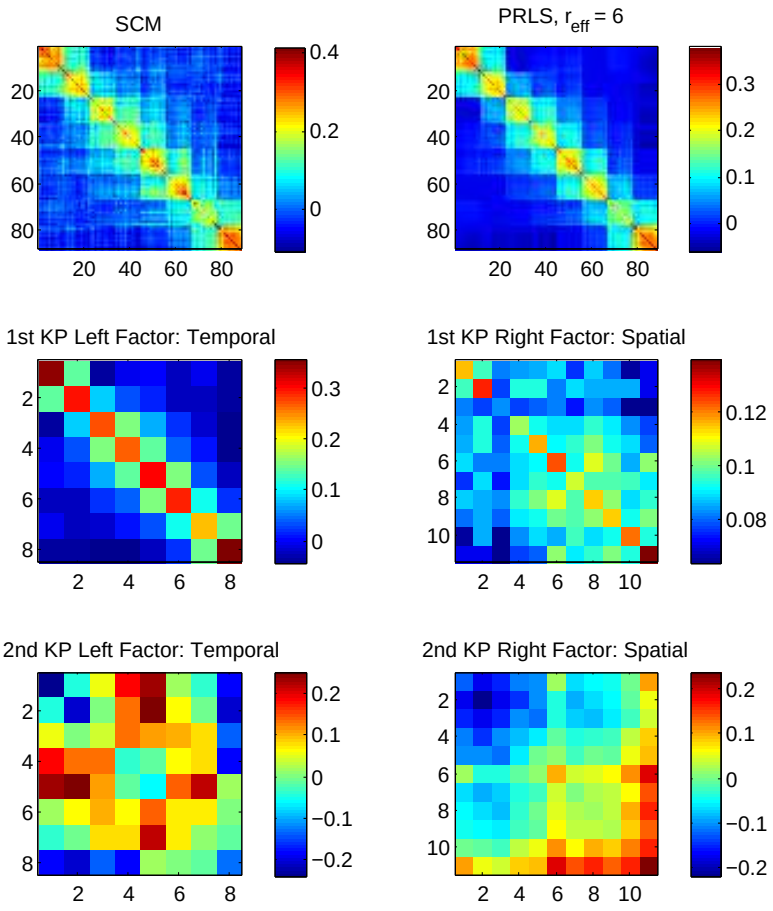


Figure 3.8: Irish wind speed data: Sample covariance matrix (SCM) (top left), PRLS covariance estimate (top right), temporal Kronecker factor for first KP component (middle left) and spatial Kronecker factor for first KP component (middle right), temporal Kronecker factor for second KP component (bottom left) and spatial Kronecker factor for second KP component (bottom right). Note that the second order factors are not necessarily positive definite, although the sum of the components (i.e., the PRLS solution) is positive definite for large enough n. Each KP factor has unit Frobenius norm. Note that the plotting scales the image data to the full range of the current colormap to increase visual contrast.

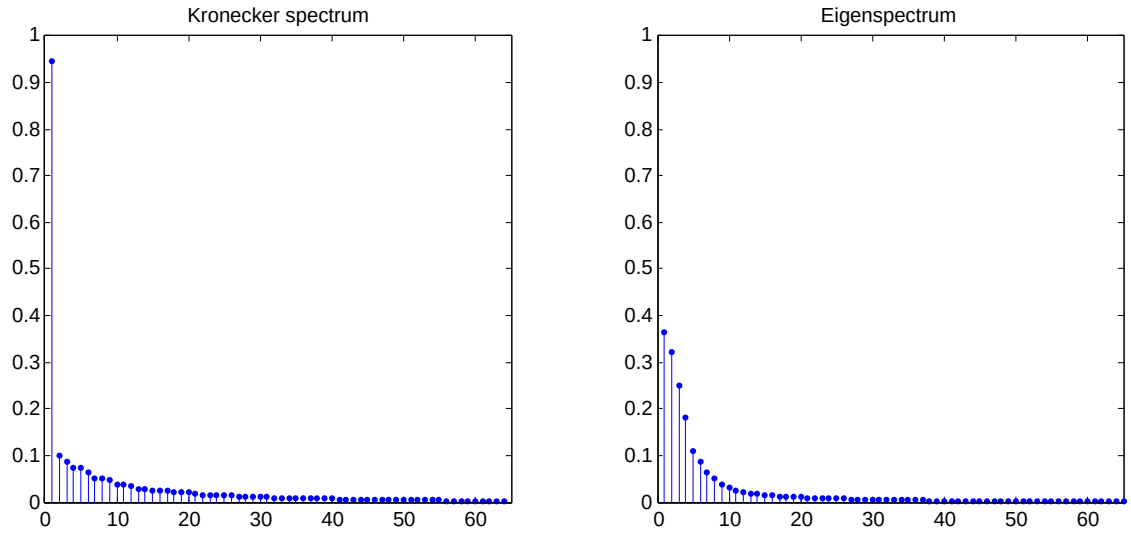


Figure 3.9: Irish wind speed data: Kronecker spectrum of SCM (left) and Eigenspectrum of SCM (right). The first and second KP components contain 94.60% and 1.07% of the spectrum energy. The first and second eigenvectors contain 36.28% and 28.76% of the spectrum energy. The KP spectrum is more compact than the eigenspectrum. Here, the eigenspectrum is truncated at $\min(p^2, q^2) = 8^2 = 64$ to match the Kronecker spectrum. Each spectrum was normalized such that each component has height equal to the percentage of energy associated with it.

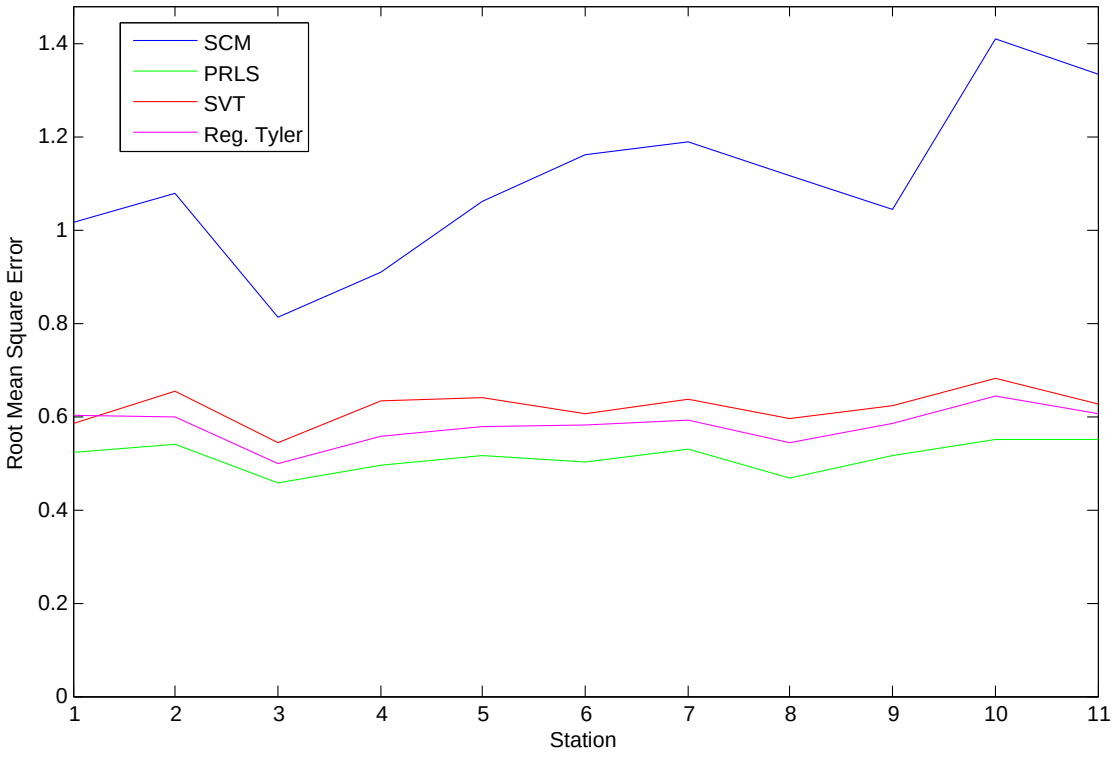


Figure 3.10: Irish wind speed data: RMSE prediction performance across q stations for linear estimators using SCM (blue), PRLS (green), SVT (red) and regularized Tyler (magenta). PRLS, SVT and regularized Tyler respectively achieve an average reduction in RMSE of 3.32, 2.50 and 2.79 dB as compared to SCM (averaged across stations).

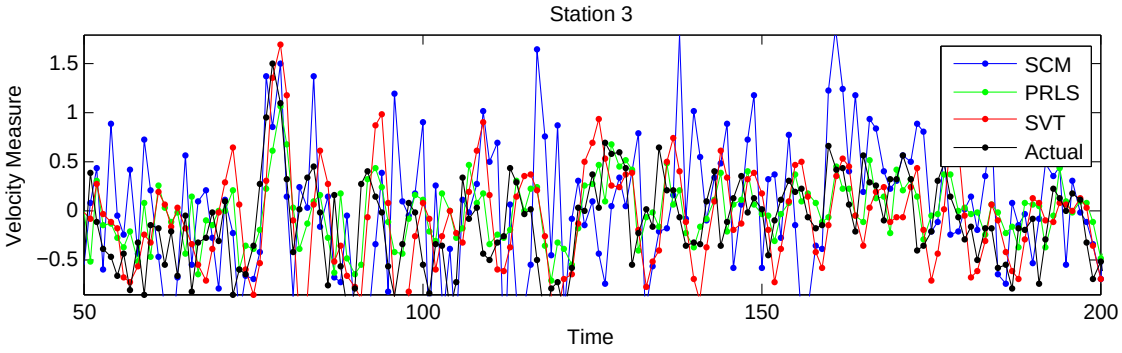


Figure 3.11: Irish wind speed data: Prediction performance for linear estimators using SCM (blue), SVT (red) and PRLS (green) for a time interval of 150 days. The actual (ground truth) wind speeds are shown in black. PRLS offers better tracking performance as compared to SVT and SCM.

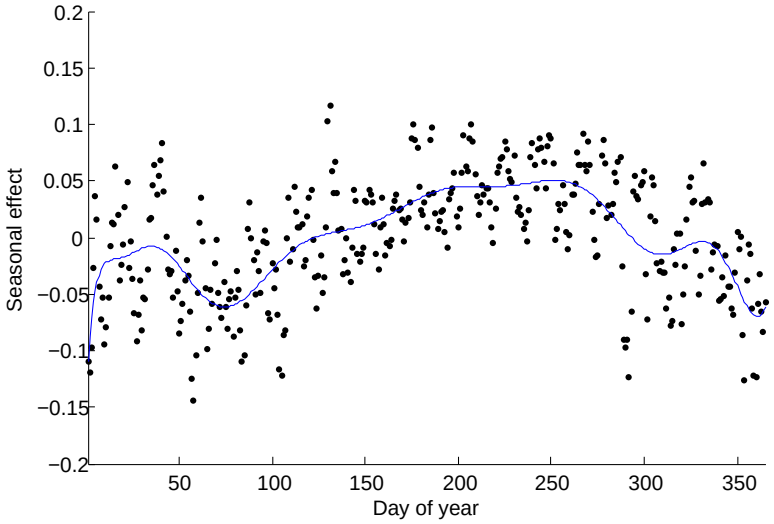


Figure 3.12: NCEP wind speed data (Continental US): Seasonal effect as a function of day of the year. A 14th order polynomial is fit by the least squares method to the average of the square root of the daily mean wind speeds over all stations and over all training years.

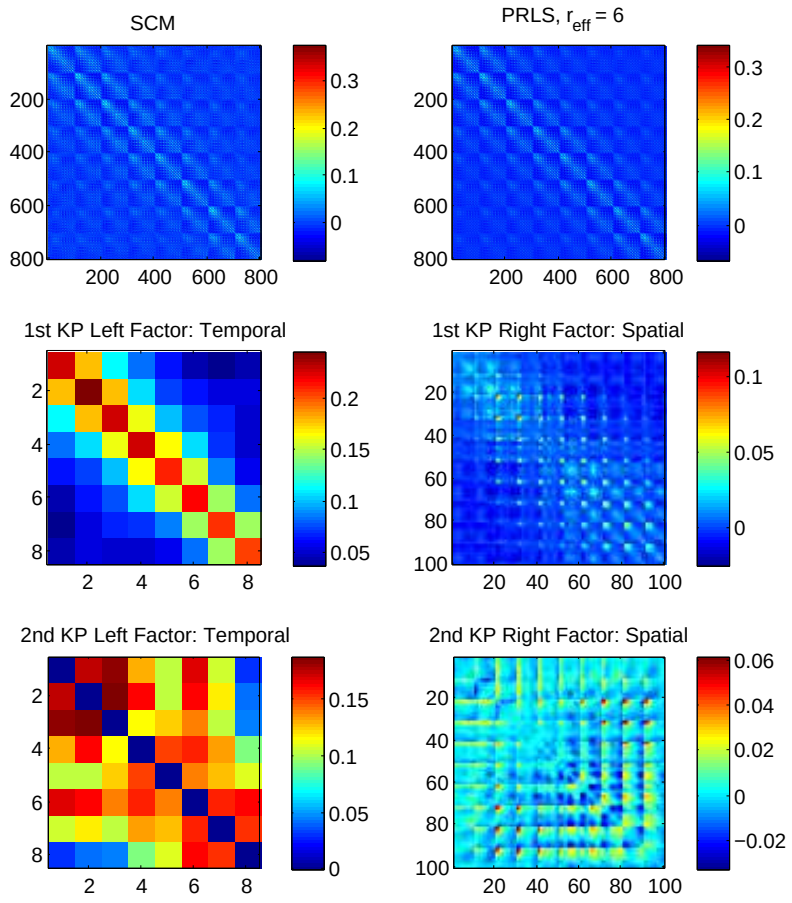


Figure 3.13: NCEP wind speed data (Continental US): Sample covariance matrix (SCM) (top left), PRLS covariance estimate (top right), temporal Kronecker factor for first KP component (middle left) and spatial Kronecker factor for first KP component (middle right), temporal Kronecker factor for second KP component (bottom left) and spatial Kronecker factor for second KP component (bottom right). Note that the second order factors are not necessarily positive definite, although the sum of the components (i.e., the PRLS solution) is positive definite for large enough n. Each KP factor has unit Frobenius norm. Note that the plotting scales the image data to the full range of the current colormap to increase visual contrast.

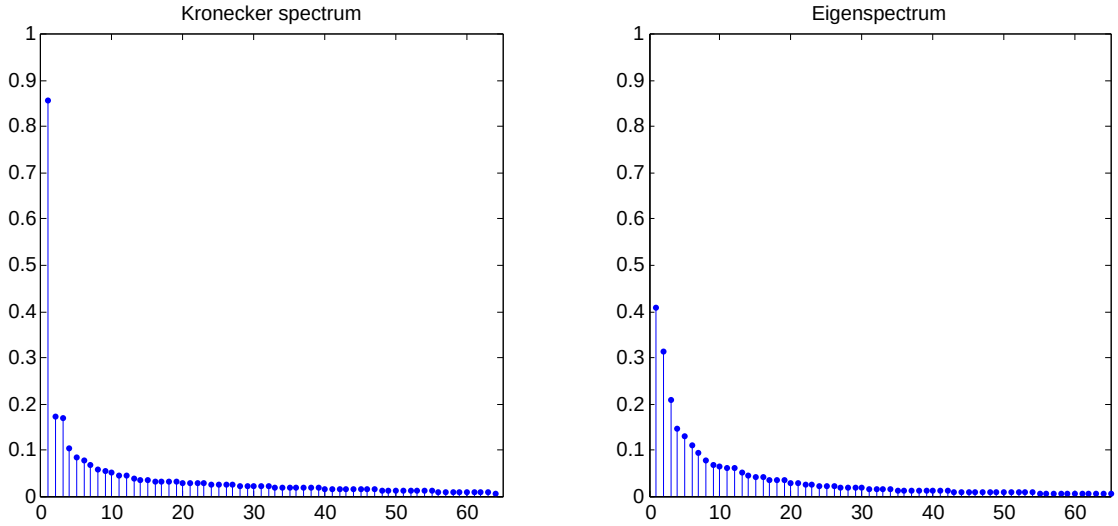


Figure 3.14: NCEP wind speed data (Continental US): Kronecker spectrum of SCM (left) and Eigenspectrum of SCM (right). The first and second KP components contain 85.88% and 3.48% of the spectrum energy. The first and second eigenvectors contain 40.93% and 23.82% of the spectrum energy. The KP spectrum is more compact than the eigenspectrum. Here, the eigenspectrum is truncated at $\min(p^2, q^2) = 8^2 = 64$ to match the Kronecker spectrum. Each spectrum was normalized such that each component has height equal to the percentage of energy associated with it.

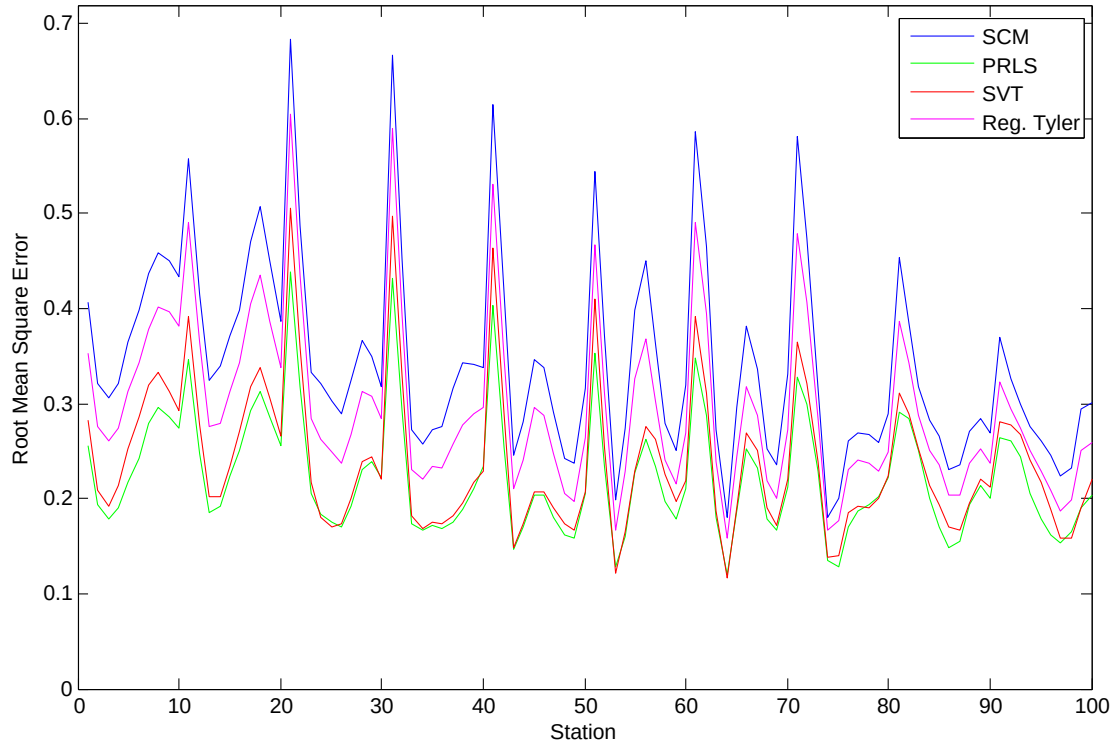


Figure 3.15: NCEP wind speed data (Continental US): RMSE prediction performance across q stations for linear estimators using SCM (blue), SVT (red), PRLS (green) and regularized Tyler (magenta). The estimators PRLS, SVT, and regularized Tyler respectively achieve an average reduction in RMSE of 1.90, 1.59, and 0.66 dB as compared to SCM (averaged across stations).

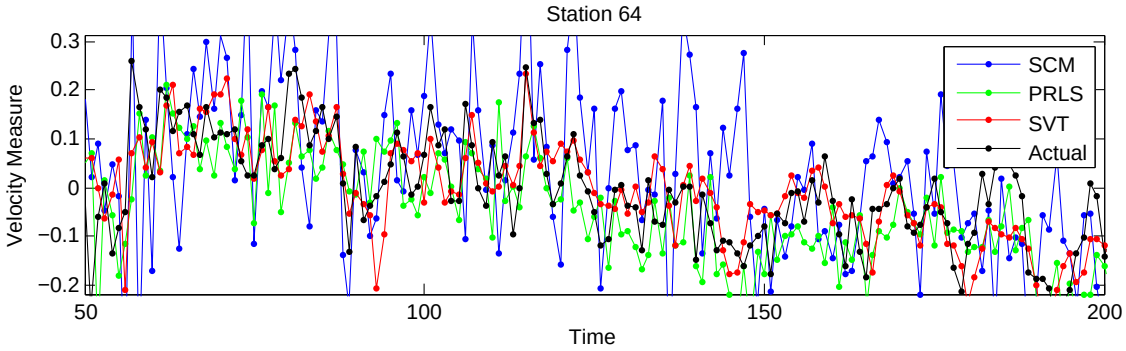


Figure 3.16: NCEP wind speed data (Continental US): Prediction performance for linear estimators using SCM (blue), SVT (red) and PRLS (green) for a time interval of 150 days. The actual (ground truth) wind speeds are shown in black. PRLS offers better tracking performance as compared to SCM and SVT.

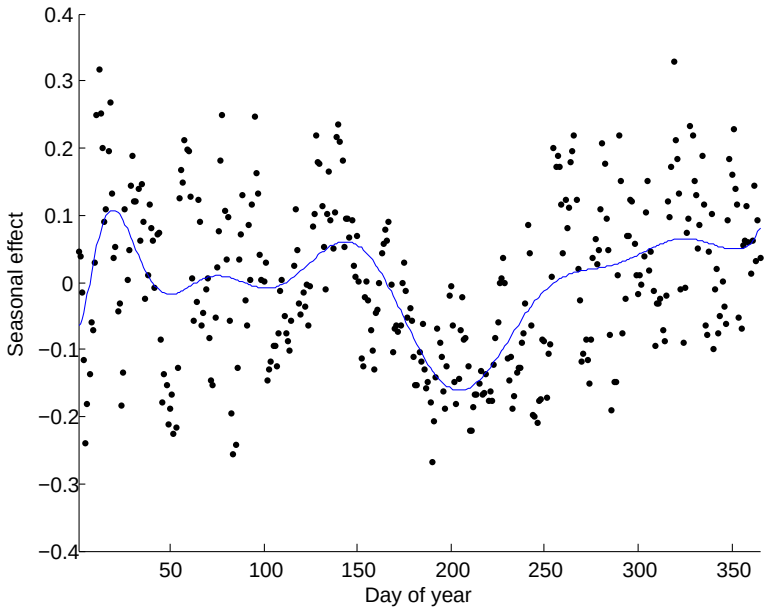


Figure 3.17: NCEP wind speed data (Arctic Ocean): Seasonal effect as a function of day of the year. A 14th order polynomial is fit by the least squares method to the average of the square root of the daily mean wind speeds over all stations and over all training years.

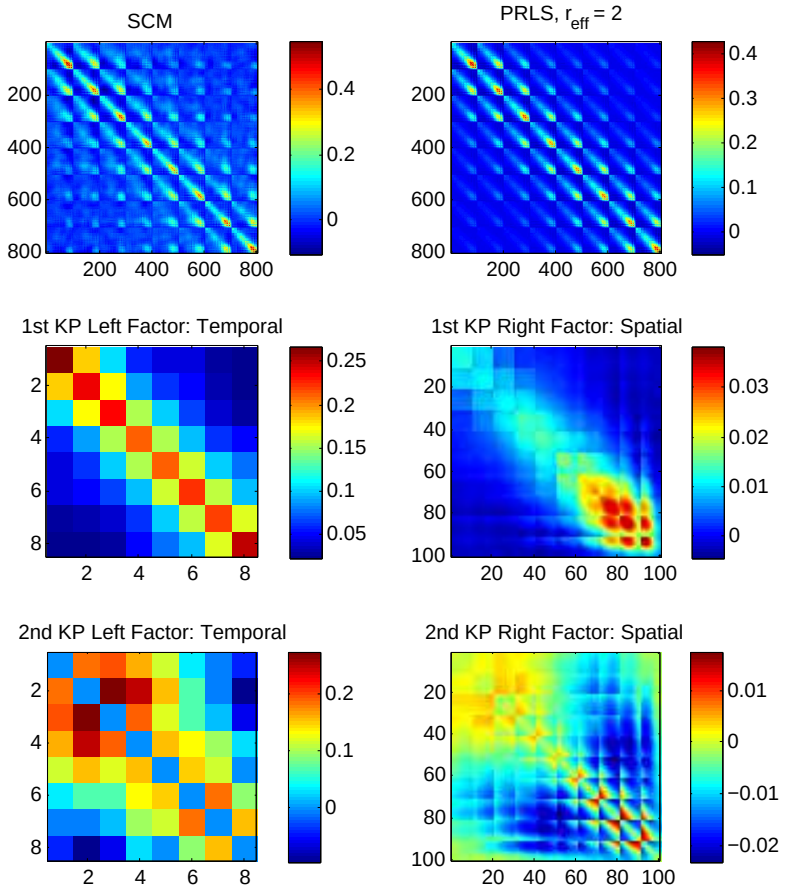


Figure 3.18: NCEP wind speed data (Arctic Ocean): Sample covariance matrix (SCM) (top left), PRLS covariance estimate (top right), temporal Kronecker factor for first KP component (middle left) and spatial Kronecker factor for first KP component (middle right), temporal Kronecker factor for second KP component (bottom left) and spatial Kronecker factor for second KP component (bottom right). Note that the second order factors are not necessarily positive definite, although the sum of the components (i.e., the PRLS solution) is positive definite for large enough n. Each KP factor has unit Frobenius norm. Note that the plotting scales the image data to the full range of the current colormap to increase visual contrast.

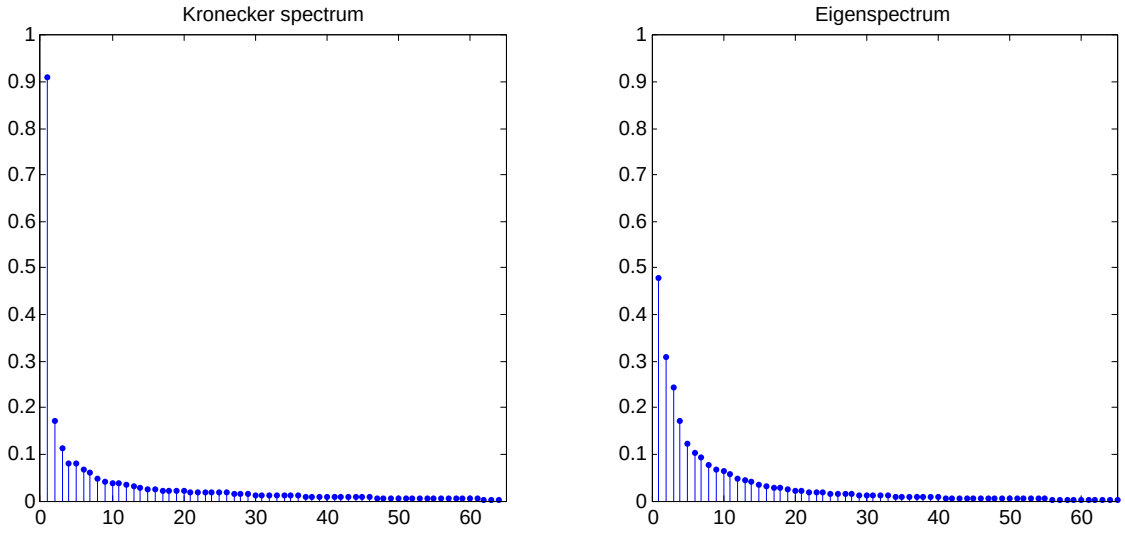


Figure 3.19: NCEP wind speed data (Arctic Ocean): Kronecker spectrum of SCM (left) and Eigenspectrum of SCM (right). The first and second KP components contain 91.12% and 3.28% of the spectrum energy. The first and second eigenvectors contain 47.99% and 19.68% of the spectrum energy. The KP spectrum is more compact than the eigenspectrum. Here, the eigenspectrum is truncated at $\min(p^2, q^2) = 8^2 = 64$ to match the Kronecker spectrum. Each spectrum was normalized such that each component has height equal to the percentage of energy associated with it.

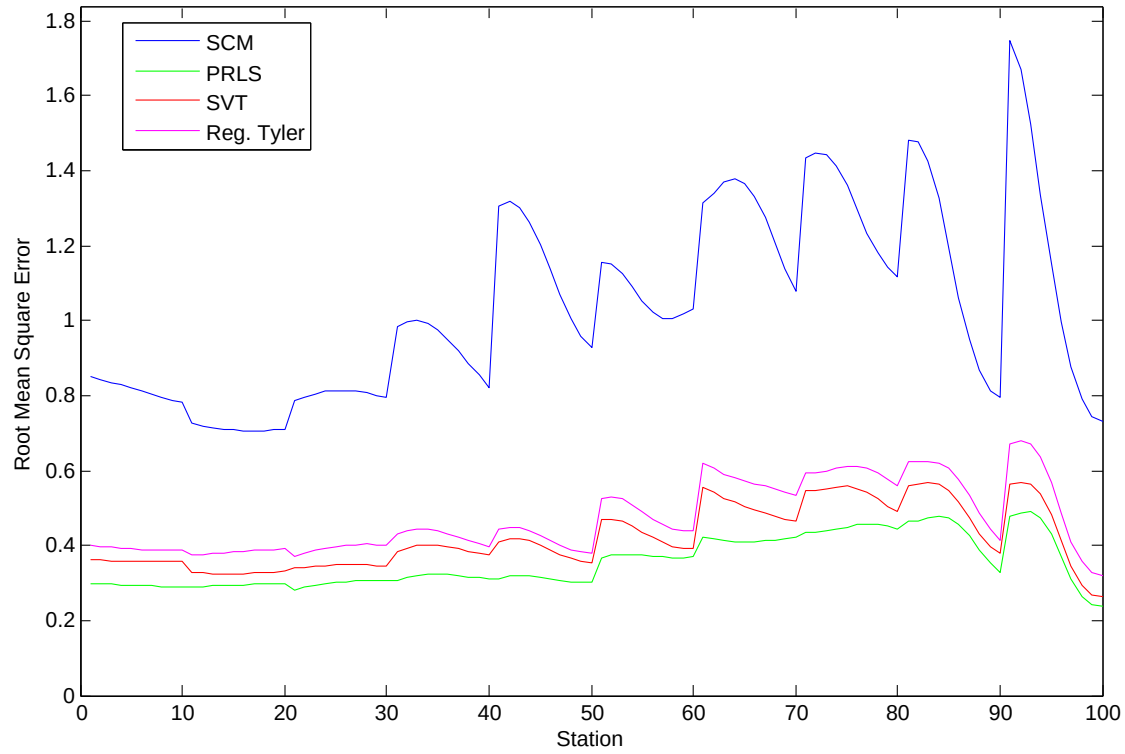


Figure 3.20: NCEP wind speed data (Arctic Ocean): RMSE prediction performance across q stations for linear estimators using SCM (blue) and PRLS (green). The estimators PRLS, SVT and regularized Tyler respectively achieve an average reduction in RMSE of 4.64, 3.91 and 3.41 dB as compared to SCM (averaged across stations).

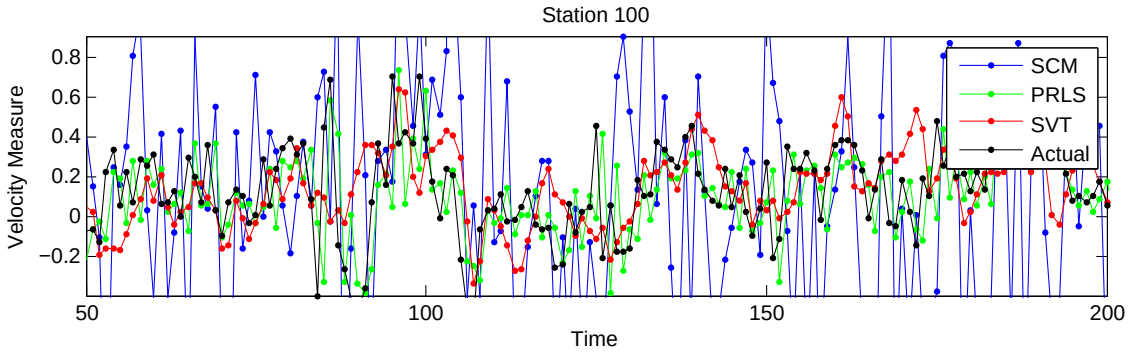


Figure 3.21: NCEP wind speed data (Arctic Ocean): Prediction performance for linear estimators using SCM (blue), SVT (red) and PRLS (green) for a time interval of 150 days. The actual (ground truth) wind speeds are shown in black. PRLS offers better tracking performance as compared to SCM and SVT.

CHAPTER IV

Centralized Collaborative 20 Questions

We consider the problem of 20 questions with noise for multiple players under the minimum entropy criterion [73] in the setting of stochastic search, with application to target localization. Each player yields a noisy response to a binary query governed by a certain error probability. First, we propose a sequential policy for constructing questions that queries each player in sequence and refines the posterior of the target location. Second, we consider a joint policy that asks all players questions in parallel at each time instant and characterize the structure of the optimal policy for constructing the sequence of questions. This generalizes the single player probabilistic bisection method [73, 30] for stochastic search problems. Third, we prove an equivalence between the two schemes showing that, despite the fact that the sequential scheme has access to a more refined filtration, the joint scheme performs just as well on average. Fourth, we establish convergence rates of the mean-square error (MSE) and derive error exponents. We also prove almost sure convergence of the estimates to the true target location. Lastly, we obtain an extension to the case of unknown error probabilities. This framework provides a mathematical model for incorporating a human in the loop for active machine learning systems.

4.1 Introduction

What is the intrinsic value of adding a human-in-the-loop to an autonomous learning machine, e.g., an automated target recognition (ATR) sensor? In the ATR setting the answer to this question could provide insight into human-aided autonomous sensing for estimating an unknown target location or identifying a target. A simple model for such a human-in-the-loop system is a collaborative multi-player 20 questions game: the human is repeatedly queried about target location in order to improve ATR performance. This chapter proposes such a 20 questions framework for studying the value of including a human-in-the-loop and for optimizing the sequence of queries.

Motivated by the approach of Jedynak et al [73], which was restricted to the single player case, we model the human-machine interaction as a noisy collaborative 20 questions game. In this framework a controller sequentially selects a set of questions about target location and uses the noisy responses of the human and the machine to formulate the next pair of questions. The query response models for the human and the machine are different, but complementary. While the machine's accuracy is constant over time, the accuracy of the human degrades over time, reflecting the human's decreased ability to resolve questions about the target location near the end of the game.

As in Jamieson et al [72], we use a simple noisy query-response model with different reliability functions for the machine and the human (called derivative-free optimizers (DFO) in [72]). Under this model we specify the optimal query policy, establish an equivalence theorem, and obtain MSE bounds and convergence rates. Our model predicts that the value of including the human-in-the-loop, as measured

by the human gain ratio (HGR), defined as a ratio of MSE's. The HGR initially increases when localization errors are large, and then slowly decreases over time as the location errors go below the human's fine resolution capability.

The paper by Jedynak et al. [73] formulates the single player 20 questions problem as follows. A controller queries a noisy oracle about whether or not a target X^* lies in a set $A_n \subset \mathbb{R}^d$. Starting with a prior distribution on the target's location $p_0(\cdot)$, the objective in [73] is to minimize the expected entropy of the posterior distribution:

$$\inf_{\pi} \mathbb{E}^{\pi} \left[H(p_N) \right] \tag{4.1}$$

where $\pi = (\pi_0, \pi_1, \dots)$ denotes the controller's query policy and the entropy is the standard differential entropy [35]:

$$H(p) = -\int_{\mathcal{X}} p(x) \log p(x) dx.$$

The posterior median of p_N is used to estimate the target location after N questions. Jedynak [73] shows the bisection policy is optimal under the minimum entropy criterion. To be concrete, in Thm. 2 of [73], optimal policies are characterized by:

$$\mathbb{P}_n(A_n) := \int_{A_n} p_n(x) dx = u^* \in \arg\max_{u \in [0,1]} \phi(u) \tag{4.2}$$

where

$$\phi(u) = H(f_1 u + (1-u) f_0) - u H(f_1) - (1-u) H(f_0)$$

is nonnegative. The densities f_0 and f_1 correspond to the noisy channel [1]:

$$\mathbb{P}(Y_{n+1} = y | Z_n = z) = f_0(y) I(z = 0) + f_1(y) I(z = 1)$$

where $Z_n = I(X^* \in A_n) \in \{0, 1\}$ is the channel input. While the framework applies to both continuous and discrete random variables y, in [73] the focus was on the

[1] The function $I(A)$ is the indicator function throughout the chapter-i.e., $I(A) = 1$ if A is true and zero otherwise.

binary case-i.e., $y \in \{0, 1\}$. The noisy channel models the conditional probability of the response to each question being correct. For the special case of a binary symmetric channel (BSC), $u^* = 1/2$ and the probabilistic bisection policy [73, 30] becomes an optimal policy.

The 20 questions framework in the single player setting is analogous to computerized adaptive testing (CAT). In CAT, the objective is to identify the unknown testing ability of the subject by asking a sequence of questions adaptively [127]. To do this, an iterative algorithm is constructed, where at each step, a question is chosen from a pool based on the current estimate of the examinee's ability, the subject responds to the question correctly or incorrectly, and the ability esitmate is updated based upon all prior answers. Computer adaptive tests tend to arrive at accurate ability estimates faster than non-adaptive tests. The ingredients of a CAT include a mathematical model for the probability of a subject with proficiency $\theta \in \mathbb{R}$ responding correctly to an item of difficulty $b \in \mathbb{R}$ and an adaptive testing algorithm [127]. In the literature, the modeling part is known as item response theory (IRT) and the testing algorithm is currently based on sequential scoring or Bayesian methods [127].

The basic assumption behind IRT is to have all items measure the same quantity of interest. In CAT, this quantity of interest is a single dimension of knowledge (e.g. mathematical ability, verbal proficiency) on which all items depend on for their correct response. Each test item's difficulty, b, is the position that it occupies on this dimension, and the subject's proficiency level, θ, is the position of each subject on this level. To get concrete, a simple IRT model is a three-parameter model based on the logistic function used to model the probability of the correct response of a subject with proficiency θ responding to an item of difficulty b:

$$\mathbb{P}(\theta; a, b, c) = c + \frac{1 - c}{1 + \exp(-a(\theta - b))} \tag{4.3}$$

The CAT objective is to estimate the proficiency level θ of a subject. The subject can answer each question correctly or incorrectly, which is modeled by the response x_j for the jth item. For simplicity, let us assume that the parameters $\beta_j = (a_j, b_j, c_j)$ of the jth item are known for all $j \in J$.

The item selection used to be done using branching methods, but now have been superseded by more efficient methods. Given the estimate of θ, say $\hat{\theta}_n$, based on the previous responses, two strategies are most widely used for selecting the next test item known as maximum information and maximum expected precision [127].

The (unconstrained) maximum information criterion chooses the item that maximizes the Fisher information of the item [127]:

$$j_{n+1} \in \arg\max_{j \in J} I(\hat{\theta}_n; j) = \frac{(\nabla_\theta \mathbb{P}(\hat{\theta}_n; j))^2}{\mathbb{P}(\hat{\theta}_n; j)(1 - \mathbb{P}(\hat{\theta}_n; j))} \tag{4.4}$$

where $I(\theta; j) = \mathbb{E}[(\nabla_\theta \log p(x_j|\theta))^2|\theta]$ is the Fisher information corresponding to the Bernoulli distribution $p(x_j|\theta; j) = \mathbb{P}(\theta; j)^{x_j}(1 - \mathbb{P}(\theta; j))^{1-x_j}$. The maximum expected precision method is based on a similar idea, but working with the posterior distribution $p(\theta|\mathcal{B}_n) = p_n(\theta)$ directly. Here, $\mathcal{B}_n$ denotes the information available about the subject after n items, which includes group memberships and previous responses. The next item is chosen to maximize the expected precision of the posterior distribution [93, 127]:

$$j_{n+1} \in \arg\max_{j \in J} \mathbb{E}\left[\mathrm{Var}(p_{n+1}(\cdot))^{-1}|j_{n+1} = j\right] \tag{4.5}$$

where $p_{n+1}(\theta) = p(\theta|\mathcal{B}_n, j_{n+1}, x_{n+1})$.

In practice, certain constraints including lack of test item repetitions, balance of item content and item rate of exposure need to be taken into account when selecting the next test item from the pool.

Given the posterior distribution, the proficiency of a subject can be estimated by calculating the maximum or the conditional mean. Like the maximum expected precision method of Owen [93], the 20 questions framework is also a sequential Bayesian approach that iteratively updates the posterior distribution of the target location given the previous questions and responses. However, the 20 questions framework adopted in this chapter differs from the CAT setup in several important respects:

- The questions chosen are associated with continuous regions A_n. In CAT, the pool of questions is discrete (although it may be uncountable).
- The objective of the 20 questions is to choose the queries sequentially that minimize the entropy of the posterior distribution after N steps (see (4.1)), while the objective in adaptive testing is to choose the items (of possibly different difficulty level) that maximize the expected inverse variance of the posterior distribution.

We conclude the comparison with CAT with a final remark. If the posterior distribution is Gaussian, then the maximum expected criterion is equivalent to the minimum entropy criterion for one-stage. This follows from the fact that the entropy of the Gaussian distribution $N(\mu, \sigma^2)$ is given by $1/2 \log(2\pi\sigma^2)$. We note however, that the posterior distribution for the 20 questions game is a piecewise constant function and thus is never Gaussian. Thus, the two approaches are loosely related but not equivalent.

In this chapter, we derive optimality conditions for optimal query strategies in the collaborative multiplayer case. We propose a sequential bisection policy for which each player responds to a single question about the location of the target, and a joint policy where all players are asked questions simultaneously. We show that even

when the collaborative players act independently, jointly optimal policies require overlapping non-identical queries. We prove that the maximum entropy reduction for the sequential bisection scheme is the same as that of the jointly optimal scheme, and is given by the sum of the capacities of all the players' channels. This is important since, while the jointly optimal scheme might be hard to implement as the number of players and dimensions increase, the sequential scheme only requires a sequence of bisections followed by intermediate posterior updates. Thus, by implementing the sequential policy, complexity is transferred from the controller to the posterior updates. Despite the fact that the optimal sequential policy has access to a more refined filtration, it achieves the same average performance as the optimal joint policy.

We extend this equivalence to the setting where the error channels associated with the players are unknown. In this case, we show that the entropy loss at each iteration is no longer constant; it is time-varying and equals the conditional expectation of the sum of the capacities of the players' channels with respect to the filtration up to the current time. In addition, we show that even for one-dimensional targets, the optimal policy for the unknown channel case is not equivalent to the probabilistic bisection policy.

The work by Castro and Nowak [30, 31] provides upper bounds on the MSE of the posterior mean of the target for the single player case. We extend their MSE bounds to the multiplayer case and provide new lower bounds on MSE by linking the information theoretic analysis to convergence rates. The combination of the upper and lower bounds sandwiches the MSE between two exponentially decaying functions of the number of plays in the 20 questions game.

Our 20 questions framework differs from other binary forced choice problems that have appeared in the literature. This includes educational testing, e.g., using

dynamic item response models [128], and active learning, e.g., using paired comparisons for ranking two objects [71]. Like the 20 questions framework, in [128, 71], a sequence of binary questions is formulated by a controller. However, the 20 questions problem considered in this chapter is quite different. The goals are not the same: in contrast with sequential testing considered in [128, 71], here as in [73] we consider sequential estimation of a continuous valued target state. Furthermore, in [128, 71] the queries are posed to a single player whereas we consider multiple players who cooperate to achieve the goal.

4.1.1 Outline

The outline of this chapter is as follows. Section 4.2 introduces the notation and collaborative player setup. This introduces the sequential bisection policy and the joint policy, and establishes that the respective optimal policies attain identical performance. Section 4.3 derives upper and lower bounds on the MSE and Section 4.5 develops similar bounds for a human error model. Section 4.6 extends the analysis to the case that the error probabilities are not known. The theory is illustrated by simulation in Section 4.7 followed by our conclusions in Section 4.8.

4.2 Noisy 20 Questions with Collaborative Players: Known Error Probability

Assume that there is a target with unknown state $X^* \in \mathcal{X} \subset \mathbb{R}^d$. We focus on the case where the target state is spatial location, i.e., in $d = 2$ or 3 dimensions. However, our results are applicable to higher dimensions also, e.g., where X^* is a kinematic state or some other multi-dimensional target feature. Starting with a prior distribution $p_0(x)$ on X^*, the aim is to find an optimal policy for querying a machine (hereafter referred to as player 1) about the target state, with the additional help of

humans. The policy's objective is to minimize the expected Shannon entropy of the posterior density $p_n(x)$ of the target location after n questions.

There are M collaborating players that can be asked questions at each time instant n. The objective of the players is to come up with the correct answer to a kind of 20 questions game. Next, we introduce two types of query design strategies. The first is a sequential strategy where the controller formulates and asks questions to each player in sequence. The second is a batch strategy where the questions are formulated and directed to all players simultaneously. For fixed n both strategies ask the same number of questions. However, the sequential strategy has the advantage of being able to use the answer of the previous player to better formulate a question to the next one. Below we show that, despite this advantage, the average entropy reduction performances of these two strategies are identical.

4.2.1 Sequential Query Design

The sequential strategy is the following coordinate-by-coordinate design: ask an optimal query to the first player, then update the posterior density and ask an optimal query to the second player, and so on (see Fig. 4.1). In [73], the optimal query policy for the case of a single player (M=1) was shown to be a bisection rule.

For each time epoch, indexed by n and called a cycle, the controller formulates and asks the M players questions $A_{n_t} = A_{n,t}$, $t = 0, \ldots, M-1$. We denote by $n_t = (n, t)$ the times at which the queries are asked.

Let the mth player's query at time $n_t = n_{m-1}$ be "does X^* lie in the region $A_{n_t} \subset \mathbb{R}^d$?". We denote the truth state of the query as the binary variable $Z_{n_t} = I(X^* \in A_{n_t}) \in \{0, 1\}$ and the noisy binary response of the mth player is $Y_{n_{t+1}} \in \{0, 1\}$.

The query region A_{n_t} chosen at time n_t depends on the information available at that time. More formally, define the multi-index (n, t) where $n = 0, 1, \ldots$ indexes

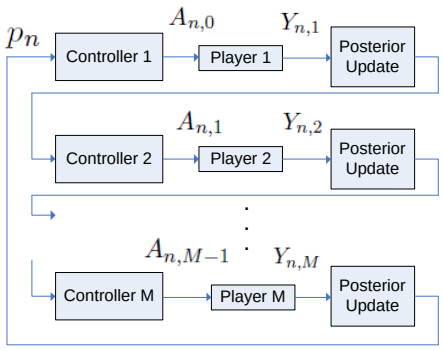


Figure 4.1: Controllers sequentially ask questions to M collaborative players about the location X^* of an unknown target. At time n, the first controller chooses the query $I(X^* \in A_{n,0})$ based on the posterior p_n. Then, player 1 yields the noisy response $Y_{n,1}$ that is used to update the posterior, and the second controller chooses the next query $I(X^* \in A_{n,1})$ for player 2 based on the updated posterior, etc.

over cycles and $t = 0, \ldots, M-1$ indexes within cycles. Define the nested sequence of sigma-algebras $\mathcal{G}_{n,t}$, $\mathcal{G}_{n,t} \subset \mathcal{G}_{n+i,t+j}$, for all $i \geq 0$ and $j \in \{0, \ldots, M-1-t\}$, generated by the sequence of queries and the players' responses. The filtration $\mathcal{G}_{n,t}$ carries all the information accumulated by the controller from time $(0,0)$ to time (n,t). The queries $\{A_{n,t}\}$ formulated by the controller are measurable with respect to this filtration.

4.2.2 Joint Query Design

Let the mth player's query at time n be "does X^* lie in the region $A_n^{(m)} \subset \mathbb{R}^d$?". We denote this query as the binary variable $Z_n^{(m)} = I(X^* \in A_n^{(m)}) \in \{0,1\}$ to which the player yields provides a possibly incorrect (i.e., noisy) binary response $Y_{n+1}^{(m)} \in \{0,1\}$. We consider a similar setting as in [73], which applied to the $M = 1$ player case, but now we have a joint controller that chooses a batch of M queries $\{A_n^{(m)}\}_{m=1}^M$ that are addressed to each of the M players at time n. A block diagram is shown in Fig. 4.2.

As in the sequential query design, the joint queries are selected based on the accumulated information available to the controller. However, since the full batch of

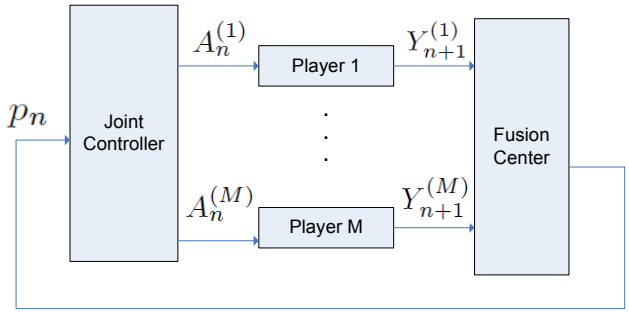


Figure 4.2: A controller asks a batch questions of M collaborative players about the location X^* of an unknown target. At time n, the controller chooses the queries $I(X^* \in A_n^{(m)})$ based on the posterior p_n. Then, the M players yield noisy responses $Y_{n+1}^{(m)}$ that are fed into the fusion center, where the posterior is updated and fed back to the controller at the next time instant $n+1$.

joint queries are determined at the beginning of the n-th cycle, the joint controller only has access to a coarser filtration $\mathcal{F}_n$, $\mathcal{F}_{n-1} \subset \mathcal{F}_n$, as compared with the filtration $\mathcal{G}_{n,t}$ of the sequential controller.

4.2.3 Definitions & Assumptions

Define the M-tuples $\mathbf{Y}_{n+1} = (Y_{n+1}^{(1)}, \ldots, Y_{n+1}^{(M)})$ and $\mathbf{A}_n = \{A_n^{(1)}, \ldots, A_n^{(M)}\}$.

Assumption IV.1. *(Conditional Independence) We assume that the players' responses are conditionally independent. In particular, for the joint controller,*

$$\begin{aligned} &\mathbb{P}(\mathbf{Y}_{n+1} = \mathbf{y} | \mathbf{A}_n, X^* = x, \mathcal{F}_n) \\ &= \prod_{m=1}^{M} \mathbb{P}(Y_{n+1}^{(m)} = y^{(m)} | A_n^{(m)}, X^* = x, \mathcal{F}_n) \end{aligned} \tag{4.6}$$

where

$$\begin{aligned} &\mathbb{P}(Y_{n+1}^{(m)} = y^{(m)} | A_n^{(m)}, X^* = x, \mathcal{F}_n) \\ &= \begin{cases} f_1^{(m)}(y^{(m)} | A_n^{(m)}, \mathcal{F}_n), & x \in A_n^{(m)} \\ f_0^{(m)}(y^{(m)} | A_n^{(m)}, \mathcal{F}_n), & x \notin A_n^{(m)} \end{cases}. \end{aligned} \tag{4.7}$$

Similar relations hold for the sequential controller under the conditional independence

assumption: in (3) and (4) simply change the subscripts n and $n+1$ to n_t and n_{t+1}, respectively, and replace the filtration $\mathcal{F}_n$ by $\mathcal{G}_{n_t}$.

Assumption IV.2. *(Memoryless Binary Symmetric Channels) We model the players' responses as independent (memoryless) binary symmetric channels (BSC) [35] with crossover probabilities $\epsilon_m \in (0, 1/2)$. In particular, for the joint query strategy, the conditional probability mass function $f_j^{(m)} = \mathbb{P}(Y_n^{(m)} = j | A_n^{(m)}, \mathcal{F}_n)$ of the response of the M-th player is:*

$$f_j^{(m)}(y^{(m)} | A_n^{(m)}, \mathcal{F}_n) = f_j^{(m)}(y^{(m)})$$

$$= \begin{cases} 1 - \epsilon_m, & y^{(m)} = j \\ \epsilon_m, & y^{(m)} \neq j \end{cases}$$

where $m = 1, \ldots, M, j = 0, 1$. A similar relation holds for the sequential query strategy: replace n by n_t and $\mathcal{F}_n$ by $\mathcal{G}_{n_t}$.

Define the set of dyadic partitions of $\mathbb{R}^d$, induced by the queries $\{A^{(m)}\}_m$:

$$\gamma(A^{(1)}, \ldots, A^{(M)}) = \left\{ \bigcap_{m=1}^{M} (A^{(m)})^{i_m} : i_m \in \{0, 1\} \right\} \tag{4.8}$$

where $(A)^0 := A^c$ and $(A)^1 := A$. The cardinality of this set of subsets is 2^M and each of these subsets partition $\mathbb{R}^d$. The objective is to localize the target within a subset $A^{(m)}$.

Define the density parameterized by $\mathbf{A}_n, p_n, i_1, \ldots, i_M$, for the joint query strategy:

$$g_{i_1:i_M}(y^{(1)}, \ldots, y^{(M)} | \mathbf{A}_n, \mathcal{F}_n) := \prod_{m=1}^{M} f_{i_m}^{(m)}(y^{(m)} | A_n^{(m)}, \mathcal{F}_n)$$

where $i_j \in \{0, 1\}$.

4.2.4 Equivalence Theorems

We first establish the structure of the optimal joint policy.

Theorem IV.3. *(Joint Optimality Conditions, Known Error Probabilities) Under Assumption IV.1, an optimal joint policy that minimizes the Shannon entropy of the posterior distribution p_n achieves the following entropy loss:*

$$G^* = \sup_{A^{(1)},\ldots,A^{(M)}} \Big\{ H\left(\sum_{i_1:i_M=0}^{1} g_{i_1:i_M}(\cdot)\mathbb{P}_n\Big(\bigcap_{m=1}^{M} (A_n^{(m)})^{i_m} \Big) \right) - \sum_{i_1:i_M=0}^{1} H\left(g_{i_1:i_M}(\cdot)\right) \mathbb{P}_n\Big(\bigcap_{m=1}^{M} (A_n^{(m)})^{i_m} \Big) \Big\}, \tag{4.9}$$

where $H(f)$ is the Shannon entropy of the probability mass function f.

Theorem IV.3 generalizes the bisection policy [73, 30] to multiple players. The fusion rule is a posterior update and by Bayes rule:

$$p_{n+1}(x) \propto \mathbb{P}(\mathbf{Y}_{n+1} = \mathbf{y}_{n+1} | \mathbf{A}_n, X^* = x, \mathcal{F}_n) \times p_n(x) \tag{4.10}$$

where $\mathbf{y}_{n+1} \in \{0,1\}^M$ are the observations at time n. Next we establish that a greedy sequential query strategy achieves the same average entropy reduction as that of the optimal joint query strategy.

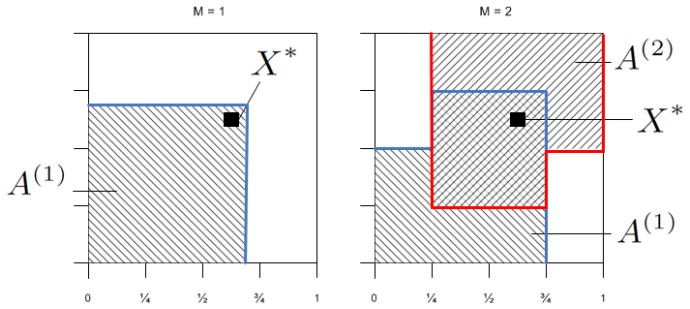


Figure 4.3: Jointly optimal queries under uniform prior for two dimensional target search. The target X^* is indicated by a black square. The one-player bisection rule (left) satisfies the optimality condition (4.12) with optimal query $A^{(1)} = [0, \frac{1}{\sqrt{2}}] \times [0, \frac{1}{\sqrt{2}}]$. The two-player bisection rule (right) satisfies (4.12) with optimal queries $A^{(1)} = [0, \frac{3}{4}] \times [0, \frac{1}{2}] \cup [\frac{1}{4}, \frac{3}{4}] \times [\frac{1}{2}, \frac{3}{4}], A^{(2)} = [\frac{1}{4}, 1] \times [\frac{1}{2}, 1] \cup [\frac{1}{4}, \frac{3}{4}] \times [\frac{1}{4}, \frac{1}{2}]$. We note that using the policy on the left, if player 1 responds that $X^* \in [0, \frac{1}{\sqrt{2}}] \times [0, \frac{1}{\sqrt{2}}]$, with high probability, then the posterior will concentrate on that region. When using the policy on the right, if player 1 and 2 respond that $X^* \in A^{(1)} \cap A^{(2)}$ with high probability, then the posterior will concentrate more on the intersection of the queries, thus better localizing the target as compared with the single player policy.

Theorem IV.4. *(Equivalence, Known Error Probabilities) Under Assumptions IV.1 and IV.2:*

1. *The expected entropy loss under an optimal joint query design is the same as the greedy sequential query design. This loss is given by:*

$$C = \sum_{m=1}^{M} C(\epsilon_m) = \sum_{m=1}^{M} (1 - h_b(\epsilon_m)) \tag{4.11}$$

where $h_b(\epsilon_m) = -\epsilon_m \log(\epsilon_m) - (1-\epsilon_m)\log(1-\epsilon_m)$ *is the binary entropy function.*

2. *All jointly optimal control laws equalize the posterior probability over the dyadic partitions induced by* $\mathbf{A}_n = \{A_n^{(1)}, \ldots, A_n^{(M)}\}$*:*

$$\mathbb{P}_n(R) = \int_R p_n(x)dx = 2^{-M}, \forall R \in \gamma(\mathbf{A}_n). \tag{4.12}$$

where the set $\gamma(\cdot)$ *was defined in (4.8).*

Thm. IV.4 shows that the optimal joint policy can be determined and implemented using the simpler greedy sequential query design. Note that, despite the fact that all players are conditionally independent, the joint policy does not decouple into separate single-player optimal policies. This is analogous to the non-separability of the optimal vector-quantizer in source coding even for independent sources [55]. In addition, the optimal queries must be overlapping-i.e., $\bigcap_{m=1}^{M} A_n^{(m)} \neq \emptyset$, but not identical. Finally, we remark that the optimal query $\mathbf{A}_n$ is not unique, so it is possible that there exists an even simpler optimal control law than the sequential greedy policy.

Equivalence: Intuition

A simple intuitive way to see the equivalence property stated in Thm. IV.4 is through the chain rule of the mutual information. Consider the joint query strategy and its associated filtration $\mathcal{F}_n$. According to Theorem IV.3, the optimal policy is to choose the queries such that the conditional mutual information is maximized. The

chain rule of conditional mutual information [35] implies:

$$I(X^*; \mathbf{Y}_{n+1}|\mathbf{A}_n, \mathcal{F}_n) = I(X^*; Y_{n+1}^{(1)}|A_n^{(1)}, \mathcal{F}_n) + \sum_{m=2}^{M} I(X^*; Y_{n+1}^{(m)}|A_n^{(m)}, \{A_n^{(k)}, Y_{n+1}^{(k)}\}_{k=1}^{m-1}, \mathcal{F}_n)$$

which relates the joint mutual information of the LHS (as in the joint scheme) to the mutual information of each player conditioned on the responses of the previous players (as in the sequential scheme). Letting $M = 2$ for concreteness, we observe:

$$\begin{aligned} &I(X^*; Y_{n+1}^{(1)}, Y_{n+1}^{(2)}|A_n^{(1)}, A_n^{(2)}, \mathcal{F}_n) \\ &\quad = I(X^*; Y_{n+1}^{(2)}|Y_{n+1}^{(1)}, A_n^{(2)}, A_n^{(1)}, \mathcal{F}_n) + I(X^*; Y_{n+1}^{(1)}|A_n^{(1)}, \mathcal{F}_n) \end{aligned}$$

This relation implies that the mutual information between the target X^* and the response $Y_{n+1}^{(2)}$ of the second player depends on the response of the first player $Y_{n+1}^{(1)}$. It follows that the information available for query design $A_n^{(2)}$ for the second player is larger than the information available for query design $A_n^{(1)}$ for the first player.

Equivalence: One-dimensional Example

As a specific example, let us consider the one-dimensional case with $M = 2$ collaborating players. Consider the query design problem for this case. We assume that the prior density p_0 is uniform over the position of a target in one dimension, i.e., the target state is in the domain $\mathcal{X} = [0, 1]$. We define the queries as intervals- i.e., $A_n^{(1)} = [a, b]$ and $A_n^{(2)} = [c, d]$. The optimal policy (4.12) requires the queries to be overlapping thus we impose the constraints $a < c, c < b$ and $b < d$. Choosing $a = 1/8, b = 1/2 + 1/8, c = 1/2 - 1/8$ and $d = 1 - 1/8$, we observe that the optimality conditions in (4.12) are satisfied over the dyadic partition set $\gamma(\mathbf{A}_n) = \{A_n^{(1)} \cap A_n^{(2)}$, $A_n^{(1)} \cap \overline{A}_n^{(2)}$, $A_n^{(1)} \cap \overline{A}_n^{(2)}$ and $\overline{A}_n^{(1)} \cap \overline{A}_n^{(2)}\}$. Thus, this is a jointly optimal law and is illustrated graphically in Fig. 4.4 (a). We note that the region of uncertainty has size 1/4 (region not covered by queries).

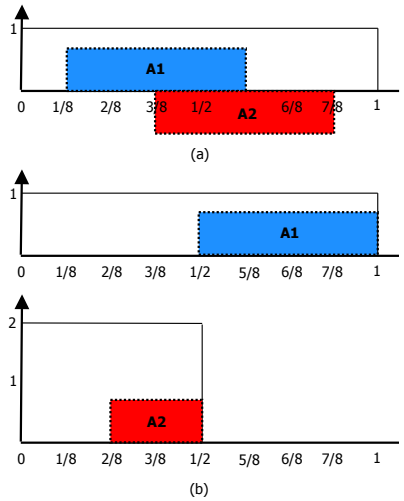


Figure 4.4: Illustration of jointly optimal policy (a) and sequential policy (b) for one-dimensional target, uniformly distributed over $[0, 1]$, and two players. In each case the total length of the intervals not covered by the queries (uncertainty) is equal to $1/4$.

The sequential policy consists of a sequence of bisections. This policy is illustrated in Fig. 4.4 (b) and the region of uncertainty also has size $1/4$.

4.3 Mean-Square Error Performance Bounds

In this section, we provide exponential lower and upper bounds on the MSE of the sequential Bayesian estimator.

4.3.1 Lower Bounds via Entropy Loss

Thm. IV.4 yields the value of the cooperative game in terms of expected entropy reduction, which is the sum of the "capacities" [2] of all the players. This value function is used next to provide a lower bound on the MSE of the sequential Bayesian estimator.

Theorem IV.5. *(Lower Bound on MSE) Let Assumptions IV.1, IV.2 hold. Assume the entropy $H(p_0)$ is finite. Then, the MSE of the joint or sequential query policies*

[2] The "capacity" of each player is the Shannon channel capacity of each BSC [35].

in Thm. 1 and 2 satisfies:

$$\frac{K}{2\pi e} d \exp\left(-\frac{2nC}{d}\right) \leq \mathbb{E}[\| X^* - X_n \|_2^2] \tag{4.13}$$

where $K = e^{2H(p_0)}$ *and* X_n *is the posterior mean. The expected entropy loss per iteration is* $C = \sum_m C(\epsilon_m)$.

Observe that the bound in (4.13) holds for any policy π and for optimal policies π^*, the bound becomes tighter since $\mathbb{E}^{\pi}[H(p_n)] = H(p_0) - nC$ for this case. We also note that the bound behaves exponentially as a function of the number of queries n with rate exponent given by the sum of the capacities C.

4.3.2 Upper Bounds

The performance analysis of the bisection method is difficult primarily due to the continuous nature of the posterior [30]. A discretized version of the probabilistic bisection method was proposed in [20], using the Burnashev-Zingagirov (BZ) algorithm, which imposes a piecewise constant structure on the posterior. A description of the BZ algorithm and its convergence rate is given in [30] (also see App. A in [29]). For simplicity of discussion, we assume the target location is constrained to the unit interval $\mathcal{X} = [0, 1]$. The generalization to $d > 1$ is a difficult open problem. A step size $\Delta > 0$ is defined such that $\Delta^{-1} \in \mathbb{N}$ and the posterior after j iterations is $p_j : \mathcal{X} \to \mathbb{R}$, given by

$$p_j(x) = \frac{1}{\Delta} \sum_{i=1}^{\Delta^{-1}} a_i(j) I(x \in I_i)$$

where $I_1 = [0, \Delta], I_i = ((i-1)\Delta, i\Delta]$ for $i = 2, \ldots, \Delta^{-1}$. We define the discretized posterior at time j as the probability vector $\mathbf{a}(j) = [a_1(j), \ldots, a_{\Delta^{-1}}(j)]$. The initial posterior is $a_i(0) = \Delta, \forall i$. The posterior is characterized completely by the discretized posterior $\mathbf{a}(j)$ which is updated at each iteration via Bayes rule [29].

Convergence rates were derived for the one-dimensional case in [30] for the bounded noise case (i.e., constant error probability) and for the unbounded noise case (i.e., error probability depends on distance from target X^* and converges to $1/2$ as the estimate reaches the target) in [31]. A modified version of this algorithm that is proven to handle unbounded noise was shown in [31]. Thm. IV.7 derives upper bounds on MSE using ideas from [31].

First, we need a simple lemma.

Lemma IV.6. *Let $\hat{X}_n$ be an estimator of target X^* lying in domain $[0,1]$. Then, for all $\Delta \in [0,1]$, we have:*

$$\mathbb{E}[(X^* - \hat{X}_n)^2] \leq \Delta^2 + (1-\Delta^2)\mathbb{P}(|X^* - \hat{X}_n| > \Delta)$$

Theorem IV.7. *(Upper Bound on MSE) Consider the sequential bisection algorithm for M players in one-dimension, where each bisection is implemented using the BZ algorithm. Then, we have:*

$$\mathbb{P}(|X^* - \hat{X}_n| > \Delta) \leq (\frac{1}{\Delta} - 1)\exp\left(-n\bar{C}\right)$$

$$\mathbb{E}[(X^* - \hat{X}_n)^2] \leq (2^{-2/3} + 2^{1/3})\exp\left(-\frac{2}{3}n\bar{C}\right) \tag{4.14}$$

where $\bar{C} = \sum_{m=1}^{M} \bar{C}(\epsilon_m)$, $\bar{C}(\epsilon) = 1/2 - \sqrt{\epsilon(1-\epsilon)}$.

The combination of the lower bound (Thm. IV.5) and the upper bound (Thm. IV.7) imply that the MSE of the BZ algorithm goes to zero at an exponential rate with rate constant between $2C$ and $\frac{2}{3}\bar{C}$.

4.4 Strong Convergence

In this section, we prove almost sure convergence of the sequential bisection schemes in the discretized and the continuous setting.

Corollary IV.8. *(Almost sure convergence for discretized PBA) Consider the sequential bisection algorithm for M players in one-dimension (i.e., $d = 1$), where each bisection is implemented via the BZ algorithm. Then, we have $\hat{X}_n \overset{a.s.}{\to} X^*$ as $n \to \infty$.*

Corollary IV.9. *(Almost sure convergence for continuous PBA) Consider the sequential bisection algorithm for M players in any dimension (i.e., $d \geq 1$), where each bisection is implemented via the standard (continous-space) PBA. Then, we have:*

$$\hat{X}_n \xrightarrow{a.s.} X^*$$

$$\frac{1}{n}\log(p_n(X^*)) \xrightarrow{a.s.} \sum_{i=1}^{M} C(\epsilon_i) = C(\boldsymbol{\epsilon})$$

as $n \to \infty$. The function $C(\epsilon_i)$ denotes the capacity of the ith BSC associated with the ith player.

Corollary IV.9 also yields a pointwise rate of convergence. Specifically, for large n, $p_n(X^*) \sim 2^{nC}$, where C is the sum of capacities. This is intuitive in the sense that the larger the sum capacity, the faster we expect the distribution on the target location to concentrate on the true target X^*.

4.5 Human-in-the-loop

In this section, we consider the 2-player case where player 1 (the machine) has a constant error probability $\epsilon_1 \in (0, 1/2)$ and player 2 (the human) has error probability increasing as the target localization error decreases:

$$\mathbb{P}(Y_{n+1}^{(2)} \neq z | Z_n^{(2)} = z) = \frac{1}{2} - \min(\delta_0, \mu|X^* - X_n|^{\kappa-1}) \tag{4.15}$$

where $\kappa > 1, 0 < \delta_0 < \mu < 1/2$ is a reliability parameter to parameterize the human [3]. Fig. 4.5 illustrates the human error model as a function of $|X^* - X_n|$. This

[3]The parameter κ controls the "resolution" of the human. It becomes increasingly difficult for the human to decide between close hypotheses as κ goes to infinity.

is a popular model used for human-based optimization [72] and active learning of threshold functions [31]. From the nature of the error probability (4.15) we expect that the answers provided by the human will be helpful in the beginning iterations but their value will go to zero as the number of iterations grows to infinity. This is because the human propensity for error becomes larger as the questions become more refined and difficult to resolve.

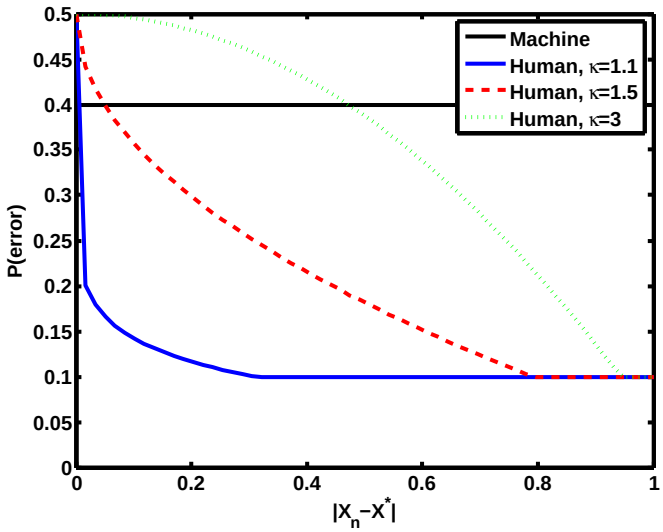


Figure 4.5: Human error probability as a function of distance from target $|X^* - X_n|$ for $\delta_0 = 0.4, \mu = 0.45$ and various $\kappa > 1$.

Using a similar argument as in the proof of Thm. IV.7, and using the modified BZ algorithm [31], from Lemma 1 in [31], we have the following. For $\kappa \geq 2$ with $\alpha_1 = \frac{\sqrt{\epsilon_1}}{\sqrt{\epsilon_1}+\sqrt{1-\epsilon_1}}, \alpha_2 = 0.09\mu(3\Delta/4)^{\kappa-1}$:

$$\mathbb{P}(|X^* - \hat{X}_n| > \Delta) \leq \Delta^{-1} \exp\left(-n\left[\bar{C}(\epsilon_1) + \frac{\mu^2}{50}\left(\frac{3\Delta}{4}\right)^{2\kappa-2}\right]\right).$$

Applying Lemma IV.6, this leads to the MSE upper bound dependent on Δ:

$$\mathbb{E}[(X^* - \hat{X}_n)^2] \leq \Delta^2 + \Delta^{-1} \exp\left(-n\left[\bar{C}(\epsilon_1) + \frac{\mu^2}{50}\left(\frac{3\Delta}{4}\right)^{2\kappa-2}\right]\right) \tag{4.16}$$

With the choice $\Delta = 2^{-1/3}e^{-n\bar{C}(\epsilon_1)/3}$,

$$\begin{aligned}\mathbb{E}[(X^* - \hat{X}_n)^2] &\leq \exp\left(-\frac{2}{3}n\bar{C}(\epsilon_1)\right) \\ &\times \left[2^{-2/3} + 2^{1/3}\exp\left(-\frac{\mu^2}{50}\left(\frac{3\cdot 2^{-1/3}}{4}\right)^{2\kappa-2} n\exp\left(-n\bar{C}(\epsilon_1)\frac{2\kappa-2}{3}\right)\right)\right] \end{aligned} \tag{4.17}$$

which is no greater than the "player 1" (machine alone) MSE bound (compare (4.17) with (4.14)). Asymptotically as $n \to \infty$, the two bounds both converge to zero at the same rate.

We define the human gain ratio (HGR) as the ratio of MSE upper bounds associated with "player 1" and "player 1 + human", respectively.

$$R_n(\kappa) = \frac{2^{-2/3} + 2^{1/3}}{2^{-2/3} + 2^{1/3}\exp\left(-\frac{\mu^2}{50}(\frac{3\cdot 2^{-1/3}}{4})^{2\kappa-2} n\exp\left(-n\bar{C}(\epsilon_1)\frac{2\kappa-2}{3}\right)\right)} \tag{4.18}$$

The HGR is plotted in Fig. 4.6 in root-scale as a function of κ. This analysis quantifies the value of including the human-in-the-loop for a sequential target localization task. We note that the larger ϵ_1 is, the larger is the HGR. Also, as κ decreases to 1, the ratio increases, meaning that the human accuracy approaches that of the machine.

4.6 Noisy 20 Questions with Collaborative Players: Unknown Error Probability

In this section we consider the setting where the error probabilities of the M players are unknown. In this case, the Bayes posterior update is not well-defined, so the probabilistic bisection algorithm cannot be directly used. In the most generic setting of having unknown $\epsilon_m \in (0, 1/2)$, we propose a joint estimation scheme to estimate the target X^* and the error probabilities $\boldsymbol{\epsilon}^* = (\epsilon_1^*, \ldots, \epsilon_M^*)$. The method propagates the joint posterior distribution of the joint random vector $(X^*, \boldsymbol{\epsilon}^*)$ forward in time given the designed queries and noisy responses. The joint posterior

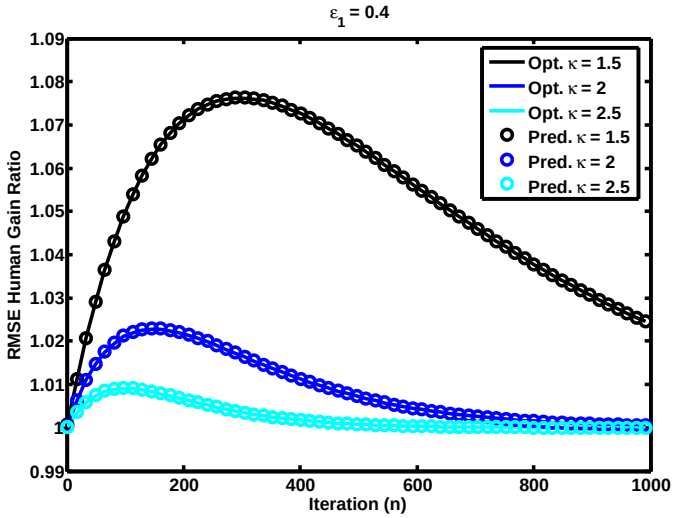


Figure 4.6: Human gain ratio $\sqrt{R_n(\kappa)}$ (see Eq. (4.18)) as a function of κ. The human provides the largest gain in the beginning few iterations and the value of information decreases as $n \to \infty$. The circles are the predicted curves according to (4.17), while the solid lines are the optimized versions of the bound (4.16) (as a fuction of Δ) for each n. The predictions well match the optimized bounds.

distribution is considered here because the error probabilities ϵ_m are coupled with the target x through the Bayesian update (see Eqns. (4.7) and (4.10)).

Define the random vector $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_M) \in [0, 1/2)^M$ and the joint posterior distribution $\mathbb{P}(X^* = x, \boldsymbol{\epsilon}^* = \boldsymbol{\epsilon} | \mathcal{F}_n) = p_n(x, \boldsymbol{\epsilon})$. We consider the minimum expected entropy criterion (4.1).

4.6.1 Assumptions

We make an analogous conditional independence assumption to Assumption IV.1 for the unknown channel case.

Assumption IV.10. *We assume that the players' responses are conditionally independent:*

$$\begin{aligned}
&\mathbb{P}(\mathbf{Y}_{n+1} = \mathbf{y} | \mathbf{A}_n, X^* = x, \boldsymbol{\epsilon}^* = \boldsymbol{\epsilon}, \mathcal{F}_n) \\
&\quad = \prod_{m=1}^{M} \mathbb{P}(Y_{n+1}^{(m)} = y^{(m)} | A_n^{(m)}, X^* = x, \epsilon_m^* = \epsilon_m, \mathcal{F}_n)
\end{aligned}$$

where

$$\mathbb{P}(Y_{n+1}^{(m)} = y^{(m)} | A_n^{(m)}, X^* = x, \epsilon_m^* = \epsilon_m, \mathcal{F}_n)$$
$$= \begin{cases} f_1^{(m)}(y^{(m)}|\epsilon_m, A_n^{(m)}, \mathcal{F}_n), & x \in A_n^{(m)} \\ f_0^{(m)}(y^{(m)}|\epsilon_m, A_n^{(m)}, \mathcal{F}_n), & x \notin A_n^{(m)} \end{cases}.$$

4.6.2 Sequential Query Design

In the sequential setup, we assume that the fusion center designs queries for each of the M sensors in sequence and refines the posterior belief of the target location given the response of each player (see Fig. 4.1). Recall the sub-time scale of sub-instants $\{n_t : t = 0, \ldots, M-1\}$ for each time instant n and consider the notation and filtration $\mathcal{G}_{n_t}$ defined in Section 4.2.A. Assuming that all sensors are queried in sequence starting from $m = 1$ and ending at $m = M$, the posterior updates (after querying the $(t+1)$th player) become:

$$p_{n_{t+1}}(x, \boldsymbol{\epsilon}) = \mathbb{P}(Y_{n_{t+1}} = y_{n_{t+1}} | A_{n_t}, X^* = x, \epsilon_{t+1}^* = \epsilon_{t+1}, \mathcal{G}_{n_t}) \times p_{n_t}(x, \boldsymbol{\epsilon})$$

$$\mathbb{P}(Y_{n_{t+1}} = y_{n_{t+1}} | A_{n_t}, X^* = x, \epsilon_{t+1}^* = \epsilon_{t+1}, \mathcal{G}_{n_t}) = \begin{cases} f_1^{(t+1)}(y_{n_{t+1}}|\epsilon_{t+1}), & x \in A_{n_t} \\ f_0^{(t+1)}(y_{n_{t+1}}|\epsilon_{t+1}), & x \notin A_{n_t} \end{cases}.$$

4.6.3 Joint Query Design

In the joint setup, we assume that the fusion center designs queries for the M sensors at each time instant n and after querying all sensors, the responses are fused by the controller and the next set of questions is formulated. Recall the notation and filtration $\mathcal{F}_n$ defined in Section 4.2.B.

Define the density parameterized by $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_M)$ and $i_1, \ldots, i_M \in \{0, 1\}$:

$$g_{i_1:i_M}(\mathbf{y}|\boldsymbol{\epsilon}) = \prod_{m=1}^{M} f_{i_m}^{(m)}(y^{(m)}|\epsilon_m)$$

At the nth time instant, the posterior update becomes:

$$p_{n+1}(x,\boldsymbol{\epsilon}) = \mathbb{P}(\mathbf{Y}_{n+1} = \mathbf{y}_{n+1}|\mathbf{A}_n, X^* = x, \boldsymbol{\epsilon}^* = \boldsymbol{\epsilon}, \mathcal{F}_n) \times p_n(x,\boldsymbol{\epsilon})$$

$$\mathbb{P}(\mathbf{Y}_{n+1} = \mathbf{y}_{n+1}|X^* = x, \boldsymbol{\epsilon}^* = \boldsymbol{\epsilon}, \mathcal{F}_n) = \prod_{m=1}^{M} \begin{cases} f_1^{(m)}(y_{n+1}^{(m)}|\epsilon_m), & x \in A_n^{(m)} \\ f_0^{(m)}(y_{n+1}^{(m)}|\epsilon_m), & x \notin A_n^{(m)} \end{cases}.$$

4.6.4 Equivalence Theorems

Since the error probabilities of sensors are unknown, the joint policy derived in Theorem IV.3 is no longer applicable or valid. The next theorem derives the jointly optimal policy for all sensors under the unknown channel case.

Theorem IV.11. *(Jointly Optimal Policy, Unknown Error Probabilities) Let Assumptions IV.1 and IV.2 hold. Consider the problem (4.1), where the joint policy is made up of the query regions for the M sensors.*

1. *Optimal policies $\mathbf{A}_n = (A_n^{(1)}, \ldots, A_n^{(M)})$ at time n satisfy:*

$$G_n^* = \sup_{A^{(1)},\ldots,A^{(M)}} \Big\{ H\left(\sum_{i_1:i_M=0}^{1} \int_{\boldsymbol{\epsilon}=0}^{1/2} g_{i_1:i_M}(\cdot|\boldsymbol{\epsilon}) \mathbb{P}_n \left(\bigcap_m (A^{(m)})^{i_m}, \boldsymbol{\epsilon} \right) d\boldsymbol{\epsilon} \right) - \sum_{i_1:i_M=0}^{1} \int_{\boldsymbol{\epsilon}=0}^{1/2} H(g_{i_1:i_M}(\cdot|\boldsymbol{\epsilon})) \mathbb{P}_n \left(\bigcap_m (A^{(m)})^{i_m}, \boldsymbol{\epsilon} \right) d\boldsymbol{\epsilon} \Big\} \tag{4.19}$$

2. *The maximum information gain at time n is:*

$$G_n^* = \sum_{m=1}^{M} \mathbb{E}[C(\epsilon_m)|\mathcal{F}_n] \tag{4.20}$$

where $\mathbb{E}[C(\epsilon_m)|\mathcal{F}_n] = \int_{\epsilon_m=0}^{1/2} C(\epsilon_m) p_n(\epsilon_m) d\epsilon_m$.

Next, we show a version of the equivalence theorem (Theorem IV.4) for the unknown channel case.

Theorem IV.12. *(Equivalence, Unknown Error Probabilities) Let Assumptions IV.10 and IV.2 hold. Consider the sequential and joint schemes described in Section*

4.6.B and Section 4.6.C. [4] *Then, it follows that* $G^*_{seq,n} = \mathbb{E}[\sum_m C(\epsilon_m)|\mathcal{G}_n]$ *and* $G^*_n = \mathbb{E}[\sum_m C(\epsilon_m)|\mathcal{F}_n]$ *for all* n.

Lower Bound on MSE Performance

The maximum entropy loss derived in Thm. IV.11 is used next to provide a lower bound on the MSE of the joint sequential estimator.

Theorem IV.13. *(Lower bound on Joint MSE) Assume* $H(p_0)$ *is finite. Then, the joint MSE of the joint query policy in Thm. IV.11 satisfies:*

$$\frac{K}{2\pi e} d \exp\left(-\frac{2n\bar{C}_n}{d}\right) \leq \mathbb{E}[\|X_n - X^*\|_2^2] + \mathbb{E}[\|\boldsymbol{\epsilon}_n - \boldsymbol{\epsilon}^*\|_2^2] \tag{4.21}$$

where $K = \exp(2H(p_0))$ *is a constant and* $X_n = \mathbb{E}[X^*|\mathcal{F}_n], \boldsymbol{\epsilon}_n = \mathbb{E}[\boldsymbol{\epsilon}^*|\mathcal{F}_n]$. *The expected entropy loss per iteration is* $\bar{C}_n = \frac{1}{n}\sum_{k=0}^{n-1} G^*_k$.

The proof follows using the result of part 2) of Theorem IV.11 and similar bounding arguments as Theorem IV.5.

4.6.5 Discussion

The jointly optimal policy derived for the unknown probability case in Thm. IV.11 is reminiscent of the jointly optimal policy of Thm. IV.3. We remark that in the unknown probability setting, the maximum entropy loss G^*_n given in (4.19) is not time-invariant, unlike in the case of known probability, in which the maximum entropy loss was the sum of the capacities of the players' channels (4.9) and (4.11). This observation motivates a sensor selection scheme; if we have the hard constraint that only one sensor may be used at a time, then, unlike in the known probability case, it may be that at different times, the maximal information gain may be obtained by different sensors.

[4] For the one-dimensional case, the sequential scheme implements (4.23) for each sub-instant to design a question for each player and the posterior is updated in sequence (see Fig. 4.1).

4.6.6 Sensor Selection Scheme

We assume that at each time instant, only one sensor can be queried. We assume that the control $u_n = u$ implies that the uth sensor is to be queried at time n and $A_n^{(u)} = A$ is the associated query region. Similarly to (4.10), the joint posterior update in this case becomes:

$$p_{n+1}(x, \boldsymbol{\epsilon}) \propto \quad \mathbb{P}(Y_{n+1}^{(u)} | u_n = u, A_n^{(u)}, X^* = x, \epsilon_u^* = \epsilon_u) p_n(x, \boldsymbol{\epsilon})$$

$$\mathbb{P}(Y_{n+1}^{(u)} = y^{(u)} | u_n = u, A_n^{(u)}, X^* = x, \epsilon_u^* = \epsilon_u) = \begin{cases} f_1^{(u)}(y^{(u)}|\epsilon_u), & x \in A_n^{(u)} \\ f_0^{(u)}(y^{(u)}|\epsilon_u), & x \notin A_n^{(u)} \end{cases}$$

Theorem IV.14. *(Sensor Selection Policy, Unknown Error Probabilities) Consider the problem (4.1), where the policy consists of which sensor to choose and the associated query region. At each time n:*

1. *All optimal query policies satisfy:*

$$\max_{u \in \{1,\ldots,M\}} G_n^*(u) = \sup_A \Big\{ H\left(\int_{\epsilon_u=0}^{1/2} f_1(\cdot|\epsilon_u) \mathbb{P}_n^{(u)}(A, \epsilon_u) + f_0(\cdot|\epsilon_u) \mathbb{P}_n^{(u)}(A^c, \epsilon_u) d\epsilon_u \right)$$

(4.22)

$$- \int_{\epsilon_u=0}^{1/2} H\left(f_1(\cdot|\epsilon_u)\right) \mathbb{P}_n^{(u)}(A, \epsilon_u) + H\left(f_0(\cdot|\epsilon_u)\right) \mathbb{P}_n^{(u)}(A^c, \epsilon_u) d\epsilon_u \Big\}$$

2. *The maximum entropy loss is:*

$$G_n^* = \max_u G_n^*(u) = \max_u \mathbb{E}[C(\epsilon_u)|\mathcal{F}_n]$$

The optimal policy for the minimum expected entropy criterion (4.1) shown in Thm. IV.14 is intuitive. The sensor u with the maximum information gain (or entropy loss) is chosen, where the entropy loss is measured as a function of the uth sub-marginal distribution $p_n^{(u)}(x, \epsilon_u)$. While the form (4.22) bears some similarity to the form (4.9), the bisection policy is no longer optimal. In addition, in this

unknown probability setting, it may not always be the case that the sensor with the largest capacity will be chosen (this would be the case in the known probability setting). The integral with respect to $d\epsilon$ over $\epsilon \in [0, 1/2)$ essentially averages out the contribution of the unknown error probabilities with respect to the observed data up to the current time n.

One-dimensional Case

The next corollary specifies the form of the optimal policy derived in Thm. IV.14 for one-dimensional targets. For simplicity, consider the unit interval $\mathcal{X} = [0, 1]$ as the target domain.

Corollary IV.15. *(Sensor Selection Policy, Unknown Error Probabilities, One-dimensional Target) Consider the problem (4.1) for the optimal sensor and query selection policy. Consider the query regions $A_n = [0, x_n]$. The optimal sensor u and associated query region $A = [0, x]$ at time n is given by:*

$$\max_u \left\{ \max_{x \in [0,1]} h_B(g_{1,n}^{(u)}(x)) - c_n^{(u)} \right\} \tag{4.23}$$

where $h_B(\cdot)$ is the binary entropy function [35] and

$$\begin{aligned} c_n^{(u)} &= \int_{\epsilon_u=0}^{1/2} h_B(\epsilon_u) p_n^{(u)}(\epsilon_u) d\epsilon_u \\ g_{1,n}^{(u)}(x) &= \int_0^x \mu_n^{(u)}(t) dt + \int_x^1 (p_n(t) - \mu_n^{(u)}(t)) dt \\ \mu_n^{(u)}(t) &= \int_{\epsilon_u=0}^{1/2} \epsilon_u p_n^{(u)}(t, \epsilon_u) d\epsilon_u \\ p_n(t) &= \int_{\epsilon_1=0}^{1/2} \cdots \int_{\epsilon_M=0}^{1/2} p_n(t, \epsilon_1, \ldots, \epsilon_M) d\epsilon_1 \cdots d\epsilon_M \end{aligned}$$

We note that the optimal policy derived for the unknown probability case in (4.23) is *not* equivalent to the probabilistic bisection policy-i.e., obtaining $\mathbb{P}_n^{(u)}([0, x_n^{(u)}]) = 1/2$ for each sensor u and then evaluating the information gain and choosing the

sensor with the maximum information gain. This heuristic scheme would yield a suboptimal information gain as compared to the maximal information gain given by (4.23). Thus, in the unknown probability setting, the optimal control law is no longer equivalent to the known probability setting (after marginalizing out the noise parameters $\epsilon_1, \ldots, \epsilon_M$). This result shows that the two settings are quite different and the answers to the unknown channel case are more complex. We empirically observed that there is a unique query point $x = x_n^* = x_n^{(u^*)}$ that maximizes the function (4.23). This is similar to the one-dimensional case for the known probability setting when the query region is of the form $A = [0, x]$; i.e., the optimal point is the median.

4.7 Simulations

This section contains a few illustrative simulations that validate the methodology presented throughout the chapter.

4.7.1 Known Error Probability

Figs. 4.7 and 4.8 show the empirical performance of the human-in-the-loop by comparing the actual MSEs of "player 1" and of "player 1 + human", for the cases of uniform and nonuniform prior distributions on the target location. It is observed that employing a human in the loop reduces the RMSE relative to only having "player 1" and to having "player 1 + player 2" for a wide range of n. We note that as $n \to \infty$, the "player 1 + human" curve will cross the "player 1 + player 2" curve, being consistent with the upper bounds shown in (4.14) and (4.17) since the human's contribution is strongest in the first few iterations, while its value decreases to zero as $n \to \infty$. Also, the human model does not yield a different exponent in the exponential rate of convergence, while adding a second player does as predicted in Thms. IV.5 and IV.7.

Next, we observe the effect of the prior distribution associated with the target location on the RMSE performance. We observe that the "player 1 + human" scheme provides a larger gain when the initial distribution is trimodal with larger variance on the true component centered at $X^* = 0.75$ (see Fig. 4.9) as shown in Fig. 4.8, as compared to the gain from starting from a uniform distribution as shown in Fig. 4.7. In fact, the human-in-the-loop combined with player 1 outperforms two players 1 and 2 for a wide range of iterations n.

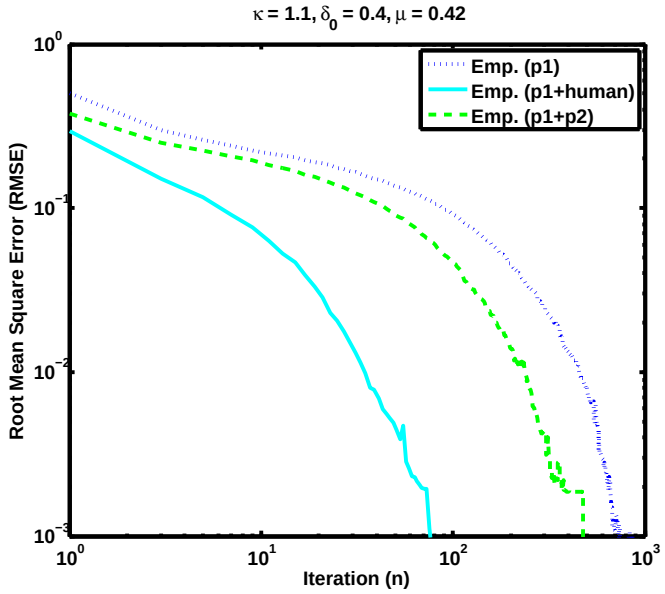


Figure 4.7: Monte Carlo simulation for RMSE performance of the sequential estimator as a function of iteration. 8000 Monte Carlo trials were used. The human parameters were set to $\kappa = 1.1, \mu = 0.42, \delta_0 = 0.4$, the players' parameters were $\epsilon_1 = \epsilon_2 = 0.4$, and the length of pseudo-posterior was $\Delta^{-1} = 1618$. The target was set to $X^* = 0.75$. The initial distribution was uniform. The parameters $0 < \mu < \delta_0 < 1/2$ were chosen such that the smallest error probability would be $1/2 - \delta_0 = 0.1$ and the resolution parameter $\kappa > 1$ was chosen close to 1 in order to show a large enough gain for including the human. As κ grows, the RMSE gain contributed by the human decreases.

Figures 4.10 and 4.11 show the empirical RMSE as a function of $\epsilon_1 \in (0, 1/2)$ for $\kappa = 2.0$ and $\kappa = 1.5$, respectively. As expected, larger MSE gains are obtained for $\kappa = 1.5$. For fixed κ, we observe from both figures that the MSE associated with just "player 1" increases as ϵ_1 increases, and in addition, the RMSE associated with

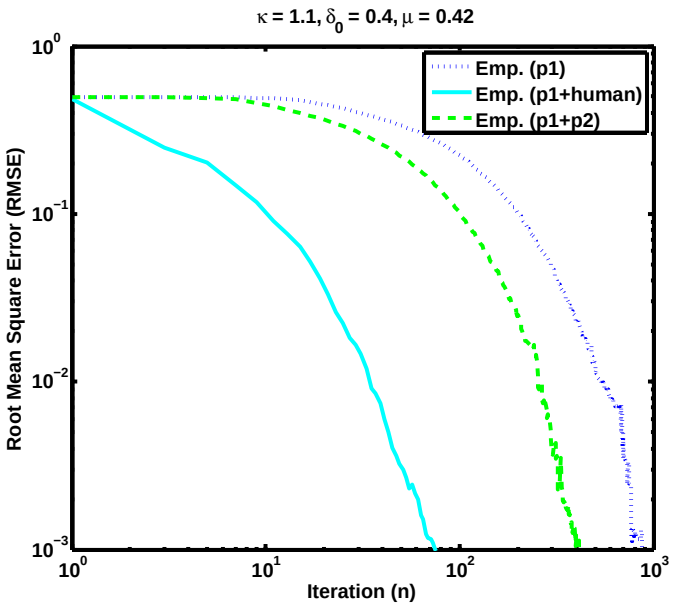


Figure 4.8: Monte Carlo simulation for RMSE performance of the sequential estimator as a function of iteration. 8000 Monte Carlo trials were used. The human parameters were set to $\kappa = 1.1, \mu = 0.42, \delta_0 = 0.4$, the players' parameters were $\epsilon_1 = \epsilon_2 = 0.4$, and the length of pseudo-posterior was $\Delta^{-1} = 1618$. The target was set to $X^* = 0.75$. The initial distribution was a mixture of three Gaussian distributions as shown in Fig. 4.9. The parameters $0 < \mu < \delta_0 < 1/2$ were chosen such that the smallest error probability would be $1/2 - \delta_0 = 0.1$ and the resolution parameter $\kappa > 1$ was chosen close to 1 in order to show a large enough gain for including the human. As κ grows, the RMSE gain contributed by the human decreases.

"player 1 + human" yields a larger improvement over just using player 1 for larger ϵ_1. In other words, the worse player 1 is, the larger the value of the human in reducing the MSE.

4.7.2 Unknown Error Probability

Fig. 4.12 numerically evaluates the MSE performance for $M = 1$ sensor with unknown error probability. This simulation implies that the binary responses obtained from one player carry enough information to both estimate the target accurately and its error probability.

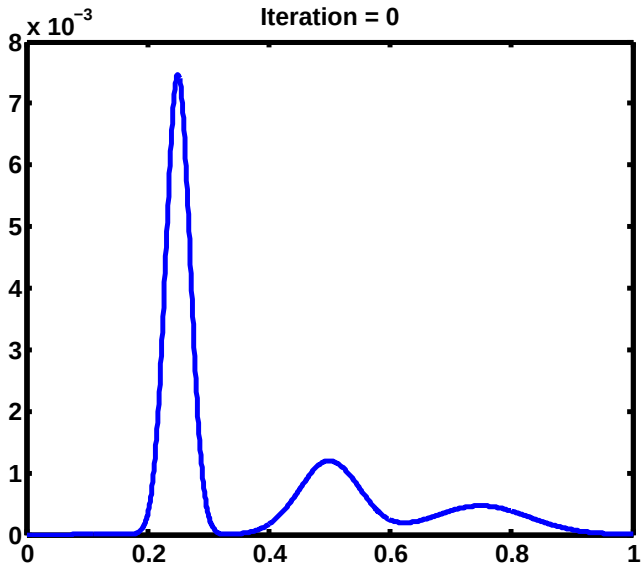


Figure 4.9: Initial distribution for BZ algorithm. The distribution is a mixture of three Gaussians with means 0.25, 0.5 and 0.75, and variances 0.02, 0.05 and 0.08, respectively. The target was set to be the center of the mode at $X^* = 0.75$ with the largest variance. The resulting MSE performance of the sequential estimator is shown in Fig. 4.8.

4.8 Conclusion

We studied the problem of collaborative 20 questions with noise for the multiplayer case. We derived an equivalence theorem that shows the joint query design has the same performance on average as the sequential bisection query design, despite the fact that the sequential bisection query design has access to a more refined filtration. In addition, the sequential bisection query design is easily implemented due to the low complexity of the controllers (unlike the jointly optimal design). Using this framework, we obtained mean-square-error bounds for the performance of the sequential estimator. The methodology was applied to human-in-the-loop target localization systems.

The framework was generalized to the case of unknown error probabilities associated with noisy players. For this case, it was shown that the maximum entropy loss per iteration is time-varying (unlike in the known probability case) and the op-

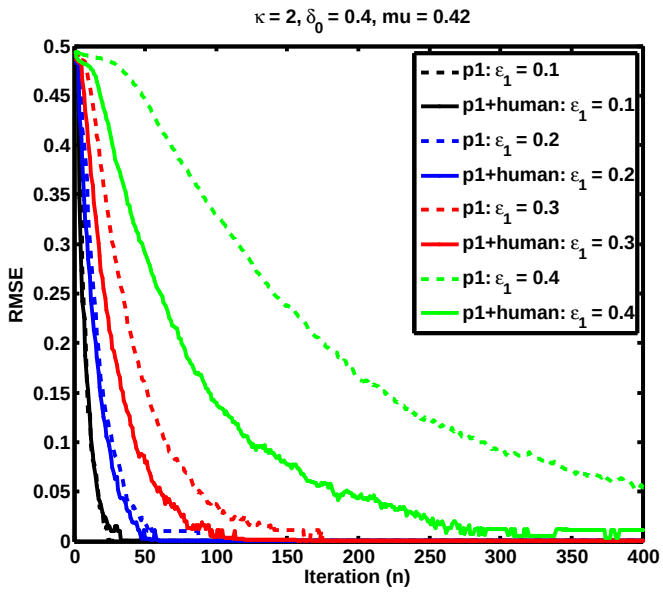


Figure 4.10: Monte Carlo simulation for RMSE performance of the sequential estimator as a function of iteration and $\epsilon_1 \in (0, 1/2)$. 2000 Monte Carlo trials were used. The human parameters were set to $\kappa = 2.0, \mu = 0.42, \delta_0 = 0.4$, the length of pseudo-posterior was $\Delta^{-1} = 1618$. The target was set to $X^* = 0.75$. The initial distribution was a mixture of three Gaussians as shown in Fig. 4.9. The parameters $0 < \mu < \delta_0 < 1/2$ were chosen such that the smallest error probability would be $1/2 - \delta_0 = 0.1$.

timal policy that achieves this gain is not equivalent to the probabilistic bisection policy. Simulations were provided to numerically evaluate the performance of the proposed sequential estimator. Worthwhile future work could include the following extensions: 1) query design for target detection and classification; 2) more sophisticated human/machine response models that account for state-dependent response (channel) errors.

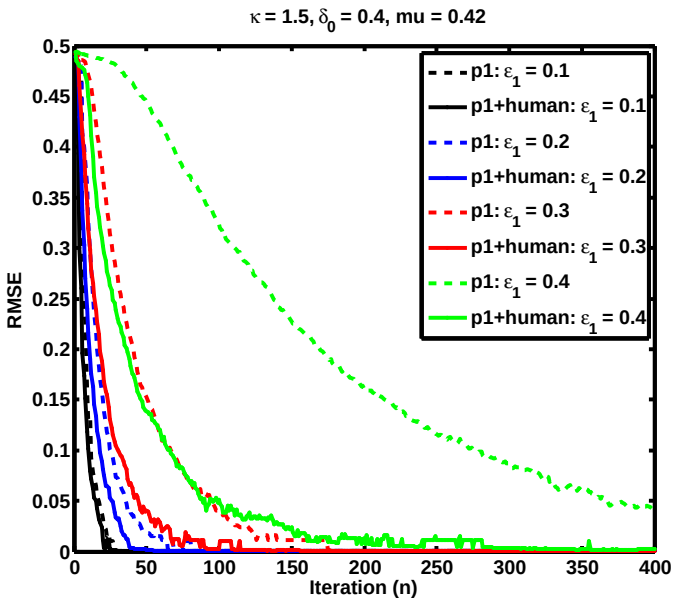


Figure 4.11: Monte Carlo simulation for RMSE performance of the sequential estimator as a function of iteration and $\epsilon_1 \in (0, 1/2)$. 2000 Monte Carlo trials were used. The human parameters were set to $\kappa = 1.5, \mu = 0.42, \delta_0 = 0.4$, the length of pseudo-posterior was $\Delta^{-1} = 1618$. The target was set to $X^* = 0.75$. The initial distribution was a mixture of three Gaussians as shown in Fig. 4.9.

4.9 Appendix

4.9.1 Proof of Theorem IV.3

Proof. Using (4.6) and (4.7), we have:

$$\begin{aligned}
&\mathbb{P}(\mathbf{Y}_{n+1} = \mathbf{y}|\mathbf{A}_n, X^* = x, \mathcal{F}_n) \\
&= \prod_{m=1}^{M} \Big\{ f_1^{(m)}(y^{(m)}|A_n^{(m)}, \mathcal{F}_n) I(x \in A_n^{(m)}) \\
&\quad + f_0^{(m)}(y^{(m)}|A_n^{(m)}, \mathcal{F}_n) I(x \notin A_n^{(m)}) \Big\} \\
&= \sum_{i_1:i_M=0}^{1} g_{i_1:i_M}(\mathbf{y}|\mathbf{A}_n, \mathcal{F}_n) I\left(x \in \bigcap_{m=1}^{M} (A_n^{(m)})^{i_m} \right). \qquad (4.24)
\end{aligned}$$

By integrating over $x \in \mathcal{X}$, we have:

$$\begin{aligned}
&\mathbb{P}(\mathbf{Y}_{n+1} = \mathbf{y}|\mathbf{A}_n, \mathcal{F}_n) = \mathbb{E}[\mathbb{P}(\mathbf{Y}_{n+1} = \mathbf{y}|\mathbf{A}_n, X^*, \mathcal{F}_n)] \\
&= \sum_{i_1:i_M=0}^{1} g_{i_1:i_M}(\mathbf{y}|\mathbf{A}_n, \mathcal{F}_n) \mathbb{P}_n \left(\bigcap_{m=1}^{M} (A_n^{(m)})^{i_m} \right). \qquad (4.25)
\end{aligned}$$

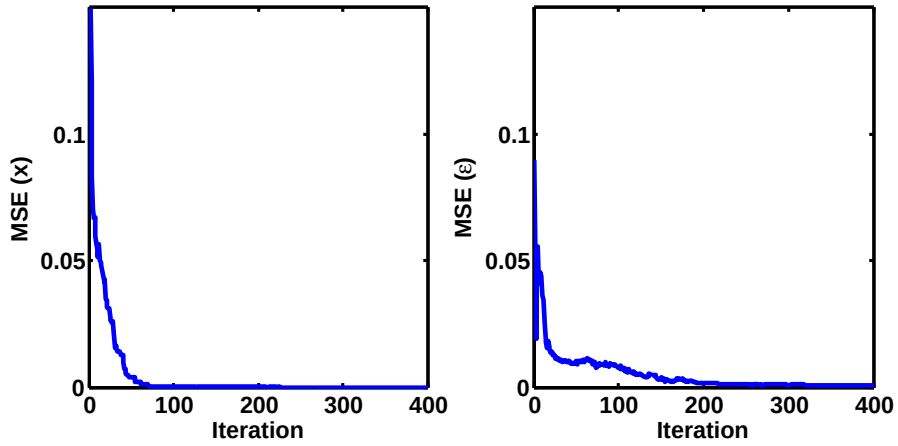


Figure 4.12: Monte Carlo simulation for MSE performance of the joint sequential estimator (of the target X^* and the error probability ϵ^*). The MSE for X is shown on the left and MSE for ϵ on the right, as a function of iteration. 100 Monte Carlo trials were used. The true error probability was set to $\epsilon^* = 0.3$ and the true target location was $X^* = 0.75$. The target was set to $X^* = 0.75$. The initial distribution was a joint uniform density $p_0(x, \epsilon)$.

Similarly to the proof of Thm. 1 in [73], we have:

$$H(p_n) - \mathbb{E}[H(p_{n+1})|\mathbf{A}_n, \mathcal{F}_n] = I(X^*; \mathbf{Y}_{n+1}|\mathbf{A}_n, \mathcal{F}_n)$$
$$= H(\mathbf{Y}_{n+1}|\mathbf{A}_n, \mathcal{F}_n) - \mathbb{E}[H(\mathbf{Y}_{n+1})|X^*, \mathbf{A}_n, \mathcal{F}_n].$$

From (4.25), we have:

$$H(\mathbf{Y}_{n+1}|\mathbf{A}_n, \mathcal{F}_n) = H\left(\sum_{i_1:i_M=0}^{1} g_{i_1:i_M}(\cdot)\mathbb{P}_n\left(\bigcap_{m=1}^{M}(A_n^{(m)})^{i_m}\right)\right)$$

and using (4.24):

$$\mathbb{E}[H(\mathbf{Y}_{n+1})|X^*, \mathbf{A}_n, \mathcal{F}_n]$$
$$= \int_{\mathcal{X}} p_n(x) H(\mathbf{Y}_{n+1}|X^* = x, \mathbf{A}_n, \mathcal{F}_n) dx$$
$$= \sum_{i_1:i_M=0}^{1} H\left(g_{i_1:i_M}\right) \mathbb{P}_n\left(\bigcap_{m=1}^{M}(A_n^{(m)})^{i_m}\right).$$

Putting this together, and using a dynamic programming argument similar to Thm. 2 in [73], it follows that the optimal query satisfies (4.9). □

4.9.2 Proof of Theorem IV.4

Proof. Let G_{seq} denote the maximum expected entropy loss after querying M players sequentially. The bisection policy yields an expected entropy loss of $C(\epsilon_m) = 1 - h_b(\epsilon_m)$ [5] after querying the mth player [73]. Thus, $G_{seq} = \sum_{m=1}^{M} C(\epsilon_m)$. The expected entropy loss at sub-time instant n_t is $H(p_{n_t}) - \mathbb{E}[H(p_{n_{t+1}})|A_{n_t}, \mathcal{G}_{n_t}] = I(X^*; Y_{n_{t+1}}|A_{n_t}, \mathcal{G}_{n_t})$. To show this rigorously, observe:

$$\begin{aligned}
G_{seq} &= \sup_{\{A_{n_t}\}_{t=0}^{M-1}} \mathbb{E}[H(p_n) - H(p_{n+1})|\mathcal{G}_n] \\
&= \sup_{\{A_{n_t}\}_{t=0}^{M-1}} \mathbb{E}\left[\sum_{t=0}^{M-1} H(p_{n_t}) - H(p_{n_{t+1}})\middle|\mathcal{G}_n\right] \\
&= \sup_{\{A_{n_t}\}_{t=0}^{M-1}} \sum_{t=0}^{M-1} \mathbb{E}\left[\mathbb{E}\left[H(p_{n_t}) - H(p_{n_{t+1}})\middle|A_{n_t}, \mathcal{G}_{n_t}\right]\middle|\mathcal{G}_n\right] \\
&= \sup_{\{A_{n_t}\}_{t=0}^{M-1}} \mathbb{E}\left[\sum_{t=0}^{M-1} I(X^*; Y_{n_{t+1}}|A_{n_t}, \mathcal{G}_{n_t})\middle|\mathcal{G}_n\right] \\
&= \mathbb{E}\left[\sum_{t=0}^{M-1} \sup_{A_{n_t}} I(X^*; Y_{n_{t+1}}|A_{n_t}, \mathcal{G}_{n_t})\middle|\mathcal{G}_n\right] \\
&= \mathbb{E}\left[\sum_{t=0}^{M-1} C(\epsilon_{t+1})\middle|\mathcal{G}_n\right] = \sum_{m=1}^{M} C(\epsilon_m).
\end{aligned}$$

To finish the proof, we show $G_{seq} = G^*$. The consequence $G_{seq} = G^*$ follows from the chain rule of conditional mutual information, but we show an argument based on convex optimization that characterizes the jointly optimal policy as well. From

[5] This is the capacity of the mth BSC [73, 35].

Thm. IV.3,

$$
\begin{aligned}
G^* &= \sup_{A^{(1)},\ldots,A^{(M)}} \Big\{ H\left(\sum_{i_1:i_M=0}^{1} g_{i_1:i_M}(\cdot) \mathbb{P}_n\Big(\bigcap_{m=1}^{M} (A_n^{(m)})^{i_m} \Big) \right) \\
&\quad - \sum_{i_1:i_M=0}^{1} H\left(g_{i_1:i_M}(\cdot) \right) \mathbb{P}_n\Big(\bigcap_{m=1}^{M} (A_n^{(m)})^{i_m} \Big) \Big\} \\
&= \sup_{\mathbf{p}} \Big\{ H\left(\sum_{i_1:i_M=0}^{1} g_{i_1:i_M}(\cdot) p_{i_1,\ldots,i_M} \right) \\
&\quad - \sum_{i_1:i_M=0}^{1} H(g_{i_1:i_M}(\cdot)) p_{i_1,\ldots,i_M} : \mathbf{p} \succeq 0, 1^T\mathbf{p} = 1 \Big\} \\
&= \sup_{\mathbf{p}} \{ H(\mathbf{p}^T\mathbf{g}) - \mathbf{p}^T H(\mathbf{g}) : \mathbf{p} \succeq 0, 1^T\mathbf{p} = 1 \} \qquad (4.26) \\
&= G_{seq}
\end{aligned}
$$

where the last equality follows by the symmetry of the BSC. The supremum in the strictly concave problem (4.26) is achieved by the uniform distribution. This is justified by noting that the second term is independent of $\mathbf{p}$ since for $1^T\mathbf{p} = 1$:

$$
\begin{aligned}
\mathbf{p}^T H(\mathbf{g}) &= \sum_{i_1:i_M=0}^{1} H\left(\prod_{m=1}^{M} f_{i_m}^{(m)}(\cdot) \right) p_{i_1,\ldots,i_M} \\
&= \sum_{i_1:i_M=0}^{1} \sum_{m=1}^{M} H\left(f_{i_m}^{(m)}(\cdot) \right) p_{i_1,\ldots,i_M} \\
&= \sum_{m=1}^{M} h_B(\epsilon_m) \cdot \sum_{i_1=0}^{1} \cdots \sum_{i_M=0}^{1} p_{i_1,\ldots,i_M} \\
&= \sum_{m=1}^{M} h_B(\epsilon_m).
\end{aligned}
$$

Thus, the supremum of (4.26) can be restricted to the first term which is achieved by $p^*_{i_1,\ldots,i_M} = 2^{-M}$ since:

$$
\begin{aligned}
H\left(\sum_{i_1:i_M=0}^{1} g_{i_1:i_M}(\mathbf{y}) p^*_{i_1,\ldots,i_M} \right) &= H\left(2^{-M} \sum_{i_1=0}^{1} \cdots \sum_{i_M=0}^{1} \prod_{m=1}^{M} (1-\epsilon_m)^{I(y^{(m)}=i_m)} \epsilon_m^{1-I(y^{(m)}=i_m)} \right) \\
&= H(u(\cdot)) = \log_2(2^M) = M
\end{aligned}
$$

where $u(\cdot)$ is the uniform distribution over $\{0,1\}^M$. □

4.9.3 Proof of Lemma IV.6

Proof. From the definition of the expectation of a bounded random variable $E_n = (X^* - \hat{X}_n)^2$:

$$\begin{aligned}
\mathbb{E}[(X^* - \hat{X}_n)^2] &= \int_0^1 \mathbb{P}((X^* - \hat{X}_n)^2 > t)dt \\
&= \int_0^{\Delta^2} \mathbb{P}((X^* - \hat{X}_n)^2 > t)dt + \int_{\Delta^2}^1 \mathbb{P}((X^* - \hat{X}_n)^2 > t)dt \\
&\leq \Delta^2 + (1 - \Delta^2)\mathbb{P}(|X^* - \hat{X}_n| > \Delta).
\end{aligned}$$

□

4.9.4 Proof of Theorem IV.5

Proof. We note from the proof of Thm. IV.3 or Thm. IV.4, for any policy π, we have $\mathbb{E}^\pi[H(p_n)] \geq H(p_0) - nC$ [6]. Let K_n denote the conditional error covariance of the random vector $e_n = X^* - \mathbb{E}[X^*|\mathbf{Y}^n]$-i.e., $K_n = \mathrm{Cov}(e_n|\mathbf{Y}^n)$. From Thm. 17.2.3 in [35] and Jensen's inequality, we have:

$$\begin{aligned}
\mathbb{E}^\pi[H(p_n)] &\leq \mathbb{E}^\pi\left[\frac{1}{2}\log((2\pi e)^d \det(K_n))\right] \\
&\leq \frac{1}{2}\log((2\pi e)^d) + \frac{1}{2}\log(\det(\mathbb{E}^\pi[K_n])) \\
&= \frac{1}{2}\log((2\pi e)^d \det(\mathbb{E}^\pi[K_n]))
\end{aligned}$$

Rewriting this:

$$\frac{Ke^{-2nC}}{(2\pi e)^d} \leq \frac{e^{2\mathbb{E}^\pi[H(p_n)]}}{(2\pi e)^d} \leq \det(\mathbb{E}^\pi[K_n]) \leq \left(\frac{\mathbb{E}^\pi[\mathrm{tr}(K_n)]}{d}\right)^d$$

where we also used the inequality of arithmetic and geometric means in the last step. Using the fact that the conditional mean minimizes the mean-square error yields the final result. □

[6] For optimal policies π, this becomes an equality.

4.9.5 Proof of Theorem IV.7

Proof. Assume the pseudo-posterior after the mth player's response is $\mathbf{a}^{(M-m)}(j+1)$, with the notation $\mathbf{a}^{(0)}(j+1) = \mathbf{a}(j+1)$. Let k^* denote the index of the bin that contains X^*-i.e., $X^* \in I_{k^*}$. Define $M(j)^{(m)} = \frac{1}{a_{k^*}^{(M-m)}(j)} - 1$, with the notation $M^{(0)}(j) = M(j)$. Define the ratio $N(j+1) = \frac{M(j+1)}{M(j)}$. Let $\{\alpha_m\}_m$ denote the parameters associated with each player's pseudo-posterior update. Similarly as in the proof of Thm. 1 in [30]:

$$\begin{aligned}
\mathbb{P}(|X^* - X_n| > \Delta) &\leq \mathbb{P}(a_{k^*}(n) < 1/2) \\
&= \mathbb{P}(M(n) > 1) \leq \mathbb{E}[M(n)] \\
&\leq M(0) \left(\max_{0 \leq j \leq n-1} \max_{\mathbf{a}(j)} \mathbb{E}[N(j+1)|\mathbf{a}(j)] \right)^n .
\end{aligned}$$

Using the bounds in the proof of Thm. 1 in [30] and the tower property of conditional expectations repeatedly:

$$\begin{aligned}
&\mathbb{E}[N(j+1)|\mathbf{a}(j)] = \mathbb{E}\left[\frac{M^{(M-1)}(j+1)}{M^{(0)}(j)} \times \prod_{k=1}^{M-1} \frac{M^{(k-1)}(j+1)}{M^{(k)}(j+1)} \Big| \mathbf{a}^{(0)}(j)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{M^{(M-1)}(j+1)}{M^{(0)}(j)} \times \prod_{k=1}^{M-1} \frac{M^{(k-1)}(j+1)}{M^{(k)}(j+1)} \Big| \mathbf{a}^{(M-1)}(j+1), \ldots, \mathbf{a}^{(1)}(j+1), \mathbf{a}^{(0)}(j)\right] \Big| \mathbf{a}^{(0)}(j)\right] \\
&= \mathbb{E}\left[\frac{M^{(M-1)}(j+1)}{M^{(0)}(j)} \mathbb{E}\left[\prod_{k=1}^{M-1} \frac{M^{(k-1)}(j+1)}{M^{(k)}(j+1)} \Big| \mathbf{a}^{(M-1)}(j+1), \ldots, \mathbf{a}^{(1)}(j+1)\right] \Big| \mathbf{a}^{(0)}(j)\right] \\
&= \mathbb{E}\left[\frac{M^{(M-1)}(j+1)}{M^{(0)}(j)} \mathbb{E}\left[\prod_{k=2}^{M-1} \frac{M^{(k-1)}(j+1)}{M^{(k)}(j+1)} \mathbb{E}\left[\frac{M^{(0)}(j+1)}{M^{(1)}(j+1)} \Big| \mathbf{a}^{(1)}(j+1)\right]\right.\right. \\
&\qquad \left.\left. \Big| \mathbf{a}^{(M-1)}(j+1), \ldots, \mathbf{a}^{(2)}(j+1)\right] \Big| \mathbf{a}^{(0)}(j)\right] \\
&\leq \left(\frac{1-\epsilon_M}{2(1-\alpha_M)} + \frac{\epsilon_M}{2\alpha_M}\right) \mathbb{E}\left[\frac{M^{(M-1)}(j+1)}{M^{(0)}(j)} \mathbb{E}\left[\prod_{k=2}^{M-1} \frac{M^{(k-1)}(j+1)}{M^{(k)}(j+1)}\right.\right. \\
&\qquad \left.\left. \Big| \mathbf{a}^{(M-1)}(j+1), \ldots, \mathbf{a}^{(2)}(j+1)\right] \Big| \mathbf{a}^{(0)}(j)\right] \\
&\leq \ldots \\
&\leq \prod_{m=1}^{M} \left(\frac{1-\epsilon_m}{2(1-\alpha_m)} + \frac{\epsilon_m}{2\alpha_m}\right).
\end{aligned}$$

To optimize the bound, we choose $\alpha_i = \frac{\sqrt{\epsilon_i}}{\sqrt{\epsilon_i}+\sqrt{1-\epsilon_i}}, i = 1, 2$ to obtain:

$$\begin{aligned}
\mathbb{P}(|X^* - X_n| > \Delta) &\leq (\frac{1}{\Delta} - 1) \left(\prod_{m=1}^{M} \left(1 - \bar{C}(\epsilon_m)\right)\right)^n \\
&\leq (\frac{1}{\Delta} - 1) \exp\left(-n \sum_{m=1}^{M} \bar{C}(\epsilon_n)\right).
\end{aligned}$$

This concludes the first part. The second part follows by applying Lemma IV.6:

$$\mathbb{E}[(X^* - \hat{X}_n)^2] \leq \Delta^2 + \Delta^{-1} e^{-n\bar{C}}.$$

Optimizing the bound, we choose $\Delta = \Delta_n = 2^{-1/3} e^{-n\bar{C}/3}$, from which we conclude the second part. □

4.9.6 Proof of Corollary IV.8

Proof. Let $\Delta \in [0,1]$ be arbitrary. The Borel-Cantelli lemma implies that if $\sum_{n=1}^{\infty} \mathbb{P}(|\hat{X}_n - X^*| > \Delta) < \infty$, then $\mathbb{P}(|\hat{X}_n - X^*| > \Delta \text{ for infinitely many } n \geq 1) = 0$. This implies $\hat{X}_n \overset{a.s.}{\to} X^*$. From Theorem IV.7, we obtain:

$$\begin{aligned}\sum_{n=1}^{\infty} \mathbb{P}(|\hat{X}_n - X^*| > \Delta) &\leq \sum_{n=1}^{\infty} (\frac{1}{\Delta} - 1) \exp\left(-n\bar{C}\right) \\ &= (\frac{1}{\Delta} - 1) \frac{\exp\left(-\bar{C}\right)}{1 - \exp\left(-\bar{C}\right)} < \infty\end{aligned}$$

The claim follows by the argument presented above. □

4.9.7 Proof of Corollary IV.9

Proof. At each iteration n of the algorithm, M posterior updates are made. The density evolution from time n to time $n+1$ can be expressed as:

$$p_{n+1}(x) = p_n(x) \prod_{i=1}^{M} \frac{l_i(Y_{i,n+1}|x, A_{n,i})}{\mathcal{Z}_{i,n}(Y_{i,n+1})} \tag{4.27}$$

where $A_{n,i}$ denote the query region of the ith player and $l_i(Y_{i,n+1}|x, A_{n,i})$ is the ith player's observation density dependent on the query region. Let $p_{i,n}(x) = p_n(x) \prod_{i'=1}^{i-1} \frac{l_{i'}(Y_{i',n+1}|x, A_{n,i'})}{\mathcal{Z}_{i',n}(Y_{i',n+1})}, i = 1, \ldots, M+1$ denote the posterior density after the $(i-1)$th player update [7]. Note that the normalizing factor $\mathcal{Z}_{i,n}(Y_{i,n+1})$ is equal to $1/2$ irrespective of $Y_{i,n+1}$ since:

$$\begin{aligned}\mathcal{Z}_{i,n}(y) &= \int_{\mathcal{X}} l_i(y|x, A_{i,n}) p_{i,n}(x) dx \\ &= \int_{\mathcal{X}} \left(f_1^{(i)}(y) I(x \in A_{i,n}) + f_0^{(i)}(y) I(x \notin A_{i,n}) \right) p_{i,n}(x) dx \\ &= f_1^{(i)}(y) \mathbb{P}_{i,n}(A_{i,n}) + f_0^{(i)}(y) \mathbb{P}_{i,n}(A_{i,n}^c) \\ &= \frac{1}{2} \left(f_1^{(i)}(y) + f_0^{(i)}(y) \right) \\ &= \frac{1}{2}\end{aligned}$$

[7] Here, the initial condition is $p_{1,n}(x) = p_n(x)$ and the terminal condition is $p_{M+1,n}(x) = p_{n+1}(x)$.

where we used the continuous PBA bisection rule $\mathbb{P}_{i,n}(A_{i,n}) = 1/2$ and the channel symmetry. Taking the logarithm in (4.27) and using the memoryless nature of the channels:

$$\begin{aligned}\log p_{n+1}(X^*) &= \log p_n(X^*) + \sum_{i=1}^{M} \log(2l_i(Y_{i,n+1}|X^*, A_{i,n})) \\ &= \log p_n(X^*) + \sum_{i=1}^{M} \log(2\mathbb{P}_i(Y_{i,n+1}|Z_{i,n}))\end{aligned}$$

where $Z_{i,n} = I(X^* \in A_{i,n})$ is the input to the ith channel. Unwrapping this recursion and using the strong law of large numbers (LLN), we further obtain:

$$\begin{aligned}\frac{1}{n}\log p_n(X^*) &= \frac{1}{n}\log p_0(X^*) + \frac{1}{n}\sum_{k=0}^{n-1}\sum_{i=1}^{M} \log(2\mathbb{P}_i(Y_{i,k+1}|Z_{i,k})) \\ &\xrightarrow{a.s.} \mathbb{E}\left[\sum_{i=1}^{M} \log_2(2\mathbb{P}_i(Y_i|Z_i))\right]\end{aligned}$$

To finish the proof, note:

$$\begin{aligned}\mathbb{E}[\log(2\mathbb{P}_i(Y_i|Z_i))] &= \sum_{z=0}^{1}\sum_{y=0}^{1} \mathbb{P}_i(y,z)\log_2\left(2\mathbb{P}_i(Y_i|Z_i)\right) \\ &= \sum_{z} \mathbb{P}_i(z) \sum_{y} \mathbb{P}_i(y|z)\log(2\mathbb{P}_i(y|z)) \\ &= \sum_{z} \mathbb{P}_i(z)\left((1-\epsilon_i)\log(2(1-\epsilon_i)) + \epsilon_i \log(2\epsilon_i)\right) \\ &= 1 - h_B(\epsilon_i) = C(\epsilon_i)\end{aligned}$$

where $h_B(\epsilon_i) = (1-\epsilon_i)\log_2(\frac{1}{1-\epsilon_i}) + \epsilon_i \log_2(\frac{1}{\epsilon_i})$ is the binary entropy function. □

4.9.8 Proof of Theorem IV.11

Proof. 1) Optimality conditions

The solution of (4.1) yields the Bellman recursion:

$$V_n(p_n) = \inf_{\mathbf{A}} \mathbb{E}\left[V_{n+1}(p_{n+1})|\mathbf{A}_n = \mathbf{A}, \mathcal{F}_n\right]$$

Using a similar argument as in Thm. 2 in [73], the optimal solution at time n is given by maximizing the entropy loss at time n:

$$G_n^* = \sup_{\mathbf{A}} I((X^*, \boldsymbol{\epsilon}^*); \mathbf{Y}_{n+1} | \mathbf{A}_n = \mathbf{A}, \mathcal{F}_n) = \sup_{\mathbf{A}} \left\{ H(p_n) - \mathbb{E}\left[H(p_{n+1}) | \mathbf{A}_n = \mathbf{A}, \mathcal{F}_n \right] \right\}$$

and the value function is given by $V_n(p_n) = H(p_n) - \sum_{k=n}^{N-1} G_k^*$ for $n < N$ and $V_N(p_N) = H(p_N)$. We can expand the mutual information:

$$I((X^*, \boldsymbol{\epsilon}^*); \mathbf{Y}_{n+1} | \mathbf{A}_n, \mathcal{F}_n) = H(\mathbf{Y}_{n+1} | \mathbf{A}_n, \mathcal{F}_n) - \mathbb{E}\left[H(\mathbf{Y}_{n+1}) | X^*, \boldsymbol{\epsilon}^*, \mathbf{A}_n, \mathcal{F}_n \right] \tag{4.28}$$

The conditional probability of $\mathbf{Y}_{n+1}$ given the query $\mathbf{A}_n = \mathbf{A}$ can be written as:

$$\begin{aligned}
&\mathbb{P}(\mathbf{Y}_{n+1} | \mathbf{A}_n = \mathbf{A}, \mathcal{F}_n) = \mathbb{E}[\mathbb{P}(\mathbf{Y}_{n+1} | \mathbf{A}_n = \mathbf{A}, X^*, \boldsymbol{\epsilon}^*, \mathcal{F}_n)] \\
&= \int_{\boldsymbol{\epsilon}=0}^{1/2} \int_{x \in \mathcal{X}} \mathbb{P}(\mathbf{Y}_{n+1} | \mathbf{A}_n = \mathbf{A}, X^* = x, \boldsymbol{\epsilon}^* = \boldsymbol{\epsilon}) p_n(x, \boldsymbol{\epsilon}) dx d\boldsymbol{\epsilon} \\
&= \int_{\boldsymbol{\epsilon}=0}^{1/2} \int_{x \in \mathcal{X}} \left(\prod_{m=1}^{M} f_1(Y_{n+1}^{(m)} | \epsilon_m) I(x \in A^{(m)}) + f_0(Y_{n+1}^{(m)} | \epsilon_m) I(x \notin A^{(m)}) \right) p_n(x, \boldsymbol{\epsilon}) dx d\boldsymbol{\epsilon} \\
&= \int_{\boldsymbol{\epsilon}=0}^{1/2} \sum_{i_1 : i_M = 0}^{1} g_{i_1 : i_M}(\mathbf{y} | \boldsymbol{\epsilon}) \left\{ \int_{x \in \mathcal{X}} I\left(\bigcap_m (A^{(m)})^{i_m} \right) p_n(x, \boldsymbol{\epsilon}) dx \right\} d\boldsymbol{\epsilon} \\
&= \sum_{i_1 : i_M = 0}^{1} \int_{\boldsymbol{\epsilon}=0}^{1/2} g_{i_1 : i_M}(\mathbf{y} | \boldsymbol{\epsilon}) \mathbb{P}_n \left(\bigcap_m (A^{(m)})^{i_m}, \boldsymbol{\epsilon} \right) d\boldsymbol{\epsilon}
\end{aligned}$$

where $p_n(x, \boldsymbol{\epsilon}) = p_n(x, \epsilon_1, \ldots, \epsilon_M)$. This gives the first term in (4.19). To obtain the second term, notice:

$$\begin{aligned}
&\mathbb{E}[H(\mathbf{Y}_{n+1}) | X^*, \boldsymbol{\epsilon}^*, \mathbf{A}_n = \mathbf{A}, \mathcal{F}_n] \\
&= \int_{\boldsymbol{\epsilon}} \int_{x \in \mathcal{X}} p_n(x, \boldsymbol{\epsilon}) H(\mathbf{Y}_{n+1} | X^* = x, \boldsymbol{\epsilon}^* = \boldsymbol{\epsilon}, \mathbf{A}_n = \mathbf{A}, \mathcal{F}_n) dx d\boldsymbol{\epsilon} \\
&= \int_{\boldsymbol{\epsilon}} \left\{ \sum_{i_1 : i_M = 0}^{1} \int_{x \in \bigcap_m (A^{(m)})^{i_m}} p_n(x, \boldsymbol{\epsilon}) H(g_{i_1 : i_M}(\cdot | \boldsymbol{\epsilon})) dx \right\} d\boldsymbol{\epsilon} \\
&= \sum_{i_1 : i_M = 0}^{1} \int_{\boldsymbol{\epsilon}=0}^{1/2} H(g_{i_1 : i_M}(\cdot | \boldsymbol{\epsilon})) \mathbb{P}_n \left(\bigcap_m (A^{(m)})^{i_m}, \boldsymbol{\epsilon} \right) d\boldsymbol{\epsilon}
\end{aligned}$$

The proof is complete.

2) Bounds on Maximum entropy loss

First, we prove the upper bound. Note that the second term in (4.19) is independent of the queries, so the supremum can be restricted to only the first term without loss of generality. This is justified by using the additivity of the entropy of a product density:

$$\begin{aligned} H(g_{i_1:i_M}(\cdot|\boldsymbol{\epsilon})) &= H\left(\prod_{m=1}^{M} f_{i_m}^{(m)}(\cdot|\epsilon_m)\right) \\ &= \sum_{m=1}^{M} H(f_{i_m}^{(m)}(\cdot|\epsilon_m)) = \sum_{m=1}^{M} h_b(\epsilon_m) \end{aligned}$$

From part 1), the maximum entropy loss can be bounded from above as:

$$\begin{aligned} G_n^* &= \sup_{A^{(1)},\ldots,A^{(M)}} H\left(\sum_{i_1:i_M=0}^{1}\int_{\boldsymbol{\epsilon}=0}^{1/2} g_{i_1:i_M}(\cdot|\boldsymbol{\epsilon})\mathbb{P}_n\left(\bigcap_m (A^{(m)})^{i_m},\boldsymbol{\epsilon}\right)d\boldsymbol{\epsilon}\right) \\ &\quad - \sum_{i_1:i_M=0}^{1}\int_{\boldsymbol{\epsilon}=0}^{1/2} H(g_{i_1:i_M}(\cdot|\boldsymbol{\epsilon}))\mathbb{P}_n\left(\bigcap_m (A^{(m)})^{i_m},\boldsymbol{\epsilon}\right)d\boldsymbol{\epsilon}\Big\} \\ &\leq \log_2(\mathrm{card}(\mathcal{Y})) - \int_{\boldsymbol{\epsilon}=0}^{1/2}\left\{\sum_m h_B(\epsilon_m)\right\}\left\{\sum_{i_1:i_M=0}^{1}\mathbb{P}_n\left(\bigcap_m (A^{(m)})^{i_m},\boldsymbol{\epsilon}\right)\right\}d\boldsymbol{\epsilon} \qquad (4.29) \\ &= M - \sum_m\left\{\int_{\boldsymbol{\epsilon}_m=0}^{1/2} h_b(\epsilon_m)p_n(\epsilon_m)d\boldsymbol{\epsilon}\right\} \\ &= \sum_m (1-\mathbb{E}[h_b(\epsilon_m)|\mathcal{F}_n]) \\ &= \mathbb{E}\left[\sum_m C(\epsilon_m)\Big|\mathcal{F}_n\right] \end{aligned}$$

where we used the fact that the capacity of a BSC is $C(\epsilon_m) = 1 - h_b(\epsilon_m)$. In (4.29), we also used the fact that the uniform distribution maximizes the entropy (see Ch.2 in [35]).

Second, we prove the lower bound. By the concavity of $H(\cdot)$, we obtain:

$$
\begin{aligned}
G_n^* = & \sup_{A^{(1)},\ldots,A^{(M)}} \Big\{ H\left(\sum_{i_1:i_M=0}^{1} \int_{\boldsymbol{\epsilon}=0}^{1/2} g_{i_1:i_M}(\cdot|\boldsymbol{\epsilon}) \mathbb{P}_n \left(\bigcap_m (A^{(m)})^{i_m}, \boldsymbol{\epsilon} \right) d\boldsymbol{\epsilon} \right) \\
& - \sum_{i_1:i_M=0}^{1} \int_{\boldsymbol{\epsilon}=0}^{1/2} H(g_{i_1:i_M}(\cdot|\boldsymbol{\epsilon})) \mathbb{P}_n \left(\bigcap_m (A^{(m)})^{i_m}, \boldsymbol{\epsilon} \right) d\boldsymbol{\epsilon} \Big\} \\
\geq & \sup_{A^{(1)},\ldots,A^{(M)}} \Big\{ \int_{\boldsymbol{\epsilon}=0}^{1/2} H\left(\sum_{i_1:i_M=0}^{1} g_{i_1:i_M}(\cdot|\boldsymbol{\epsilon}) \mathbb{P}_n \left(\bigcap_m (A^{(m)})^{i_m} \Big| \boldsymbol{\epsilon} \right) \right) p_n(\boldsymbol{\epsilon}) d\boldsymbol{\epsilon} \\
& - \int_{\boldsymbol{\epsilon}=0}^{1/2} \sum_{i_1:i_M=0}^{1} H(g_{i_1:i_M}(\cdot|\boldsymbol{\epsilon})) \mathbb{P}_n \left(\bigcap_m (A^{(m)})^{i_m} \Big| \boldsymbol{\epsilon} \right) p_n(\boldsymbol{\epsilon}) d\boldsymbol{\epsilon} \Big\} \\
= & \sup_{A^{(1)},\ldots,A^{(M)}} \mathbb{E}\Big[H\left(\sum_{i_1:i_M=0}^{1} g_{i_1:i_M}(\cdot|\boldsymbol{\epsilon}) \mathbb{P}_n \left(\bigcap_m (A^{(m)})^{i_m} \Big| \boldsymbol{\epsilon} \right) \right) \\
& - \sum_{i_1:i_M=0}^{1} H(g_{i_1:i_M}(\cdot|\boldsymbol{\epsilon})) \mathbb{P}_n \left(\bigcap_m (A^{(m)})^{i_m} \Big| \boldsymbol{\epsilon} \right) \Big| \mathcal{F}_n \Big] \\
= & \sup_{\mathbf{p}:\mathbf{p}\geq 0, 1^T\mathbf{p}=1} \mathbb{E}\Big[H\left(\sum_{i_1:i_M=0}^{1} g_{i_1:i_M}(\cdot|\boldsymbol{\epsilon}) p_{i_1,\ldots,i_M} \right) \qquad (4.30) \\
& - \sum_{i_1:i_M=0}^{1} H(g_{i_1:i_M}(\cdot|\boldsymbol{\epsilon})) \mathbb{P}_n \left(\bigcap_m (A^{(m)})^{i_m} \Big| \boldsymbol{\epsilon} \right) \Big| \mathcal{F}_n \Big] \\
= & \, \mathbb{E}\Big[\sup_{\mathbf{p}:\mathbf{p}\geq 0, 1^T\mathbf{p}=1} H\left(\sum_{i_1:i_M=0}^{1} g_{i_1:i_M}(\cdot|\boldsymbol{\epsilon}) p_{i_1,\ldots,i_M} \right) \\
& - \sum_{i_1:i_M=0}^{1} H(g_{i_1:i_M}(\cdot|\boldsymbol{\epsilon})) p_{i_1,\ldots,i_M} \Big| \mathcal{F}_n \Big] \\
= & \, \mathbb{E}\left[\sum_{m=1}^{M} C(\epsilon_m) \Big| \mathcal{F}_n \right] \qquad (4.31)
\end{aligned}
$$

where we used the consistent reparameterization $\mathbb{P}_n\left(\bigcap_m (A^{(m)})^{i_m} \Big| \boldsymbol{\epsilon}\right) = p_{i_1,\ldots,i_M}$ in (4.30) and Thm. IV.4 in (4.31). □

4.9.9 Proof of Theorem IV.12

Proof. After querying all M players in sequence, the entropy loss is:

$$
\begin{aligned}
G^*_{seq,n} &= \sup_{\{A_{n_t}\}_{t=0}^{M-1}} \mathbb{E}[H(p_n) - H(p_{n+1})|\mathcal{G}_n] \\
&= \sup_{\{A_{n_t}\}_{t=0}^{M-1}} \mathbb{E}\left[\sum_{t=0}^{M-1} H(p_{n_t}) - H(p_{n_{t+1}})\middle|\mathcal{G}_n\right] \qquad (4.32) \\
&= \sup_{\{A_{n_t}\}_{t=0}^{M-1}} \sum_{t=0}^{M-1} \mathbb{E}\left[\mathbb{E}\left[H(p_{n_t}) - H(p_{n_{t+1}})\middle|A_{n_t}, \mathcal{G}_{n_t}\right]\middle|\mathcal{G}_n\right] \qquad (4.33) \\
&= \sup_{\{A_{n_t}\}_{t=0}^{M-1}} \mathbb{E}\left[\sum_{t=0}^{M-1} I((X^*, \boldsymbol{\epsilon}^*); Y_{n_{t+1}}|A_{n_t}, \mathcal{G}_{n_t})\middle|\mathcal{G}_n\right] \qquad (4.34) \\
&= \mathbb{E}\left[\sum_{t=0}^{M-1} \sup_{A_{n_t}} I((X^*, \boldsymbol{\epsilon}^*); Y_{n_{t+1}}|A_{n_t}, \mathcal{G}_{n_t})\middle|\mathcal{G}_n\right] \\
&= \mathbb{E}\left[\sum_{t=0}^{M-1} C(\epsilon_{t+1})\middle|\mathcal{G}_n\right] = \mathbb{E}\left[\sum_{m=1}^{M} C(\epsilon_m)\middle|\mathcal{G}_n\right]
\end{aligned}
$$

where we used a telescoping sum in (4.32) and the tower property of expectation with $\mathcal{G}_{n_t} \supseteq \mathcal{G}_n$ in (4.33). In (4.34), we used the optimality condition of maximum entropy loss by applying Thm. IV.11 with M=1 for each sub-instant n_t with $m = t + 1$:

$$\sup_{A_{n_t}} \left\{H(p_{n_t}) - \mathbb{E}[H(p_{n_{t+1}})|A_{n_t}, \mathcal{G}_{n_t}]\right\} = \sup_{A_{n_t}} I((X^*, \boldsymbol{\epsilon}^*); Y_{n_{t+1}}|A_{n_t}, \mathcal{G}_{n_t}) = \mathbb{E}[C(\epsilon_{t+1})|\mathcal{G}_{n_t}]$$

The second part follows from Thm. IV.11 part 2). □

4.9.10 Proof of Theorem IV.14

Proof. The solution of (4.1) yields the Bellman recursion:

$$V_n(p_n) = \inf_{u,A} \mathbb{E}\left[V_{n+1}(p_{n+1})|u_n = u, A_n = A, \mathcal{F}_n\right]$$

Using a similar argument as in Thm. 2 in [73], the optimal solution at time n is given by maximizing the entropy loss at time n:

$$G_n = \max_u \sup_A I((X^*, \boldsymbol{\epsilon}^*); Y^{(u)}_{n+1}|u_n = u, A^{(u)}_n = A, \mathcal{F}_n) = H(p_n) - \mathbb{E}\left[H(p_{n+1})|u_n = u, A^{(u)}_n = A, \mathcal{F}_n\right]$$

and the value function is given by $V_n(p_n) = H(p_n) - \sum_{k=n}^{N-1} G_k$ for $n < N$ and $V_N(p_N) = H(p_N)$. We can expand the mutual information:

(4.35)

$$I((X^*, \boldsymbol{\epsilon}^*); Y_{n+1}^{(u)} | u_n, A_n^{(u)}, \mathcal{F}_n) = H(Y_{n+1}^{(u)} | u_n, A_n^{(u)}, \mathcal{F}_n) - \mathbb{E}\left[H(Y_{n+1}^{(u)}) | X^*, \boldsymbol{\epsilon}^*, u_n, A_n^{(u)}, \mathcal{F}_n\right]$$

The conditional probability of $Y_{n+1}^{(u)}$ given the selection $u_n = u$ and the query $A_n^{(u)} = A$:

$$\begin{aligned}
&\mathbb{P}(Y_{n+1}^{(u)} | u_n = u, A_n^{(u)} = A, \mathcal{F}_n) = \mathbb{E}[\mathbb{P}(Y_{n+1}^{(u)} | u_n = u, A_n^{(u)} = A, X^*, \boldsymbol{\epsilon}^*, \mathcal{F}_n)] \\
&\quad = \int_{\boldsymbol{\epsilon}} \int_{x \in \mathcal{X}} \Big(f_1(Y_{n+1}^{(u)} | \epsilon_u) I(x \in A) + f_0(Y_{n+1}^{(u)} | \epsilon_u) I(x \notin A) \Big) p_n(x, \boldsymbol{\epsilon}) dx d\boldsymbol{\epsilon} \\
&\quad = \int_{\epsilon_u = 0}^{1/2} \int_{x \in \mathcal{X}} \Big(f_1(Y_{n+1}^{(u)} | \epsilon_u) I(x \in A) + f_0(Y_{n+1}^{(u)} | \epsilon_u) I(x \notin A) \Big) p_n^{(u)}(x, \epsilon_u) dx d\epsilon_u \\
&\quad = \int_{\epsilon_u = 0}^{1/2} f_1(Y_{n+1}^{(u)} | \epsilon_u) \mathbb{P}_n^{(u)}(A, \epsilon_u) + f_0(Y_{n+1}^{(u)} | \epsilon_u) \mathbb{P}_n^{(u)}(A^c, \epsilon_u) d\epsilon_u
\end{aligned}$$

where $p_n^{(u)}(x, \epsilon_u) = \int_{\{\epsilon_m \in [0,1/2): m \neq u\}} p_n(x, \boldsymbol{\epsilon}) d\{\epsilon_m : m \neq u\}$ denotes the uth sub-marginal. This gives the first term in (4.22). To obtain the second term, notice:

$$\begin{aligned}
&\mathbb{E}[H(Y_{n+1}^{(u)}) | X^*, \boldsymbol{\epsilon}^*, u_n = u, A_n^{(u)} = A, \mathcal{F}_n] \\
&\quad = \int_{\boldsymbol{\epsilon}} \int_{x \in \mathcal{X}} p_n(x, \boldsymbol{\epsilon}) H(Y_{n+1}^{(u)} | X^* = x, \boldsymbol{\epsilon}^* = \boldsymbol{\epsilon}, u_n = u, A_n^u = A, \mathcal{F}_n) dx d\boldsymbol{\epsilon} \\
&\quad = \int_{\boldsymbol{\epsilon}} \left\{ \int_{x \in A} p_n(x, \boldsymbol{\epsilon}) H(f_1(Y_{n+1}^{(u)} | \epsilon_u)) dx + \int_{x \notin A} p_n(x, \boldsymbol{\epsilon}) H(f_0(Y_{n+1}^{(u)} | \epsilon_u)) dx \right\} d\boldsymbol{\epsilon} \\
&\quad = \int_{\epsilon_u = 0}^{1/2} H(f_1(\cdot | \epsilon_u)) \mathbb{P}_n^{(u)}(A, \epsilon_u) + H(f_0(\cdot | \epsilon_u)) \mathbb{P}_n^{(u)}(A^c, \epsilon_u) d\epsilon_u
\end{aligned}$$

The proof of the first part is complete. The second part follows from part (2) of Theorem IV.11. □

4.9.11 Proof of Corollary IV.15

Proof. From Thm. IV.14, we have the optimality condition shown in (4.22). Under Assumption IV.2, we have $H(f_0(\cdot | \epsilon_u) = H(f_1(\cdot | \epsilon_u)) = h_B(\epsilon_u)$. Using this in the

second term in the supremum of (4.22):

$$\begin{aligned}
&\int_{\epsilon_u=0}^{1/2} H\left(f_1(\cdot|\epsilon_u)\right)\mathbb{P}_n^{(u)}(A,\epsilon_u) + H\left(f_0(\cdot|\epsilon_u)\right)\mathbb{P}_n^{(u)}(A^c,\epsilon_u)d\epsilon_u \\
&\qquad = \int_{\epsilon_u=0}^{1/2} h_B(\epsilon_u)\left(\mathbb{P}_n^{(u)}(A,\epsilon_u) + \mathbb{P}_n^{(u)}(A^c,\epsilon_u)\right)d\epsilon_u \\
&\qquad = \int_{\epsilon_u=0}^{1/2} h_B(\epsilon_u)p_n^{(u)}(\epsilon_u)d\epsilon_u = c_n^{(u)} \qquad (4.36)
\end{aligned}$$

Thus, we conclude that the second term in (4.22) is independent of the query region A, but still depends on the sensor u.

Rewriting the first term in the supremum of (4.22), we have for $A = [0, x]$:

$$\begin{aligned}
&H\left(\int_{\epsilon_u=0}^{1/2} f_1(\cdot|\epsilon_u)\mathbb{P}_n^{(u)}(A,\epsilon_u) + f_0(\cdot|\epsilon_u)\mathbb{P}_n^{(u)}(A^c,\epsilon_u)d\epsilon_u\right) \\
&= H\left(\int_{\epsilon_u=0}^{1/2} f_1(\cdot|\epsilon_u)\left\{\int_0^x p_n^{(u)}(t,\epsilon_u)dt\right\} + f_0(\cdot|\epsilon_u)\left\{\int_x^1 p_n^{(u)}(t,\epsilon_u)dt\right\}d\epsilon_u\right) \\
&= H\left(\int_0^x \left\{\int_{\epsilon_u=0}^{1/2} f_1(\cdot|\epsilon_u)p_n^{(u)}(t,\epsilon_u)d\epsilon_u\right\}dt + \int_x^1 \left\{\int_{\epsilon_u=0}^{1/2} f_0(\cdot|\epsilon_u)p_n^{(u)}(t,\epsilon_u)d\epsilon_u\right\}dt\right) \\
&\qquad (4.37) \\
&= h_B(g_{1,n}^{(u)}(x))
\end{aligned}$$

where $g_{1,n}^{(u)}(x)$ is defined in the statement of the theorem. □

CHAPTER V

Decentralized Collaborative 20 Questions

We consider the problem of decentralized 20 questions with noise for multiple players/agents under the minimum entropy criterion [73] in the setting of stochastic search, with application to target localization. We propose decentralized extensions of the active query-based search strategy that combines elements from the 20 questions approach of [118] and the social learning algorithm of [70]. Although agents do not have knowledge of their neighbors' statistics, using martingale theory, we prove asymptotic convergence (to the true target location) of the semi-Bayesian estimation strategy. This framework provides a flexible and tractable mathematical model for active decentralized target estimation systems. We illustrate the effectiveness and robustness of the proposed decentralized collaborative 20 questions model for several different network topologies.

5.1 Introduction

Consider a set of agents in a graph that try to localize a target collectively. In this chapter, we address the question: What is the value of collaboration when there is no central authority? In the decentralized framework, there is no fusion center that can perform centralized inference to come up with a sequential Bayesian estimate of the target location, e.g., as studied in [118]. A simple model for such a decentralized

estimation problem is that agents are connected according to some network topology and that they can perform local updating and information sharing. In [118], a framework was proposed in which each agent plays a cooperative 20 questions game: each agent is repeatedly queried about a target state and communicates its binary response to a centralized controller that determines the next set of queries. While the centralized controller requires global knowledge of the agents' error probabilities, this chapter proposes an extension of [118] to the decentralized framework where agents share their information with neighbors and only need to know their own error probability.

There exist many methods for decentralized information sharing in multi-agent systems that include consensus, gossip algorithms and distributed averaging. In each of these approaches, messages are distributed around the network through local processing and local communication. Consensus has broad applications including distributed optimization [121, 122], load-balancing [37], and distributed detection [106]. For example, the early seminal work of Tsitsiklis [121] studied averaging in the context of distributed estimation and detection.

Gossip algorithms have gained interest lately primarily due to their robustness and flexibility, and are directly related to consensus. The randomized gossip formulation proposed by Boyd et al. [17] adopted a randomized gossip model. It was shown that the convergence rate is controlled by the second largest eigenvalue of a doubly stochastic matrix defining the algorithm, making evident a natural relation between mixing times of random walks on the graph defined by a matrix of transition probabilities and averaging time of a gossip algorithm. However, the slow convergence of randomized gossip on random graphs sparked further research, including geographic gossip [46], where nodes pair up with geographically distant nodes and exhange infor-

maiton via multihop routing methods, yielding faster convergence. Further refining of these methods included randomized path averaging [8], where routing nodes contributed their own estimates along the way, requiring only a number of transmissions on ther order of the number of nodes in the network.

The survey paper by Dimakis et al. [45] reviews gossip algorithms for sensor networks in the context of estimation, source localization and compression. A large body of literature exists on gossip algorithms. In [3], randomized gossip broadcast algorithms for consensus were proposed and conditions for reaching consensus towards the average value of the initial node measurements were presented. The mean-square error of the randomized averaging procedure was also studied and shown to decay monotonically to a steady-state value. In [77], gossip for linear parameter estimation was studied and it was shown that, under appropriate conditions on the network structure and observation models, the distributed estimator achieves the same performance as the best centralized linear estimator (in terms of asymptotic variance). In [91], consensus aspects are studied specifically for the wireless medium and a new consensus algorithm called hierarchical averaging is proposed to improve tradeoffs between resource consumption and quantization error. Our work differs from these approaches since our observations obey noisy query-response models where the queries are functions of agents' local information and successive queries are determined by a feedback control policy.

Motivated by the approach of Jedynak et al. [73], which was restricted to the single agent, centralized collaborative multi-agent estimation of a target state was studied in [118] in the context of a noisy collaborative 20 questions game. In this framework a controller sequentially selects a set of questions about the target state and uses the noisy responses of the sensors to formulate the next set of questions. The

query response model for each agent are different to account for heterogeneity, e.g., a mixture of human and cyber agents. Under certain assumptions on the observation processes, it was shown that the optimal, entropy-minimizing joint query policy is equivalent to a sequential query policy.

In another body of work focused on social learning by Jadbabaie et al. [70], a dynamic model of opinion formation is studied. It is shown that when agents use a simple updating rule that linearly combines their personal belief with the neighbors' beliefs, as long as the agents' private signal are incorporated in a Bayesian manner, repeated interactions lead to successful learning of the true state of the world, which is assumed to be discrete-valued. The non-Bayesian learning model used in this work dates back to the models for opinion formation of DeGroot [43], under which each individual agent initially receives one signal about the state of the world and then shares its belief of the state with its neighbors. The key result is specification of conditions that guarantee asymptotic agreement among agents in connected components of the social network.

In this work, we study an alternative non-Bayesian estimation framework, as contrasted to the Bayesian framework proposed in [118], that consists of an updating stage and local belief sharing as proposed in [70]. Our work differs from the work of Jadbabaie et al [70] in several important respects:

- We consider continuous-valued target space as contrasted with the discrete case studied in [70].
- We consider controlled observations that violates the independent identically distributed assumptions in [70].

Our work also differs from the works on 20 questions/active stochastic search of Je-

dynak et al [73], Castro & Nowak [30], Waeber et al [126], and Tsiligkaridis et al [118] because we consider intermediate local belief sharing between agents after each local bisection and update. In addition, our work differs since each agent incorporates the beliefs of its neighbors in a way that is agnostic of its neighbors' error probabilities. We finally remark that convergence of the proposed algorithm is non-trivial since the entropy of the belief for each agent in the network is no longer guaranteed to be monotonically decreasing as a function of iteration.

The main convergence result is built on lemmas and Fig. 5.1 serves as a guide for the flow of the analysis presented in this paper.

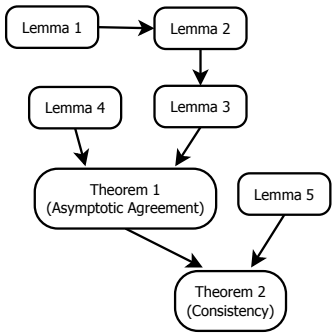


Figure 5.1: The flow of the convergence analysis.

5.1.1 Outline

The outline of this paper is as follows. Section 5.2 introduces the notation. Section 5.3 briefly reviews some related prior work. Section 5.4 introduces the decentralized estimation algorithm and its convergence properties are studied in Section 5.5. Several simulations are presented in Section 5.6 followed by our conclusions in Section 5.7.

5.2 Notation

We define X^* the true target location and its domain as the unit hypercube $\mathcal{X} = [0,1]^d$. Let $\mathcal{B}(\mathcal{X})$ be the set of all Borel-measurable subsets $B \subseteq \mathcal{X}$. Let $\mathcal{N} = \{1, \ldots, M\}$ denote the agent set of the network. The agents of the network are indexed by a vertex set V and the directed edges joining agents are captured by E. The directed graph $G = (\mathcal{N}, E)$ captures the possible interactions between agents. Define the neighborhood of agent i as $\mathcal{N}_i = \{j \in \mathcal{N} : (j,i) \in E\}$.

Define the belief of the ith agent at time t on $\mathcal{X}$ as the density $p_{i,t}(x)$. Define the $M \times 1$ vector $p_t(x) = [p_{1,t}(x), \ldots, p_{M,t}(x)]^T$ for each $x \in \mathcal{X}$. For any $B \in \mathcal{B}(\mathcal{X})$, define $\mathbb{P}_t(B)$ as the vector with ith element equal to $\int_B p_{i,t}(x)dx$. The interaction matrix $A = \{a_{i,j}\}$ (as in [70]) is defined to be any matrix A consisting of nonnegative entries where each row sums to 1. We define the query point/target estimate of the ith agent as $\hat{X}_{i,t}$. In the one-dimensional case $d = 1$, the query point is the right boundary of the region $A_{i,t} = [0, \hat{X}_{i,t}]$. Let $F_{i,t}(a) = \mathbb{P}_{i,t}([0,a]) = \int_0^a p_{i,t}(x)dx$ denote the CDF operator associated with the density $p_{i,t}(\cdot)$.

We assume that each agent i constructs a query at time t of the form "does X^* lie in the region $A_{i,t} \subset \mathcal{X}$?". We denote this query with the binary variable $Z_{i,t} = I(X^* \in A_{i,t})$ to which each agent i responds with a binary response $Y_{i,t+1}$, which is correct with probability $1 - \epsilon_i$, and by assumption $\epsilon_i < 1/2$. This model for the error channel is equivalent a binary symmetric channel (BSC) with crossover probability ϵ_i. The query region $A_{i,t}$ at time t depends on the accumulated information up to time t at agent i. Define a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω is the space containing sequences of realizations of the observations $\{y_{i,t} : 1 \leq i \leq M\}_{t \geq 0}$, and $\mathbb{P}(\cdot)$ is the probability measure associated with the sample paths in Ω. The expectation operator

$\mathbb{E}[\cdot]$ is taken with respect to the probability measure $\mathbb{P}(\cdot)$. Define the nested sequence of σ-algebras $\mathcal{F}_t$, $\mathcal{F}_{t-1} \subset \mathcal{F}_t$, for all $t \geq 0$, generates by the sequence of queries and responses. The queries $\{A_{i,t} : 1 \leq i \leq M\}_{t \geq 0}$ are measurable with respect to this filtration.

5.3 Prior Work

5.3.1 20 Questions & Stochastic search

The paper by Jedynak et al. [73] formulates the single player 20 questions problem as follows. A controller queries a noisy oracle about whether or not a target X^* lies in a set $A_n \subset \mathbb{R}^d$. Starting with a prior distribution on the target's location $p_0(\cdot)$, the objective in [73] is to minimize the expected entropy of the posterior distribution:

$$\inf_{\pi} \mathbb{E}^{\pi} \left[H(p_N) \right] \tag{5.1}$$

where $\pi = (\pi_0, \pi_1, \dots)$ denotes the controller's query policy and the entropy is the standard differential entropy [35]:

$$H(p) = -\int_{\mathcal{X}} p(x) \log p(x) dx.$$

The posterior median of p_N is used to estimate the target location after N questions. Jedynak [73] shows the bisection policy is optimal under the minimum entropy criterion. To be concrete, in Thm. 2 of [73], optimal policies are characterized by:

$$\mathbb{P}_n(A_n) := \int_{A_n} p_n(x) dx = u^* \in \arg\max_{u \in [0,1]} \phi(u) \tag{5.2}$$

where

$$\phi(u) = H(f_1 u + (1-u) f_0) - u H(f_1) - (1-u) H(f_0)$$

is nonnegative. The densities f_0 and f_1 correspond to the noisy channel [1]:

$$\mathbb{P}(Y_{n+1} = y | Z_n = z) = \begin{cases} f_1(y), & z = 1 \\ f_0(y), & z = 0 \end{cases}$$

where $Z_n = I(X^* \in A_n) \in \{0, 1\}$ is the channel input. The noisy channel models the conditional probability of the response to each question being correct. For the special case of a binary symmetric channel (BSC), $u^* = 1/2$ and the probabilistic bisection policy [73, 30] becomes an optimal policy.

In [118], optimality conditions are derived for optimal query strategies in the collaborative multiplayer case where observations are communicated to a fusion center (or centralized controller) and were shown to generalize the probabilistic bisection policy. Two policies were studied; a sequential bisection policy for which each player responds to a single question about the location of the target, and a joint policy where all players are asked questions simultaneously. It was proven that the maximum entropy reduction for the sequential bisection scheme is the same as that of the jointly optimal scheme, and is given by the sum of the capacities of all the players' channels. Thus, the centralized controller is equivalent to a cascade of low-complexity controllers. Despite the fact that the optimal sequential policy has access to a more refined filtration, it achieves the same average performance as the optimal joint policy. This equivalence was also extended to the setting where the error channels associated with the players are unknown.

5.3.2 Non-Bayesian Social Learning

In the work by Jadbabaie et al [70], it is assumed that Θ denotes a finite set of possible states of the world and the objective is to study conditions for asymptotic agreement on the true state of the world. A set $\mathcal{N} = \{1, \ldots, M\}$ of agents interacting

[1] The function $I(A)$ is the indicator function throughout the paper-i.e., $I(A) = 1$ if A is true and zero otherwise.

over a social network (directed graph) $G = (\mathcal{N}, E)$ is considered, where E encodes the edges between agents. An edge connecting agent i and agent j is denotes as the ordered pair $(i, j) \in E$, denoting that agent j has access to the belief of agent i. The interactions are captured by an interaction matrix A, where $a_{i,j}$ denotes the strength associated the communication of agent j's belief to agent i.

The beliefs of agent i at time $t \geq 0$, defined on Θ, is denoted by $p_{i,t}(\theta)$. Conditioned on the state of the world θ, at each time $t \geq 1$, an observation set $\mathbf{y}_t = (y_{1,t}, \ldots, y_{M,t})$ is generated by the likelihood function $l(\cdot|\theta)$. The signal $y_{i,t} \in \mathcal{Y}$ is a private signal observed by agent j at time t and $\mathcal{Y}$ is a finite set. Independence across time is also assumed.

The notion of observational equivalence is key to the results derived in [70], which are related to identifiability. Two states are observationally equivalent from the point of view of an agent if the conditional distributions of its signals under the two states coincide. More specifically, elements of the set $\Theta_i^\theta = \{\tilde{\theta} \in \Theta : l_i(y|\tilde{\theta}) = l_i(y|\theta) \forall y \in \mathcal{Y}\}$ are observationally equivalent to state θ from the point of view of agent i.

The belief update of each agent i is given by:

$$p_{i,t+1}(\theta) = a_{i,i} p_{i,t}(\theta) \frac{l_i(y_{i,t+1}|\theta)}{\mathcal{Z}_{i,t}(y_{i,t+1})} + \sum_{j \in \mathcal{N}_i} a_{i,j} p_{j,t}(\theta) \tag{5.3}$$

where $\mathcal{N}_i = \{j \in \mathcal{N} : (j, i) \in E\}$ is the neighborhood set of agent i. The denominator $\mathcal{Z}_{i,t}(y_{i,t+1})$ is the normalizing factor of the Bayesian update given by $\mathcal{Z}_{i,t}(y_{i,t+1}) = \sum_{\theta \in \Theta} p_{i,t}(\theta) l_i(y_{i,t+1}|\theta)$. The parameters $a_{i,i}$ are called the self-reliances. As noted in [70], we note that although the private signals are incorporated in a Bayesian manner, the belief update is non-Bayesian: agents treat the beliefs generated through linear combinations with their neighbors as Bayesian priors when incorporating their private signals.

In Proposition 3 of [70], it is shown that assuming:

- strong network connectivity (i.e., there exists a directed path from every agent to any other agent)
- $a_{i,i} > 0, \forall i$.
- $\exists i$ such that $p_{i,0}(\theta^*) > 0$.
- $\nexists \theta \neq \theta^*$ that is observationally equivalent to θ^* from the point of view of all agents in the network.

it follows that all agents in the network learn the true state of the world (assuming the true state of the world generates the observations) almost surely-i.e., $p_{i,t}(\theta^*) \to 1$ with probability 1 for all $i \in \mathcal{N}$ as $t \to \infty$.

This result is important because in spite of the non-Bayesian nature of the belief updates (significantly less computationally demanding than its Bayesian updating counterpart) and constant weights $a_{i,j}$, every agent in the social network will eventually learn the true state of the world. This holds even though the truth may not be recognizable to any individual.

In the controlled sensing problem studied in this paper, the true target location can be perfectly learned by any agent as the number of iterations grow (without any averaging required). It is shown numerically that estimation performance can be improved on average through decentralized averaging in addition to local repeated querying. In other words, decentralized averaging improves the uniformity over all sensors.

5.4 Decentralized Estimation Algorithm

Starting with a prior distribution $p_{i,0}(x)$ on X^*, the aim is to reach consensus across the network through repeated querying and information sharing. Motivated

by the optimality of the bisection rule for symmetric channels proved by Jedynak et al [73], the first stage of the decentralized estimation algorithm is to bisect the posterior of each agent $i \in \mathcal{N}$ at $\hat{X}_{i,t}$ and refine its belief through Bayes' rule. The second stage consists of each agent averaging its neighbor's beliefs and its own. This is repeated until convergence. The matrix A contains the weights for collaboration between agents and are allowed to be zero; if $a_{i,j} = 0$, then agent i cannot observe information from agent j at any time. The exact details are summarized in Algorithm 3.

Algorithm 3 Decentralized Estimation Algorithm

Input: $G = (\mathcal{N}, E), A = \{a_{i,j} : (i,j) \in \mathcal{N} \times \mathcal{N}\}, \{\epsilon_i : i \in \mathcal{N}\}$
Output: $\{\hat{X}_{i,t}, \check{X}_{i,t} : i \in \mathcal{N}\}$
Initialize $p_{i,0}(\cdot)$ to be positive everywhere.
repeat
  For each agent $i \in \mathcal{N}$:
    Bisect posterior density: $\mathbb{P}_{i,t}(A_{i,t}) = 1/2$.
    Obtain (noisy) binary response $y_{i,t+1} \in \{0, 1\}$.
    Belief update:

$$p_{i,t+1}(x) = a_{i,i} p_{i,t}(x) \frac{l_i(y_{i,t+1}|x, A_{i,t})}{\mathcal{Z}_{i,t}(y_{i,t+1})} + \sum_{j \in \mathcal{N}_i} a_{i,j} p_{j,t}(x), \qquad x \in \mathcal{X} \tag{5.4}$$

  where the observation probability mass function (p.m.f.) is:

$$l_i(y|x, A_{i,t}) = f_1^{(i)}(y) I(x \in A_{i,t}) + f_0^{(i)}(y) I(x \notin A_{i,t}), \qquad y \in \mathcal{Y}$$

  and $f_1^{(i)}(y) = (1-\epsilon_i)^{I(y=1)} \epsilon_i^{I(y=0)}, f_0^{(i)}(y) = 1 - f_1^{(i)}(y)$.
    Calculate target estimate: $\check{X}_{i,t} = \int_{\mathcal{X}} x p_{i,t}(x) dx$.
until convergence

We note that the normalizing factor $\mathcal{Z}_{i,t}(y)$ is given by $\int_{\mathcal{X}} p_{i,t}(x) l_i(y|x, \hat{X}_{i,t}) dx$ and can be shown to be equal to $1/2$ (see proof of Lemma V.6). In one dimension, $d = 1$, the query points are the medians $\hat{X}_{i,t} = F_{i,t}^{-1}(1/2)$ and the observation p.m.f. becomes:

$$l_i(y|x, \hat{X}_{i,t}) = f_1^{(i)}(y) I(x \leq \hat{X}_{i,t}) + f_0^{(i)}(y) I(x > \hat{X}_{i,t}).$$

We note two important differences between our density update (5.4) and the

update (5.3). The density $l_i(y|x, \hat{X}_{i,t})$ depends on the query point $\hat{X}_{i,t}$, which is time-varying and as a result, the density $l_i(y|x, \hat{X}_{i,t})$ is time-varying, unlike the time-invariant case in (5.3). Thus, the identifiability assumptions made in [70] do not make sense for our problem. In addition the update (5.4) holds pointwise for every $x \in \mathcal{X}$ and may not be bounded, unlike the discrete case in (5.3).

5.5 Convergence Analysis

5.5.1 Assumptions

To simplify the analysis of the algorithm, we make the following mild assumptions.

Assumption V.1. *(Conditional Independence) We assume that the players' responses are conditionally independent. In particular,*

$$\mathbb{P}(\mathbf{Y}_{t+1} = \mathbf{y}|\mathcal{F}_t) = \prod_{i=1}^{M} \mathbb{P}(Y_{i,t+1} = y_i|\mathcal{F}_t) \tag{5.5}$$

and each players response is governed by the observation density:

$$l_i(y_i|x, A_{i,t}) := \mathbb{P}(Y_{i,t+1} = y_i|A_{i,t}, X^* = x) = \begin{cases} f_1^{(i)}(y_i), & x \in A_{i,t} \\ f_0^{(i)}(y_i), & x \notin A_{i,t} \end{cases} \tag{5.6}$$

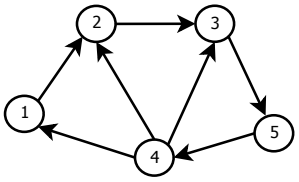
Assumption V.2. *(Memoryless Binary Symmetric Channels) We model the players' responses as independent (memoryless) binary symmetric channels (BSC) [35] with crossover probabilities $\epsilon_i \in (0, 1/2)$. The probability mass function $f_z^{(i)}(Y_{i,t+1}) = \mathbb{P}(Y_{i,t+1}|Z_{i,t} = z)$ is:*

$$f_z^{(i)}(y_i) = \begin{cases} 1 - \epsilon_i, & y_i = z \\ \epsilon_i, & y_i \neq z \end{cases}$$

for $i = 1, \ldots, M, z = 0, 1$. The assumption $\epsilon_i < 1/2$ implies that the response of each agent i is probably correct.

Assumption V.3. *(Strong Connectivity & Positive Self-reliances) As in [70], we also assume that the network is strongly connected and all self-reliances $a_{i,i}$ are strictly*

positive. The strong connectivity assumption implies that the interaction matrix A is irreducible. An example of a strongly connected network is shown in Fig. 5.5.1.



$$A = \begin{bmatrix} a_{1,1} & 0 & 0 & a_{1,4} & 0 \\ a_{2,1} & a_{2,2} & 0 & a_{2,4} & 0 \\ 0 & a_{3,2} & a_{3,3} & a_{3,4} & 0 \\ 0 & 0 & 0 & a_{4,4} & a_{4,5} \\ 0 & 0 & a_{5,3} & 0 & a_{5,5} \end{bmatrix}$$

5.5.2 Analysis

The density evolution (5.4) can be concisely written in matrix form as:

$$p_{t+1}(x) = (A + D_t(x))p_t(x), \qquad x \in \mathcal{X} \tag{5.7}$$

where A is the time-invariant interaction matrix and $D_t(x)$ is a diagonal time-varying matrix dependent on the responses $\mathbf{y}_{t+1} = (y_{1,t+1}, \ldots, y_{M,t+1})$, the query regions $A_{i,t} \subset \mathcal{X}$ [2] and the state $x \in \mathcal{X}$. The ith diagonal entry of $D_t(x)$ is given by:

$$[D_t(x)]_{i,i} = a_{i,i} \left(\frac{l_i(y_{i,t+1}|x, A_{i,t})}{\mathcal{Z}_{i,t}(y_{i,t+1})} - 1 \right)$$

We remark that the results of Jadbabaie et al [70] are not applicable here since the distributions $l_i(\cdot|x, A_{i,t})$ are time-varying because the query regions $A_{i,t}$ are time-varying.

To begin the analysis, we prove certain technical lemmas.

The next proposition provides bounds on the dynamic range of Ax, where x is any arbitrary vector.

Recall the coefficient of ergodicity of the interaction matrix A [107, 68]:

$$\tau_1(A) = \frac{1}{2} \max_{i \neq j} \|A^T(e_i - e_j)\|_1 = \frac{1}{2} \max_{i \neq j} \sum_{l=1}^{M} |a_{i,l} - a_{j,l}| \tag{5.8}$$

[2] In one-dimension $d = 1$, the query regions take the form $A_{i,t} = [0, \hat{X}_{i,t}]$.

The "more ergodic" the matrix A is, the closer $\tau_1(A)$ is to zero. One extreme is the identity matrix $A = I_M$, for which $\tau_1(A) = 1$. This makes intuitive sense since the identity matrix allow no information sharing. A matrix with fixed self-reliances $\alpha \in (0, 1)$ and uniform off-diagonal weights-i.e., $\frac{1-\alpha}{M-1}$, it is easy to check that $\tau_1(A) = |\alpha - \frac{1-\alpha}{M-1}|$.

Proposition V.4. *(Contraction Property of A) Assume $A = \{a_{i,j}\}$ is a $M \times M$ stochastic matrix. Let x be an arbitrary non-negative vector. Then, we have for all pairs (i, j):*

$$[Ax]_i - [Ax]_j \leq \tau_1(A) \left(\max_i x_i - \min_i x_i \right)$$

For a proof, see Theorem 3.1 in [107].

While the positivity assumption of Proposition V.4 might be restrictive for our problem, irreducibility of the matrix A implies that there exists r such that A^r is a stochastic matrix with positive entries [107]. This fact will be used in the analysis to follow.

Next, we recall a tight smooth approximation to the non-smooth maximum and minima operators. Similar results have appeared in Prop. 1 in [32] and p. 72 in [18].

Proposition V.5. *(Tight Smooth Approximation to Maximum/Minimum Operator) Let $a \in \mathbb{R}^M$ be an arbitrary vector. Then, we have for all $k > 0$:*

$$\max_i a_i \leq \frac{1}{k} \log \left(\sum_{i=1}^{M} \exp(k a_i) \right) \leq \max_i a_i + \frac{\log M}{k} \tag{5.9}$$

and

$$\min_i a_i \geq -\frac{1}{k} \log \left(\sum_{i=1}^{M} \exp(-k a_i) \right) \geq \min_i a_i - \frac{\log M}{k} \tag{5.10}$$

Lemma V.6. *Consider Algorithm 3. Let $B \in \mathcal{B}(\mathcal{X})$. Then, we have:*

$$\mathbb{E}\left[\int_B D_t(x) p_t(x) dx \middle| \mathcal{F}_t \right] = 0.$$

Proof. See Appendix A. □

Lemma V.7. *Consider Algorithm 3. Let $B \in \mathcal{B}(\mathcal{X})$. Then, we have $\mathbb{E}[v^T \mathbb{P}_{t+1}(B)|\mathcal{F}_t] = v^T \mathbb{P}_t(B)$ for some positive vector $v \succ 0$, and $\lim_{t\to\infty} v^T \mathbb{P}_t(B)$ exists almost surely.*

Proof. See Appendix B. □

The results to follow assume that the target lies in a bounded set $\mathcal{X} = [0, 1]$ for simplicity. While Lemmas V.6, V.7 and V.9 hold for any dimension $d \geq 1$, we remark that while the one-dimensional restriction of the target space $\mathcal{X} = [0, 1]$ might seem restrictive at first, the extensions of the proof techniques to higher dimensional spaces is a non-trivial problem. Similar problems are stated by Waeber et al. [126] in the context of extending the convergence theory of the probabilistic bisection algorithm (PBA) to higher dimensions that remain open to the best of our knowledge.

Lemma V.8. *Consider Algorithm 3. Let $B = [0, b] \in \mathcal{B}(\mathcal{X})$. Then, there exists $v_i > 0$ such that:*

$$\prod_{i=1}^{M} \cosh\left(v_i a_{i,i}(1 - 2\epsilon_i) \min\left\{\mathbb{P}_{i,t}(B), 1 - \mathbb{P}_{i,t}(B)\right\}\right) \xrightarrow{a.s.} 1 \tag{5.11}$$

as $t \to \infty$.

Proof. See Appendix C. □

Define the dynamic range (with respect to all agents in the network) of the posterior distribution integrated over the set B as:

$$V_t(B) = \max_i \mathbb{P}_{i,t}(B) - \min_i \mathbb{P}_{i,t}(B) \tag{5.12}$$

Also, define the innovation term:

$$d_{i,t+1}(B) = \left[\int_B D_t(x) p_t(x) dx\right]_i = \int_B [D_t(x)]_{i,i} p_{i,t}(x) dx$$

We next prove a lemma that shows that the dynamic range $V_t(B)$ follows an exponential decay law up to a perturbation given by the dynamic range of the innovation terms.

Lemma V.9. *Consider Algorithm 3. Let $B = [0, b] \in \mathcal{B}([0, 1])$. Then, for all $r \in \mathbb{N}$:*

$$V_{t+r}(B) \leq \tau_1(A^r) V_t(B) + \sum_{k=0}^{r-1} \left(\max_i d_{i,t+r-k}(B) - \min_i d_{i,t+r-k}(B) \right) \tag{5.13}$$

In addition, there exists a finite $r \in \mathbb{N}$ such that $\tau_1(A^r) < 1$.

Proof. See Appendix D. □

Theorem V.10. *(Asymptotic Agreement/Consensus) Consider Algorithm 3. Let $B = [0, b] \in \mathcal{B}([0, 1])$. Then, consensus of the agents' beliefs is asymptotically achieved across the network:*

$$V_t(B) = \max_i \mathbb{P}_{i,t}(B) - \min_i \mathbb{P}_{i,t}(B) \xrightarrow{p.} 0$$

as $t \to \infty$.

Proof. See Appendix E. □

To proceed, we need another lemma.

Lemma V.11. *Consider Algorithm 3. Let v be the left eigenvector of A corresponding to the unit eigenvalue. Assume that for all agents i, $p_{i,0}(X^*) > 0$. Then, the posteriors evaluated at the true target location X^* have the following asymptotic behavior:*

$$\lim_{t \to \infty} \frac{1}{t} \sum_{i=1}^{M} v_i \log(p_{i,t}(X^*)) \geq \sum_{i=1}^{M} v_i a_{i,i} C(\epsilon_i) = K(\boldsymbol{\epsilon})$$

where $C(\epsilon)$ is the capacity of the BSC.

Proof. See Appendix F. □

Now, we are ready to prove the main consistency result of the asymptotic beliefs.

Theorem V.12. *(Convergence of Beliefs to a Deterministic Limit & Consistency) Consider Algorithm 3. Let $B = [0, b] \in \mathcal{B}([0, 1])$. Then, we have for each $i \in \mathcal{N}$:*

$$F_{i,t}(b) = \mathbb{P}_{i,t}(B) \xrightarrow{p.} F_{\infty}(b) = \begin{cases} 0, & b < X^* \\ 1, & b > X^* \end{cases}$$

as $t \to \infty$. In addition, for all $i \in \mathcal{N}$:

$$\check{X}_{i,t} := \int_{x=0}^{1} x p_{i,t}(x) dx \xrightarrow{p.} X^*$$

Proof. See Appendix G. □

The next Corollary generalizes the result of Theorem V.12.

Corollary V.13. *Consider Algorithm 3. Let $B = \cup_{k=1}^{K} I_k \in \mathcal{B}([0, 1])$ be a finite union of disjoint intervals $I_k = [a_k, b_k)$. Then, for each $i \in \mathcal{N}$:*

$$\mathbb{P}_{i,t}(B) \xrightarrow{p.} \begin{cases} 0, & X^* \notin B \\ 1, & X^* \in B \end{cases}$$

as $t \to \infty$.

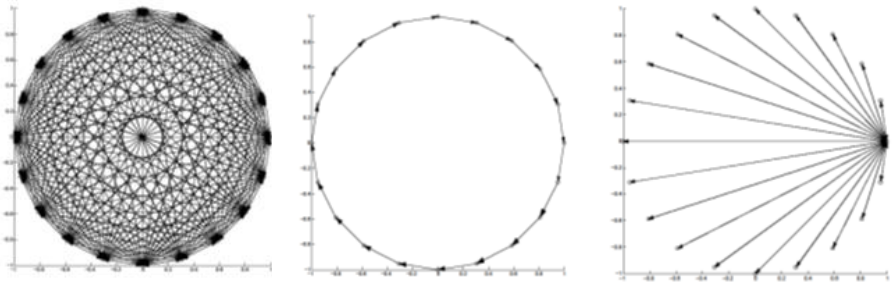
Borel sets in one-dimension can be represented as countable union of disjoint intervals, so Cor. V.13 almost holds for all Borel sets.

5.6 Simulations

This section contains a few simulations that validate the methodology presented throughout the paper and illustrate the benefits of belief sharing.

Three graph topologies were considered in this paper to test the robustness of the methodology and are shown in Fig. 5.2. We consider $M = 20$ agents implementing Algorithm 3 for 1000 iterations. The mean-squared error (MSE) was chosen as a

performance metric. For convergence $\hat{X}_{i,t} \to X^*$, we expect the MSE to converge to zero.



a) Fully connected graph b) Cyclic graph c) Star graph

Figure 5.2: Graph topologies considered in this paper.

The instantaneous MSE for agent i was calculated using $MSE_{i,t} = (\hat{X}_{i,t} - X^*)^2$. The min, max and avg RMSE metrics were calculated as:

$$RMSE_{min} = \sqrt{\frac{1}{T}\sum_{t=1}^{T} \min_i MSE_{i,t}}$$

$$RMSE_{max} = \sqrt{\frac{1}{T}\sum_{t=1}^{T} \max_i MSE_{i,t}}$$

$$RMSE_{avg} = \sqrt{\frac{1}{T}\sum_{t=1}^{T} \frac{1}{M}\sum_{i=1}^{M} MSE_{i,t}}$$

The min and max metrics represent the worst and best performance over all the agents. In the plots, the legends I:min, I:max and I:avg denote the min, max and avg performance for the special case of $A = I$ (i.e., no information sharing), while the legends A:min, A:max and A:avg denote the min, max and avg performance for the case of $A \neq I$ (i.e., decentralized averaging).

5.6.1 Uniformly bad sensors

In this setup, all agents in the network have the same error probability $\epsilon = \epsilon_i = 0.4$. The self-reliance parameters of each agent were set to 0.95 and the rest of the

parameters were made equal such that each row of A sums to unity. Across all graph

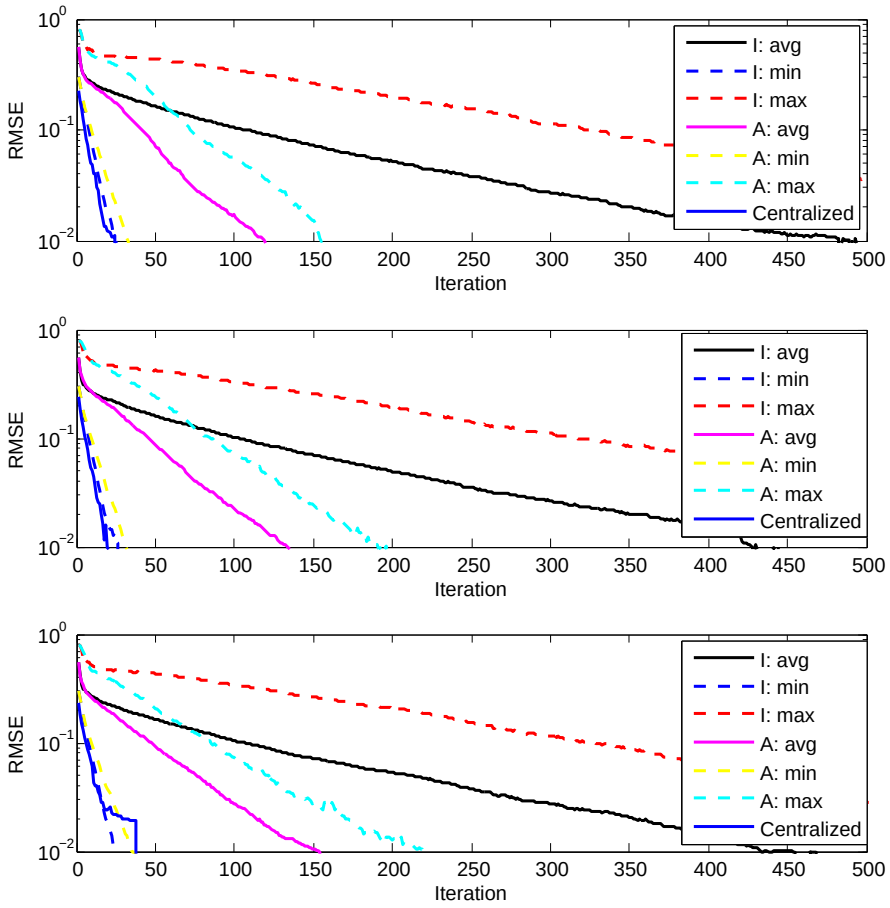


Figure 5.3: RMSE performance of the estimator for the fully connected network (top), cyclic network (middle) and star network (bottom). The average and worst-case MSE across the network is lower for the case of averaging vs. the case of no information sharing. The target location was set to $X^* = 0.8$. The curves plotted are results of averaging error performance over 500 Monte Carlo runs.

topologies, the major trends are the same. Fig. 5.3 shows that the average and best case performance of the network in terms of RMSE is improved by averaging beliefs in the network. Of course, this occurs at a slight reduction in performance for the best case performance. In terms of average RMSE performance, the decentralized averaging seems to be linear in a logarithmic scale, which implies exponential decay in the MSE as a function of iterations. In addition, the decentralized averaging algorithm seems to have a different slope than the corresponding algorithm with no

averaging, which implies a better rate exponent. The posterior entropy criterion is interesting since we observe a phase transition-i.e., after enough iterations, the decentralized estimation algorithm (with averaging) begins to outperform the case of no information sharing. In terms of MSE, the decentralized estimation algorithm with averaging uniformly outperforms the estimation algorithm with no information sharing for all iterations.

We empirically observed that larger gains were obtained in the low SNR regime (i.e. larger error probabilities). In the high SNR regime, averaging tends to appear like noise and thus may hurt performance instead of helping. Interesting phenomena occur when the network operates in a middle regime, where some sensors have high SNR and some have low SNR.

5.6.2 A good sensor injected in a set of bad sensors

In this setup, one agent has an error probability $\epsilon = 0.05$ (good agent) and the rest of the agents in the network have an error probability $\epsilon = \epsilon_i = 0.45$ (bad agents). The self-reliance parameters of all agents were set to 0.95. The rest of the parameters were made equal such that each row of A sums to unity.

The centralized fully Bayesian estimation algorithm, which knows the error probabilities of all agents and has access to all of the agents' observations and queries, is also implemented. Due to the intractability of the jointly optimal query design, we make use of the basic equivalence principle derived in [118], and implement the centralized method using a series of bisections (one per agent). The equivalence principle shows that this sequential bisection algorithm achieves the same performance as the jointly optimal algorithm on average. The RMSE performance is plotted in Fig. 5.4 for the three different topologies of Fig. 5.2. It is observed that the decentralized performance mimics the centralized performance quite well, and the performance of

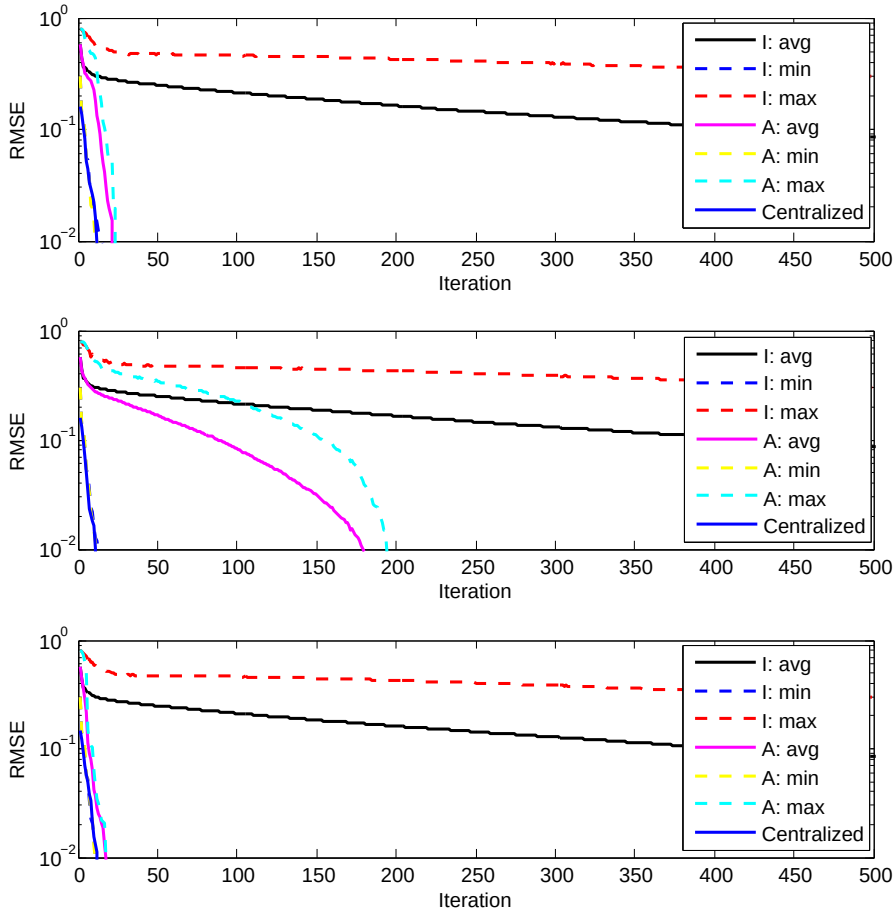


Figure 5.4: RMSE performance of the estimator for the fully connected network (top), cyclic network (middle) and star network (bottom). The MSE across the network is lower for the case of averaging vs. the case of no information sharing. Decentralized averaging tends to match the centralized performance, while the algorithm with no averaging lags quite a bit behind. The target location was set to $X^* = 0.8$. The curves plotted are results of averaging error performance over 500 Monte Carlo runs.

the algorithm with no averaging lags quite a bit behind the decentralized averaging and centralized algorithms. Thus, we conclude that the one good sensor tends to have a significant influence on the beliefs of the the bad agents in the network, almost matching the performance of the centralized estimator. In addition, we note that the best MSE performance of the algorithm with no averaging (I:min) corresponds to the performance of the good sensor in the network and is fairly close to the average performance of the decentralized averaging algorithm (A:avg), thus improving the

uniformity of the bad sensors. This significant result on robustness is due to the fact that the decentralized algorithm is implemented in parallel and each agent is completely agnostic of the error probabilities of all the other agents in the network. Out of the three topologies, we empirically observe that the cyclic network has the largest performance gap with respect to the centralized algorithm.

5.6.3 Random Error Probabilities

In this setup, the error probabilities of all agents are chosen i.i.d. uniformly from the feasible set $(0, 1/2)$. For the simulation shown here, the smallest error probability was 0.0197 and the largest error probability was 0.4909, thus the network contains at least one good agent and one bad agent. The self-reliance parameters of all agents were set to 0.95. The rest of the parameters were made equal such that each row of A sums to unity. The RMSE performance is plotted in Fig. 5.5 for the three different topologies of Fig. 5.2. It is observed that the decentralized performance is not too far behind the centralized performance, and the performance of the algorithm with no averaging lags quite a bit behind the decentralized averaging and centralized algorithms. Thus, although the off-diagonal weights for each row of the interaction matrix A are uniformly spread and the self-reliances are identical (i.e. the diagonals of A), the decentralized estimation algorithm is robust to the fluctuations of the error probabilities. Out of the three topologies, we empirically observe that the cyclic network has the largest performance gap with respect to the centralized algorithm.

5.7 Conclusion

We introduced the problem of decentralized 20 questions with noise and illustrated several benefits over the case of no information sharing through analysis and simulations. At each iteration of our proposed decentralized algorithm, agents query and

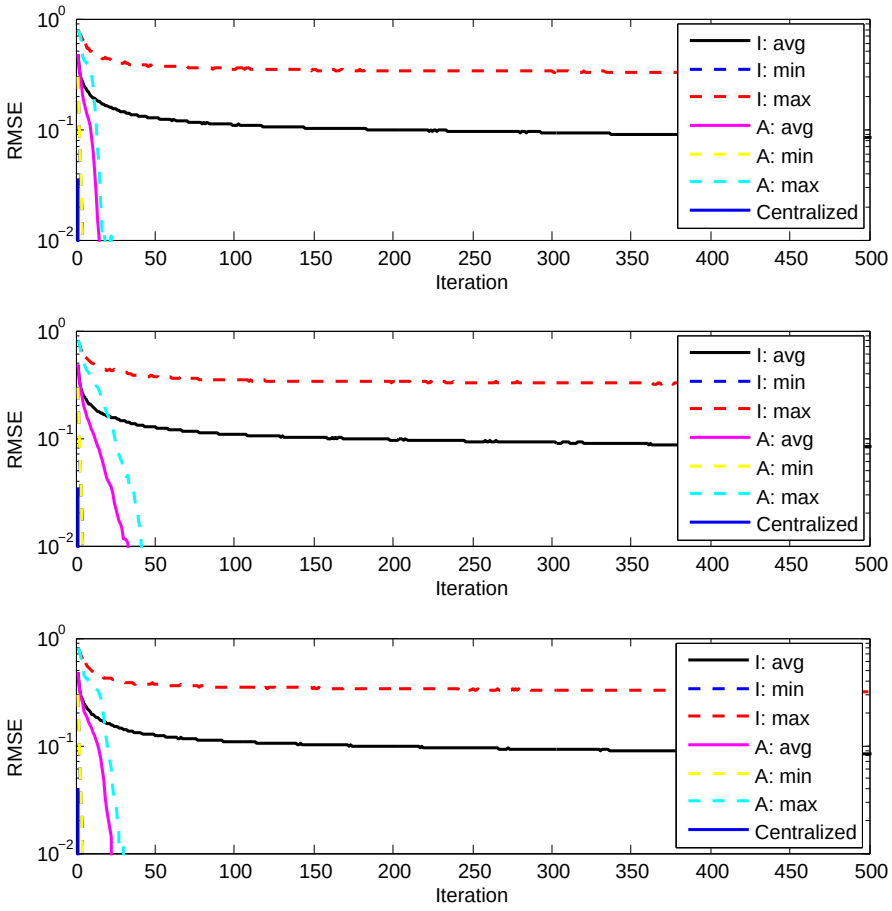


Figure 5.5: RMSE performance of the estimator for the fully connected network (top), cyclic network (middle) and star network (bottom). The MSE across the network is lower for the case of averaging vs. the case of no information sharing. Decentralized averaging tends to match the centralized performance, while the algorithm with no averaging lags quite a bit behind. The target location was set to $X^* = 0.8$. The curves plotted are results of averaging error performance over 500 Monte Carlo runs.

respond based on their local beliefs and average information through their neighbors. Asymptotic convergence properties of the agents' beliefs were derived, showing that they reach consensus to the true belief. Simulations were presented to validate the convergence properties of the algorithm and test the robustness of the methodology.

5.8 Appendix

5.8.1 Proof of Lemma V.6

Proof. Without loss of generality, fix $i \in \mathcal{N}$. From direct substitution and integration, we have:

$$\begin{aligned}\int_B [D_t(x)]_{i,i} p_{i,t}(x)dx &= a_{i,i}\left(\frac{\int_B l_i(y_{i,t+1}|x, A_{i,t})p_{i,t}(x)dx}{\mathcal{Z}_{i,t}(y_{i,t+1})} - \int_B p_{i,t}(x)dx\right) \\ &= a_{i,i}\left(2\int_B l_i(y_{i,t+1}|x, A_{i,t})p_{i,t}(x)dx - \mathbb{P}_{i,t}(B)\right)\end{aligned}$$

where we used the fact that $\mathcal{Z}_{i,t}(y) = 1/2$ for all $y \in \mathcal{Y}$. This follows from the probabilistic bisection property:

$$\begin{aligned}\mathcal{Z}_{i,t}(y) &= \int_{\mathcal{X}} p_{i,t}(x)\left(f_1^{(i)}(y)I(x \in A_{i,t}) + f_0^{(i)}(y)I(x \notin A_{i,t})\right)dx \\ &= f_1^{(i)}(y)\mathbb{P}_{i,t}(A_{i,t}) + f_0^{(i)}(y)\mathbb{P}_{i,t}(A_{i,t}^c) \\ &= f_1^{(i)}(y)(1/2) + f_0^{(i)}(y)(1/2) \\ &= 1/2\end{aligned}$$

where we used the fact $f_1^{(i)}(y) + f_0^{(i)}(y) = 1$. From the definition of $l_i(y|x, A_{i,t})$, it follows that:

$$\int_B [D_t(x)]_{i,i} p_{i,t}(x)dx = a_{i,i}\left(2\left(f_1^{(i)}(y_{i,t+1})\mathbb{P}_{i,t}(B \cap A_{i,t}) + f_0^{(i)}(y_{i,t+1})\mathbb{P}_{i,t}(B \cap A_{i,t}^c)\right) - \mathbb{P}_{i,t}(B)\right)$$

Taking the conditional expectation of both sides, we obtain:

$$\begin{aligned}
&\mathbb{E}\left[\int_B [D_t(x)]_{i,i} p_{i,t}(x)dx \middle| \mathcal{F}_t\right] \\
&= a_{i,i}\left(2\mathbb{E}\left[f_1^{(i)}(Y_{i,t+1})\mathbb{P}_{i,t}(B\cap A_{i,t}) + f_0^{(i)}(Y_{i,t+1})\mathbb{P}_{i,t}(B\cap A_{i,t}^c)|\mathcal{F}_t\right] - \mathbb{P}_{i,t}(B)\right) \\
&= a_{i,i}\left(2\sum_{y=0}^{1}\left(f_1^{(i)}(y)\mathbb{P}_{i,t}(B\cap A_{i,t}) + f_0^{(i)}(y)\mathbb{P}_{i,t}(B\cap A_{i,t}^c)\right)\mathbb{P}(Y_{i,t+1}=y|\mathcal{F}_t) - \mathbb{P}_{i,t}(B)\right) \\
&= a_{i,i}\left(\sum_{y=0}^{1}\left(f_1^{(i)}(y)\mathbb{P}_{i,t}(B\cap A_{i,t}) + f_0^{(i)}(y)\mathbb{P}_{i,t}(B\cap A_{i,t}^c)\right) - \mathbb{P}_{i,t}(B)\right) \\
&= a_{i,i}\left(\left(\mathbb{P}_{i,t}(B\cap A_{t,i}) + \mathbb{P}_{i,t}(B\cap A_{i,t}^c)\right) - \mathbb{P}_{i,t}(B)\right) \\
&= a_{i,i}\left(\mathbb{P}_{i,t}(B) - \mathbb{P}_{i,t}(B)\right) = 0
\end{aligned}$$

where we used the fact that under the probabilistic bisection, $\mathbb{P}(Y_{i,t+1}=y|\mathcal{F}_t)=1/2$ for all y. This follows from:

$$\begin{aligned}
&\mathbb{P}(Y_{i,t+1}=y|\mathcal{F}_t) \\
&\quad= \sum_{z=0}^{1}\mathbb{P}(Y_{i,t+1}=y|Z_{i,t}=z,\mathcal{F}_t)\mathbb{P}(Z_{i,t}=z|\mathcal{F}_t) \\
&\quad= \mathbb{P}(Y_{i,t+1}=y|Z_{i,t}=0)\mathbb{P}(Z_{i,t}=0|\mathcal{F}_t) + \mathbb{P}(Y_{i,t+1}=y|Z_{i,t}=1)\mathbb{P}(Z_{i,t}=1|\mathcal{F}_t) \\
&\quad= f_0(y)\mathbb{P}(X^*\notin A_{i,t}|\mathcal{F}_t) + f_1(y)\mathbb{P}(X^*\in A_{i,t}|\mathcal{F}_t) \\
&\quad= f_0(y)\mathbb{P}_{i,t}(A_{i,t}^c) + f_1(y)\mathbb{P}_{i,t}(A_{i,t}) \\
&\quad= 1/2
\end{aligned}$$

Since i was arbitrarily chosen, the proof is complete. □

5.8.2 Proof of Lemma V.7

Proof. From strong connectivity (i.e., Assumption 3), it follows that A is an irreducible stochastic matrix. Thus, there exists a left eigenvector $v\in\mathbb{R}^M$ with strictly positive entries corresponding to a unit eigenvalue-i.e., $v^T = v^T A$ [9].

Integrating (5.7) and left-multiplying by v^T:

$$v^T \int_B p_{t+1}(x)dx = v^T A \int_B p_t(x)dx + v^T \int_B D_t(x)p_t(x)dx$$

$$\Leftrightarrow v^T \mathbb{P}_{t+1}(B) = v^T \mathbb{P}_t(B) + \sum_{i=1}^{M} v_i \int_B [D_t(x)]_{i,i} p_{i,t}(x)dx \tag{5.14}$$

Taking the conditional expectation of both sides and using Lemma V.6, we obtain $\mathbb{E}[v^T \mathbb{P}_{t+1}(B)|\mathcal{F}_t] = v^T \mathbb{P}_t(B)$. Thus, the process $\{v^T \mathbb{P}_t(B) : t \geq 0\}$ is a martingale with respect to the filtration $\mathcal{F}_t$. We note that it is bounded below by zero and above by $\|v\|_1$ almost surely. From the martingale convergence theorem, it follows that it converges almost surely. □

5.8.3 Proof of Lemma V.8

Proof. Define the tilted measure variable $\zeta_t(B) = \exp(v^T \mathbb{P}_t(B))$. From Lemma V.7 and Jensen's inequality, it follows that

$$\mathbb{E}[\zeta_{t+1}(B)|\mathcal{F}_t] \geq \zeta_t(B)$$

so the process $\{\zeta_t(B) : t \geq 0\}$ is a submartingale with respect to the filtration $\mathcal{F}_t$. From the proof of Lemma V.7, it follows that $\zeta_t(B)$ is bounded a.s., so by the martingale convergence theorem, it follows that $\lim_{t\to\infty} \zeta_t(B)$ exists and is finite almost surely. As a result, we have from (5.14):

$$\lim_{t\to\infty} \frac{\zeta_{t+1}(B)}{\zeta_t(B)} \overset{a.s.}{=} 1 \overset{a.s.}{=} \lim_{t\to\infty} \exp\left(v^T \int_B D_t(x)p_t(x)dx\right)$$

Since the variables in the limit on the RHS are bounded a.s.-i.e.,

$$\begin{aligned}
&\left| v^T \int_B D_t(x) p_t(x) dx \right| \\
&\quad \leq \|v\|_1 \max_i \left| \int_B [D_t(x)]_{i,i} p_{i,t}(x) dx \right| \\
&\quad = \|v\|_1 \max_i (2(f_1^{(i)}(Y_{i,t+1}) \mathbb{P}_{i,t}(B \cap A_{i,t}) + f_0^{(i)}(Y_{i,t+1}) \mathbb{P}_{i,t}(B \cap A_{i,t}^c)) - \mathbb{P}_{i,t}(B)) \\
&\quad \leq \|v\|_1 \max_i (2(1-\epsilon_i)\mathbb{P}_{i,t}(B) - \mathbb{P}_{i,t}(B)) \\
&\quad \leq \|v\|_1 (1 - 2\min_i \epsilon_i) \leq \|v\|_1 < \infty
\end{aligned}$$

the dominated convergence theorem for conditional expectations [48] implies:

$$\mathbb{E}\left[\exp\left(v^T \int_B D_t(x) p_t(x) dx \right) \middle| \mathcal{F}_t \right] \xrightarrow{a.s.} 1 \tag{5.15}$$

as $t \to \infty$. Substituting the definition of $D_t(x)$ into (5.15) and using Assumption 1, it follows after some algebra that (5.15) is equivalent to:

$$\prod_{i=1}^{M} \frac{\mathbb{E}\left[\exp\left(v_i a_{i,i} \int_B 2 l_i(Y_{i,t+1}|x, A_{i,t}) p_{i,t}(x) dx \right) \middle| \mathcal{F}_t \right]}{\exp(v_i a_{i,i} \mathbb{P}_{i,t}(B))} \xrightarrow{a.s.} 1 \tag{5.16}$$

Next, we analyze the ratio of exponentials for two separate cases. First, consider the case $\mathbb{P}_{i,t}([0,b]) = \int_0^b p_{i,t}(x)dx \leq 1/2$. Using the definition of $\hat{X}_{i,t}$, it follows that $b \leq \hat{X}_{i,t}$. This implies that $l_i(y|x, A_{i,t}) = f_1^{(i)}(y)$ for all $x \leq b$. Using this fact and $\mathbb{P}(Y_{i,t+1} = y|\mathcal{F}_t) = 1/2$:

$$\begin{aligned}
&\frac{\mathbb{E}\left[\exp\left(v_i a_{i,i} \int_B 2 l_i(Y_{i,t+1}|x, A_{i,t}) p_{i,t}(x) dx \right) \middle| \mathcal{F}_t \right]}{\exp(v_i a_{i,i} \mathbb{P}_{i,t}(B))} \\
&\quad = \frac{1}{2} \frac{\exp(v_i a_{i,i} 2(1-\epsilon_i)\mathbb{P}_{i,t}(B)) + \exp(v_i a_{i,i} 2\epsilon_i \mathbb{P}_{i,t}(B))}{\exp(v_i a_{i,i} \mathbb{P}_{i,t}(B))} \\
&\quad = \frac{1}{2} \left(\exp(v_i a_{i,i}(1-2\epsilon_i)\mathbb{P}_{i,t}(B)) + \exp(-v_i a_{i,i}(1-2\epsilon_i)\mathbb{P}_{i,t}(B)) \right) \\
&\quad = \cosh(v_i a_{i,i}(1-2\epsilon_i)\mathbb{P}_{i,t}(B))
\end{aligned} \tag{5.17}$$

where we used the fact that $(e^a + e^{-a})/2 = \cosh(a)$. Second, consider the complementary case $\mathbb{P}_{i,t}([0,b]) > 1/2$. In this case, we have $b > \hat{X}_{i,t}$ and as a result:

$$\begin{aligned}\int_0^b 2l_i(Y_{i,t+1}|x, A_{t,i})p_{i,t}(x)dx &= \int_0^{\hat{X}_{i,t}} 2f_1^{(i)}(Y_{i,t+1})p_{i,t}(x)dx + \int_{\hat{X}_{i,t}}^b 2f_0^{(i)}(Y_{i,t+1})p_{i,t}(x)dx \\ &= 2f_1^{(i)}(Y_{i,t+1})\mathbb{P}_{i,t}(A_{i,t}) + 2f_0^{(i)}(Y_{i,t+1})(\mathbb{P}_{i,t}(B) - \mathbb{P}_{i,t}(A_{i,t})) \\ &= f_1^{(i)}(Y_{i,t+1}) + f_0^{(i)}(Y_{i,t+1})(2\mathbb{P}_{i,t}(B) - 1) \\ &= \begin{cases} (1-2\epsilon_i) + 2\epsilon_i\mathbb{P}_{i,t}(B), & Y_{i,t+1} = 1 \\ 2(1-\epsilon_i)\mathbb{P}_{i,t}(B) + (2\epsilon_i - 1), & Y_{i,t+1} = 0 \end{cases}\end{aligned}$$

Using this result and $\mathbb{P}(Y_{i,t+1} = y|\mathcal{F}_t) = 1/2$:

$$\begin{aligned}&\frac{\mathbb{E}\left[\exp\left(v_i a_{i,i}\int_B 2l_i(Y_{i,t+1}|x, A_{i,t})p_{i,t}(x)dx\right)\middle|\mathcal{F}_t\right]}{\exp(v_i a_{i,i}\mathbb{P}_{i,t}(B))} \\ &\quad= \frac{1}{2}\frac{\exp(v_i a_{i,i}((1-2\epsilon_i) + 2\epsilon_i\mathbb{P}_{i,t}(B))) + \exp(v_i a_{i,i}(2(1-\epsilon_i)\mathbb{P}_{i,t}(B) + (2\epsilon_i - 1)))}{\exp(v_i a_{i,i}\mathbb{P}_{i,t}(B))} \\ &\quad= \frac{1}{2}\left(\exp(v_i a_{i,i}(1-2\epsilon_i)(1-\mathbb{P}_{i,t}(B))) + \exp(-v_i a_{i,i}(1-2\epsilon_i)(1-\mathbb{P}_{i,t}(B)))\right)\end{aligned}$$

(5.18)

$$= \cosh(v_i a_{i,i}(1-2\epsilon_i)\mathbb{P}_{i,t}(B^c))$$

Combining the two cases (5.17) and (5.18) by noting that

$$\min\{\mathbb{P}_{i,t}(B), 1-\mathbb{P}_{i,t}(B)\} = \begin{cases} \mathbb{P}_{i,t}(B), & \mathbb{P}_{i,t}(B) \le 1/2 \\ 1-\mathbb{P}_{i,t}(B), & \mathbb{P}_{i,t}(B) > 1/2 \end{cases},$$

we have:

$$\begin{aligned}&\frac{\mathbb{E}\left[\exp\left(v_i a_{i,i}\int_B 2l_i(Y_{i,t+1}|x, A_{i,t})p_{i,t}(x)dx\right)\middle|\mathcal{F}_t\right]}{\exp(v_i a_{i,i}\mathbb{P}_{i,t}(B))} \\ &\quad= \cosh\left(v_i a_{i,i}(1-2\epsilon_i)\min\{\mathbb{P}_{i,t}(B), 1-\mathbb{P}_{i,t}(B)\}\right)\end{aligned}$$

The proof is completed by substituting this expression into (5.16). □

5.8.4 Proof of Lemma V.9

Proof. Integrating both sides of the recursion (5.7):

$$\mathbb{P}_{t+1}(B) = A\mathbb{P}_t(B) + d_{t+1}(B) \tag{5.19}$$

Unrolling (5.19) over r steps:

$$\mathbb{P}_{t+r}(B) = A^r\mathbb{P}_t(B) + \sum_{k=0}^{r-1} A^k d_{t+r-k}(B) \tag{5.20}$$

Since A is a stochastic matrix, Proposition V.4 implies:

$$\begin{aligned}
V_{t+r}(B) &= \max_i \mathbb{P}_{i,t+r}(B) - \min_i \mathbb{P}_{i,t+r}(B) \\
&\leq \tau_1(A^r)V_t(B) + \max_{i,j} \sum_{k=0}^{r-1} \left([A^k d_{t+r-k}(B)]_i - [A^k d_{t+r-k}(B)]_j\right) \\
&\leq \tau_1(A^r)V_t(B) + \sum_{k=0}^{r-1} \left(\max_i [A^k d_{t+r-k}(B)]_i - \min_i [A^k d_{t+r-k}(B)]_i\right) \\
&\leq \tau_1(A^r)V_t(B) + \sum_{k=0}^{r-1} \left(\max_i d_{i,t+r-k}(B) - \min_i d_{i,t+r-k}(B)\right)
\end{aligned}$$

It is known that $\tau_1(A^r) \in [0, 1]$ for any $r \in \mathbb{N}$ [68, 107]. The irreducibility of the matrix A implies the existence of a positive r such that $\tau_1(A^r) < 1$ [107]. □

5.8.5 Proof of Theorem V.10

Proof. To show convergence of the integrated beliefs of all agents in the network to a common limiting belief, it suffices to show $V_t(B) \xrightarrow{p_1} 0$. While this method of proof does not allow identification of the limiting belief, it shows a global equilibrium exists and yields insight into the rate of convergence through the ergodicity properties of A. The structure of the limiting belief is studied in Theorem V.12.

Without loss of generality, we consider the case $r = 1$ in Lemma V.9. The case $r > 1$ follows similarly. From Lemma V.9, we obtain:

$$\mathbb{E}[V_{t+1}(B)|\mathcal{F}_t] \leq \tau_1(A)V_t(B) + \mathbb{E}\left[\max_i d_{i,t+1}(B) - \min_i d_{i,t+1}(B)\middle|\mathcal{F}_t\right] \tag{5.21}$$

where $\tau_1(A) < 1$. To continue, we need to show that the remainder is asymptotically negligible-i.e.,

$$\mathbb{E}\left[\max_i d_{i,t+1}(B) - \min_i d_{i,t+1}(B)\middle|\mathcal{F}_t\right] \to 0.$$

Using Proposition V.5, we obtain for any $k > 0$:

$$\begin{aligned}
&\mathbb{E}\left[\max_i d_{i,t+1}(B) - \min_i d_{i,t+1}(B)\middle|\mathcal{F}_t\right] \\
&\leq \frac{1}{k}\mathbb{E}\left[\log\left(\sum_{i=1}^{M}\exp(kd_{i,t+1}(B))\right) + \log\left(\sum_{i=1}^{M}\exp(-kd_{i,t+1}(B))\right)\middle|\mathcal{F}_t\right] \\
&\leq \frac{1}{k}\left[\log\left(\sum_{i=1}^{M}\mathbb{E}[\exp(kd_{i,t+1}(B))|\mathcal{F}_t]\right) + \log\left(\sum_{i=1}^{M}\mathbb{E}[\exp(-kd_{i,t+1}(B))|\mathcal{F}_t]\right)\right]
\end{aligned} \tag{5.22}$$

where we used Jensen's inequality and the linearity of expectation.

Using similar analysis as in the proof of Lemma V.8, the (conditional) moment generating functions of the innovation terms can be written as hyperbolic cosines:

$$\begin{aligned}
\mathbb{E}[e^{kd_{i,t+1}(B)}|\mathcal{F}_t] &= \cosh\left(ka_{i,i}(1-2\epsilon_i)\min\left\{\mathbb{P}_{i,t}(B), 1-\mathbb{P}_{i,t}(B)\right\}\right) \\
\mathbb{E}[e^{-kd_{i,t+1}(B)}|\mathcal{F}_t] &= \cosh\left(-ka_{i,i}(1-2\epsilon_i)\min\left\{\mathbb{P}_{i,t}(B), 1-\mathbb{P}_{i,t}(B)\right\}\right)
\end{aligned}$$

Using the even symmetry of the $\cosh(\cdot)$ function, substituting these expressions into (5.22) and using Proposition V.5 again, we obtain:

$$\begin{aligned}
&\mathbb{E}\left[\max_i d_{i,t+1}(B) - \min_i d_{i,t+1}(B)\middle|\mathcal{F}_t\right] \\
&\leq \frac{2}{k}\log\left(\sum_{i=1}^{M}\cosh\left(ka_{i,i}(1-2\epsilon_i)\min\left\{\mathbb{P}_{i,t}(B), 1-\mathbb{P}_{i,t}(B)\right\}\right)\right) \\
&\leq \frac{2}{k}\log\left(\sum_{i=1}^{M}\exp\left(ka_{i,i}(1-2\epsilon_i)\min\left\{\mathbb{P}_{i,t}(B), 1-\mathbb{P}_{i,t}(B)\right\}\right)\right) \\
&\leq 2\left(\max_i\left\{a_{i,i}(1-2\epsilon_i)\min\left\{\mathbb{P}_{i,t}(B), 1-\mathbb{P}_{i,t}(B)\right\}\right\} + \frac{\log M}{k}\right)
\end{aligned}$$

Taking the limit $k \to \infty$ to tighten the bound and using (5.21):

(5.23)
$$\mathbb{E}[V_{t+1}(B)|\mathcal{F}_t] \leq \tau_1(A)V_t(B) + 2\max_i \left\{ a_{i,i}(1-2\epsilon_i)\min\left\{\mathbb{P}_{i,t}(B), 1-\mathbb{P}_{i,t}(B)\right\} \right\}$$

Lemma V.8 implies that $\cosh\left(v_i a_{i,i}(1-2\epsilon_i)\min\left\{\mathbb{P}_{i,t}(B), 1-\mathbb{P}_{i,t}(B)\right\}\right) \to 1$ for all $i \in \mathcal{N}$. Since $1 = \cosh(0) \leq \cosh(x)$ for all $x \in \mathbb{R}$, it follows that $\min\{\mathbb{P}_{i,t}(B), 1 - \mathbb{P}_{i,t}(B)\} \to 0$ almost surely. Note that here we used the positivity of the v_i and the self-reliances $a_{i,i}$ (i.e., Assumption 3) along with the fact that $\epsilon_i < 1/2$. Define the non-negative sequence $\delta_t := 2\max_i \left\{a_{i,i}(1-2\epsilon_i)\min\left\{\mathbb{P}_{i,t}(B), 1-\mathbb{P}_{i,t}(B)\right\}\right\}$. The above implies $\delta_t \overset{a.s.}{\to} 0$ as $t \to \infty$.

Taking the unconditional expectation of both sides in (5.23):

$$\mathbb{E}[V_{t+1}(B)] \leq \tau_1(A)\mathbb{E}[V_t(B)] + \mathbb{E}[\delta_t] \tag{5.24}$$

where $\mathbb{E}[\delta_t] \to 0$ by the dominated convergence theorem. Using induction on (5.24), we obtain for all $t \geq 0$:

$$\mathbb{E}[V_t(B)] \leq \tau_1(A)^t\mathbb{E}[V_0(B)] + \sum_{l=0}^{t-1} \tau_1(A)^l\mathbb{E}[\delta_{t-1-l}] \tag{5.25}$$

Taking the limits of both sides of (5.25) and using the fact that $\tau_1(A) < 1$ and $\mathbb{E}[V_0(B)] < \infty$:

$$\lim_{t\to\infty}\mathbb{E}[V_t(B)] \leq \lim_{t\to\infty}\tau_1(A)^t\mathbb{E}[V_0(B)] + \lim_{t\to\infty}\sum_{l=0}^{t-1}\tau_1(A)^l\mathbb{E}[\delta_{t-1-l}] = 0 \tag{5.26}$$

It follows that $\mathbb{E}[V_t(B)] \to 0$ since $V_t(B) \geq 0$. Markov's inequality further implies $V_t(B) \overset{p}{\to} 0$. The proof is complete. □

5.8.6 Proof of Lemma V.11

Proof. From (5.4), we evaluate at $x = X^*$ and obtain:

$$p_{i,t+1}(X^*) = a_{i,i}p_{i,t}(X^*)\left(\frac{l_i(Y_{i,t+1}|X^*, A_{i,t})}{\mathcal{Z}_{i,t}(Y_{i,t+1})}\right) + \sum_{j\neq i} a_{i,j}p_{j,t}(X^*)$$

$$= a_{i,i}p_{i,t}(X^*)\left(2\mathbb{P}(Y_{i,t+1}|Z_{i,t})\right) + \sum_{j\neq i} a_{i,j}p_{j,t}(X^*)$$

where $Z_{i,t} = I(X^* \in A_{i,t})$ is the query input to the noisy channel and $\mathbb{P}(Y_{i,t+1}|Z_{i,t})$ models the binary symmetric channel for the ith agent. Taking the logarithm of both sides and using Jensen's inequality, we obtain for each agent i:

$$\log p_{i,t+1}(X^*) \geq \sum_{j=1}^{M} a_{i,j}\log p_{j,t}(X^*) + a_{i,i}\log\left(2\mathbb{P}(Y_{i,t+1}|Z_{i,t})\right)$$

Writing this in vector form with the understanding that the logarithm of a vector is taken component-wise:

$$\log p_{t+1}(X^*) \succeq A\log p_t(X^*) + \mathrm{diag}(A)\log U_{t+1} \tag{5.27}$$

where the vector U_{t+1} is given component-wise by $[U_{t+1}]_i = 2\mathbb{P}(Y_{i,t+1}|Z_{i,t})$. Left-multiplying (5.27) by v^T and using the eigenrelation $v^T = v^T A$, we obtain:

$$v^T\log(p_{t+1}(X^*)) \geq v^T\log(p_t(X^*)) + v^T\mathrm{diag}(A)\log U_{t+1} \tag{5.28}$$

Using induction on (5.28), we obtain:

$$v^T\log(p_t(X^*)) \geq v^T\log(p_0(X^*)) + \sum_{k=0}^{t-1} v^T\mathrm{diag}(A)\log(U_{k+1})$$

This implies by the strong law of large numbers (LLN):

$$\begin{aligned}\lim_{t\to\infty}\frac{1}{t}v^T\log(p_t(X^*)) &\geq \lim_{t\to\infty}\frac{1}{t}v^T\log(p_0(X^*)) + \lim_{t\to\infty}\frac{1}{t}\sum_{k=0}^{t-1}v^T\mathrm{diag}(A)\log(U_{k+1})\\ &= \mathbb{E}\left[\sum_{i=1}^{M}v_i a_{i,i}\log(2\mathbb{P}(Y_i|Z_i))\right]\\ &= \sum_i v_i a_{i,i}\mathbb{E}\left[\log(2\mathbb{P}(Y_i|Z_i))\right]\end{aligned}$$

To finish the proof, note:

$$\begin{aligned}\mathbb{E}\left[\log_2(2\mathbb{P}(Y_i|Z_i))\right] &= \sum_{Z_i}\mathbb{P}(Z_i)\sum_{Y_i}\mathbb{P}(Y_i|Z_i)\log_2(2\mathbb{P}(Y_i|Z_i))\\ &= \sum_{Z_i}\mathbb{P}(Z_i)\left((1-\epsilon_i)\log_2(2(1-\epsilon_i)) + \epsilon_i\log_2(2\epsilon_i)\right)\\ &= 1 - h_B(\epsilon_i) = C(\epsilon_i)\end{aligned}$$

□

5.8.7 Proof of Theorem V.12

Proof. From Theorem V.10 we obtain for each agent i,

$$\mathbb{P}_{i,t}([0,b]) \xrightarrow{p.} \mathbb{P}_\infty(B). \tag{5.29}$$

as $t\to\infty$, where $\mathbb{P}_\infty(B)$ is a common limiting random variable. To finish the proof, we show that $\mathbb{P}_\infty(B)$ is the constant $I(b > X^*)$. Lemma V.11 implies that for t large (as $t\to\infty$):

$$\sum_i v_i a_{i,i}\log(p_{i,t}(X^*)) \gtrsim tK(\boldsymbol{\epsilon})$$

which implies $\sum_i v_i a_{i,i}\log(p_{i,t}(X^*)) \xrightarrow{a.s.} +\infty$. This further implies that there exists an agent $\tilde{i}$ such that $p_{\tilde{i},t}(X^*)\to\infty$ almost surely. For that agent, it follows that $\mathbb{P}_{\tilde{i},t}([0,b])\to I(b > X^*)$ almost surely. From (5.29), it follows that $F_{i,t}(b) = \mathbb{P}_{i,t}([0,b]) \xrightarrow{p.} F_\infty(b) = I(b > X^*)$ for all $i\in\mathcal{N}$.

To conclude the proof, we show the conditional mean estimators $\check{X}_{i,t}$ converge to the correct target location X^* in probability (i.e., consistency). From the definition of the conditional expectation, we obtain:

$$\begin{aligned}\check{X}_{i,t} &= \int_{u=0}^{1} \mathbb{P}_{i,t}((u,1])du \\ &= 1 - \int_{u=0}^{1} F_{i,t}(u)du\end{aligned}$$

where the random variables $F_{i,t}(u)$ are uniformly bounded in $[0,1]$. To finish the proof it suffices to show

$$\int_{u=0}^{1} F_{i,t}(u)du \xrightarrow{p.} \int_{u=0}^{1} F_{\infty}(u)du$$

since $\int_{u=0}^{1} F_{\infty}(u)du = 1 - X^*$. This is accomplished by a variant of the dominated convergence theorem, where the limits are taken in probability. We prove this here for completeness. The first part of the theorem implies

$$\limsup_{t\to\infty} |F_{i,t}(u) - F_{\infty}(u)| \overset{p.}{=} 0 \tag{5.30}$$

for each $u \in [0,1]\backslash X^*$. Also, we have with probability 1:

$$|F_{i,t}(u) - F_{\infty}(u)| \leq 2 \tag{5.31}$$

for all $u \in [0,1]\backslash X^*$ and all t. The reverse Fatou lemma along with (5.31) and (5.30) imply:

$$\limsup_{t\to\infty} \int_0^1 |F_{i,t}(u) - F_{\infty}(u)|\, du \leq \int_0^1 \limsup_{t\to\infty} |F_{i,t}(u) - F_{\infty}(u)|\, du \overset{p.}{=} 0$$

Thus, we conclude that:

$$\lim_{t\to\infty} \int_0^1 |F_{i,t}(u) - F_{\infty}(u)|\, du \overset{p.}{=} 0$$

This concludes the proof. □

CHAPTER VI

Conclusion and Future Work

Many modern systems involving inference and/or control have a high dimensional character that makes optimization of such systems challenging. For example, in spatio-temporal data sets with many sensors and/or time points, estimation of the data covariance matrix over the joint space is intractable in sample-starved settings. In multisensor controlled sensing systems with target localization as a task, although in principle large gains can be obtained by asking multiple queries at each time instant, the implementation of optimal query policies becomes highly nontrivial as the number of sensors or dimensionality of target space gets large. In such systems, the target estimate is refined by updating the posterior distribution of the target using the sensors' noisy responses to sequentially designed questions involving the region where the target may lie. In practical settings, due to limited resources and time-varying phenomena, the sensor classifier will make an error with a certain probability. To improve covariance-based classifier performance, covariance estimation accuracy becomes a fundamental issue.

In Chapter II, under the standard i.i.d. Gaussian sample assumption, we derive high dimensional MSE convergence rates for covariance estimation under Kronecker product (KP) covariance model. The novelty is that through a greedy method for

optimizing the nonconvex maximum likelihood problem that arises, the high dimensional convergence rates improve upon the SCM rate and even faster rates can be obtained for sparse Kronecker product factors using ℓ_1 penalization methods. The methodology heavily relies on factorization properties of the Kronecker product under smooth functionals, which makes such a nonconvex optimization approach difficult to generalize to more complex models consisting of sums of KP's.

Chapter III extends the single term KP model to a series of KP terms, which also paves the way to approximating general covariance matrices of low separation rank. In constrast to Chapter II, the methodology here relies on a convex optimization approach and high dimensional MSE convergence rates are obtained. The key to enforcing structure in the solution is a permutation operator (related to the Kronecker factor dimensions) and a nuclear norm penalty. It is shown that for models with low separation rank, the proposed estimator, PRLS, outperforms the standard SCM in high dimensions.

Chapter IV studies the problem of joint controller design for target localization with multiple sensors. This problem arises in centralized collaborative stochastic search. In this setup, the controller (i.e., a fusion center) asks questions on the presence of the target in a given region to each sensor and each sensor/classifier provides a noisy response on the presence of the target. Using tools from stochastic control, the structure of jointly optimal policies is derived, which shows the design of such policies is highly nontrivial and can be expensive for many sensors or high dimensional targets, even if the sensors are conditionally independent. Thus, a sequential bisection policy that is easy to implement is proposed and is shown to obtain the same average performance gain as the jointly optimal scheme. From another point of view, despite the fact that the sequential scheme has access to a more refined filtration,

the joint scheme performs just as well on average. MSE rates are derived that show fast convergence to the target and the theory is extended to the case of unknown error probabilities associated with the sensors. Suprisingly, it is shown that even in the one-dimensional case and one player, under the setup of unknown probabilities, the optimal policy is not the probabilistic bisection policy (after marginalizing out the noise).

Chapter V extends the collaborative stochastic search ideas to scenarios where a central authority is not present. In this context, a set of low complexity controllers asks questions on the presence of the target in a given region to the sensors, and the sensors provides noisy responses to the queries. Unlike in the centralized setting, each query is solely a function of the local belief. Using each response, the local belief of each agent is updated via Bayes' rule and then linearly combined with its neighbors' beliefs from the previous time instant, giving rise to a semi-Bayesian sequential estimation scheme. A question of primary importance is global convergence of the sensors' beliefs under this scheme. It is proven that as the number of iterations grow to infinity, the sequence of beliefs across all sensors in the network converge to a common Dirac measure centered at the true target location, i.e., a consensus of beliefs is achieved and the limit is correct.

Future work on this thesis, related to Kronecker product covariance estimation, may include estimation of a sum of Kronecker product model in the inverse covariance domain, proving positive definiteness of the PRLS estimator for the case $n < pq$ (as this is observed to be true empirically), and studying sums of sparse Kronecker product decompositions for the covariance to further reduce the dimensionality of the covariance.

Future work related to the decentralized collaborative stochastic search may in-

clude rate-of-convergence analysis. This would yield the value of information of collaboration and new information. In addition, it may allow the practitioner to design the network such that the convergence rate is fastest by optimizing over the parameters of the interaction matrix. The characterization of the spread of the error distribution would also yield insight into the rate of convergence. Another related open problem is the almost sure convergence of the target estimates to the true target location. Further extensions may include convergence for vector-valued targets, and convergence for unbounded noise models (e.g., human error models).

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] D. Acemoglu, A. Ozdaglar, and A. Parandeh-Gheibi. Spread of (mis)information in social networks. *Games and Economic Behavior*, 70(2):194–227, 2010.

[2] G. I. Allen and R. Tibshirani. Transposable regularized covariance models with an application to missing data imputation. *The Annals of Applied Statistics*, 4(2):764–790, 2010.

[3] T. C. Aysal, M. E. Yildiz, A. D. Sarwate, and A. Scaglione. Broadcast gossip algorithms for consensus. *IEEE Transactions on Signal Processing*, 57(7), July 2009.

[4] J. Bai and S. Shi. Estimating high dimensional covariance matrices and its applications. *Annals of Economics and Finance*, 12(2):199–215, 2011.

[5] O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516, March 2008.

[6] R. G. Baraniuk. Compressive sensing. *IEEE Signal Processing Magazine*, pages 118–124, July 2007.

[7] Robert G. Bartle and Donald R. Sherbert. *Introduction to Real Analysis*. John Wiley & Sons, 3 edition, 2000.

[8] F. Benezit, A. Dimakis, P. Thiran, and M. Vetterli. Order-optimal consensus through randomized path averaging. *IEEE Transactions on Information Theory*, 56(10):5150–5167, October 2010.

[9] A. Berman and R. J. Plemmons. *Nonnegative matrices in the Mathematical Sciences*. Academic Press, New York, 1979.

[10] D. A. Berry and B. Fristedt. *Bandit Problems: Sequential Allocation of Experiments*. Chapman & Hall, London, 1985.

[11] D. P. Bertsekas and S. E. Shreve. *Stochastic Optimal Control: The Discrete-Time Case*. Athena Scientific, 1996.

[12] G. Beylkin and M. J. Mohlenkamp. Algorithms for numerical analysis in high dimensions. *SIAM Journal on Scientific Computing*, 26(6):2133–2159, 2005.

[13] P. Bickel and E. Levina. Covariance regularization by thresholding. *Annals of Statistics*, 36(6):2577–2604, 2008.

[14] P. Bickel and E. Levina. Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199–227, 2008.

[15] Fetsje Bijma, Jan de Munck, and Rob Heethaar. The spatiotemporal meg covariance matrix modeled as a sum of kronecker products. *NeuroImage*, 27:402–415, 2005.

[16] E. Bonilla, K. M. Chai, and C. Williams. Multi-task gaussian process prediction. *Advances in Neural Information Processing Systems*, pages 153–160, 2008.

[17] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah. Randomized Gossip Algorithms. *IEEE Transactions on Information Theory*, 52(6):2508–2530, June 2006.

[18] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[19] J. P. Burg, D. G. Luenberger, and D. L. Wenger. Estimation of structured covariance matrices. *Proceedings of the IEEE*, 70(9), September 1982.

[20] M. V. Burnashev and K. Sh. Zigangirov. An interval estimation problem for controlled observations. *Problems in Information Transmission*, 10:223–231, 1974.

[21] J-F. Cai, E. J. Candes, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal of Optimization*, 20(4):1956–1982, 2010.

[22] J-F. Cai and S. Osher. Fast singular value thresholding without singular value decomposition. Technical report, UCLA, 2010.

[23] T. Cai, C. Zhang, and H. Zhou. Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, 38(4):2118–2144, 2010.

[24] E. Candès and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98:925–936, 2010.

[25] E. Candès and B. Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9(6):717–772, 2009.

[26] E. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

[27] E. Candès and T. Tao. The power of convex relaxation: near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

[28] E. Candès and M. B. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, pages 21–30, March 2008.

[29] R. Castro. *Active Learning and Adaptive Sampling for Non-parametric Inference*. PhD thesis, Rice University, August 2007.

[30] R. Castro and R. Nowak. Active learning and sampling. In *Foundations and Applications of Sensor Management*. Springer, 2007.

[31] R. Castro and R. D. Nowak. Upper and lower bounds for active learning. In *44th Annual Allerton Conference on Communication, Control and Computing*, 2006.

[32] M. Chen, S. C. Liew, Z. Shao, and C. Kai. Markov approximation for combinatorial network optimization. *IEEE Transactions on Information Theory*, 59(10), October 2013.

[33] Y. Chen, A. Wiesel, and A. Hero. Robust shrinkage estimation of high dimensional covariance matrices. *IEEE Transactions on Signal Processing*, 59(9):4097–4107, September 2011.

[34] S. Choi, D. Gale, and S. Kariv. Sequential equilibrium in monotone games: A theory-based analysis of experimental data. *Journal of Economic Theory*, 143(1):302–330, 2008.

[35] T. D. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 2006.

[36] N. Cressie. *Statistics for Spatial Data*. Wiley, New York, 1993.

[37] G. Cybenko. Dynamic load balancing for distributed memory multiprocessors. *Journal of Parallel and Distributed Computing*, 7(2):279–301, 1989.

[38] A. Dawid. Some matrix-variate distribution theory: notational considerations and a bayesian application. *Biometrika*, 68:265–274, 1981.

[39] X. de Luna and M. Genton. Predictive spatio-temporal models for spatially sparse environmental data. *Statistica Sinica*, 15:547–568, 2005.

[40] J. C. de Munck, F. Bijma, P. Gaura, C. A. Sieluzycki, M. I. Branco, and R. M. Heethaar. A maximum-likelihood estimator for trial-to-trial variations in noisy meg/eeg data sets. *IEEE Transactions on Biomedical Engineering*, 51(12), 2004.

[41] J. C. de Munck, H. M. Huizenga, L. J. Waldorp, and R. M. Heethaar. Estimating stationary dipoles from meg/eeg data contaminated with spatially and temporally correlated background noise. *IEEE Transactions on Signal Processing*, 50(7), July 2002.

[42] M. H. DeGroot. *Optimal Statistical Decisions*. McGraw Hill, New York, 1970.

[43] M. H. DeGroot. Reaching a consensus. *J. Amer. Statist. Assoc.*, 69(345):118–121, 1974.

[44] G. Derado, F. D. Bowman, and C. D. Kilts. Modeling the spatial and temporal dependence in fmri data. *Biometrics*, 66(3):949–957, September 2010.

[45] A. Dimakis, S. Kar, J. M. F. Moura, M. G. Rabbat, and A. Scaglione. Gossip algorithms for distributed signal processing. *Proceedings of the IEEE*, 98(11), November 2010.

[46] A. Dimakis, A. Sarwate, and M. Wainwright. Geographic gossip: Efficient averaging for sensor networks. *IEEE Transactions on Signal Processing*, 56(3):1205–1216, March 2006.

[47] D. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[48] R. Durrett. *Probability: Theory and Examples*. Duxbury Press, Belmont, CA, 3 edition, 2005.

[49] P. Dutilleul. The mle algorithm for the matrix normal distribution. *J. Statist. Comput. Simul.*, 64:105–123, 1999.

[50] J. Fan, Y. Fan, and J. Lv. High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147(1):1348–1360, 2008.

[51] G. Fitzmaurice, N. Laird, and J. Ware. *Applied longitudinal analysis*. Wiley-Interscience, 2004.

[52] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[53] D. Gale and S. Kariv. Bayesian learning in social networks. *Games and Economic Behavior*, 45(2):329–346, 2003.

[54] M. G. Genton. Separable approximations of space-time covariance matrices. *Environmetrics*, 18:681–695, 2007.

[55] A. Gersho and R. M. Gray. *Vector quantization and Signal Compression*. Kluwer Academic Press/Springer, 1992.

[56] J. C. Gittins and D. M. Jones. A dynamic allocation index for the sequential design of experiments. In *Progress in Statistics*, pages 241–266. Amsterdam, North-Holland, 1974.

[57] T. Gneiting. Nonseparable, stationary covariance functions for space-time data. *Journal of the American Statistical Association (JASA)*, 97(458):590–600, 2002.

[58] B. Golub and M. Jackson. Naive learning in social networks and the wisdom of crowds. *Amer. Econ. J.: Microecon.*, 2(1):112–149, 2010.

[59] G. H. Golub and C. Van Loan. *Matrix Computations*. JHU Press, 1996.

[60] A. K. Gupta and D. K. Nagar. *Matrix Variate Distributions*. Chapman Hill, 1999.

[61] N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, May 2011.

[62] J. Haslett and A. E. Raftery. Space-time modeling with long-memory dependence: assessing ireland's wind power resource. *Applied Statistics*, 38(1):1–50, 1989.

[63] A. Hero and B. Rajaratnam. Hub discovery in partial correlation graphs. *IEEE Transactions on Information Theory*, 58(9):6064–6078, September 2012.

[64] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 1 edition, 1990.

[65] M. Horstein. Sequential transmission using noiseless feedback. *IEEE Transactions on Information Theory*, pages 136–143, July 1963.

[66] C.-J. Hsieh, M. A. Sustik, I. S. Dhillon, and P. Ravikumar. Sparse inverse covariance matrix estimation using quadratic approximation. *Advances in Neural Information Processing Systems*, 24, 2011.

[67] J. Huang, N. Liu, M. Pourahmadi, and L. Liu. Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93(1), 2006.

[68] I. C. F. Ipsen and T. M. Selee. Ergodicity coefficients defined by vector norms. *SIAM J. Matrix Anal. Appl.*, 32(1):153–200, 2011.

[69] A. Jadbabaie, J. Lin, and A. Morse. Coordination of groups of mobile autonomous agents using nearest neighbor rules. *IEEE Transactions on Automatic Control*, 48(6):988–1001, June 2003.

[70] A. Jadbabaie, P. Molavi, A. Sandroni, and A. Tahbaz-Salehi. Non-bayesian social learning. *Games and Economic Behavior*, 76:210–225, 2012.

[71] K. G. Jamieson and R. D. Nowak. Active ranking using pairwise comparisons. In J. Shawe-Taylor, R.S. Zemel, P. Bartlett, F.C.N. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2240–2248. MIT Press, 2011.

[72] K. G. Jamieson, R. D. Nowak, and B. Recht. Query complexity of derivative-free optimization. *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[73] B. Jedynak, P. I. Frazier, and R. Sznitman. Twenty questions with noise: Bayes optimal policies for entropy loss. *Journal of Applied Probability*, 49:114–136, 2012.

[74] I. Johnstone and A. Lu. On consistency and sparsity for principal component analysis in high dimensions. *Journal of the American Statstical Association*, 104(486):682–693, 2009.

[75] S. C. Jun, S. M. Plis, D. M. Ranken, and D. M. Schmidt. Spatiotemporal noise covariance estimation from limited empirical magnetoencephalographic data. *Physics in Medicine and Biology*, 51:5549–5564, 2006.

[76] E. Kalnay, M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, M. Chelliah, W. Ebisuzaki, W.Higgins, J. Janowiak, K. C. Mo, C. Ropelewski, J. Wang, A. Leetmaa, R. Reynolds, Roy Jenne, and Dennis Joseph. The ncep/ncar 40-year reanalysis project. *Bulletin of the American Meteorological Society*, 77(3):437–471, 1996.

[77] S. Kar and J. M. F. Moura. Covergence rate analysis of distributed gossip (linear parameter) estimation: Fundamental limits and tradeoffs. *IEEE Journal of Selected Topics in Signal Processing*, 5(4), August 2011.

[78] H. J. Kushner and G. G. Yin. *Stochastic Approximation adn Recursive Algorithms and Applications*. Springer, 2003.

[79] Clifford Lam and Jianqing Fan. Sparsistency and rates of convergence in large covariance matrices estimation. *The Annals of Statistics*, 37:4254–4278, 2009.

[80] Steffen L. Lauritzen. *Graphical Models*. Oxford University Press US, first edition, 1996.

[81] M. Ledoux and M. Talagrand. *Probability in Banach spaces. Isoperimetry and processes.* Springer-Verlag, Berlin, Heidelberg, New York, 1991.

[82] Seunggeun Lee, Fei Zou, and Fred A Wright. Convergence and prediction of principal component scores in high-dimensional settings. *The Annals of Statistics*, 38(6):3605–3629, 2010.

[83] Chenlei Leng and Cheng Yong Tang. Sparse matrix graphical models. *Journal of the American Statistical Association*, 107:1187–1200, October 2012.

[84] Charles Van Loan and Nikos Pitsianis. Approximation with kronecker products. In *Linear Algebra for Large Scale and Real Time Applications*, pages 293–314. Kluwer Publications, 1993.

[85] K. Lounici. High-dimensional covariance matrix estimation with missing observations. *arXiv:1201.2577v5*, May 2012.

[86] N. Lu and D. Zimmerman. On likelihood-based inference for a separable covariance matrix. Technical report, Statistics and Actuarial Science Dept., Univ. of Iowa, Iowa City, IA, 2004.

[87] N. Lu and D. Zimmerman. The likelihood ratio test for a separable covariance matrix. *Statistics and Probability Letters*, 73(5):449–457, May 2005.

[88] Nicolai Meinshausen and Peter Buhlmann. High dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, 34:1436–1462, 2006.

[89] L. Moreau. Stability of multiagent systems with time-dependent communication links. *IEEE Transactions on Automatic Control*, 50(2):169–182, February 2005.

[90] E. Mossel and O. Tamuz. Efficient Bayesian learning in social networks with Gaussian estimators. *Preprint, arXiv: 1002.0747*, 2010.

[91] M. Nokleby, W. Bajwa, R. Calderbank, and B. Aazhang. Toward resource-optimal consensus over the wireless medium. *IEEE Journal of Selected Topics in Signal Processing*, 7(2), April 2013.

[92] R. D. Nowak. The geometry of generalized binary search. *IEEE Transactions on Information Theory*, 57(12), December 2011.

[93] R. J. Owen. A bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, 70:351–356, 1975.

[94] A. Papoulis and S. U. Pillai. *Probability, Random Variables and Stochastic Processes*. Mc Graw Hill, 2002.

[95] D. Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17:1617–1642, 2007.

[96] M. I. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley-Interscience, 2005.

[97] N Raj Rao, James A Mingo, Roland Speicher, and Alan Edelman. Statistical eigen-inference from large wishart matrices. *The Annals of Statistics*, pages 2850–2885, 2008.

[98] H. Rauhut, K. Schnass, and P. Vandergheynst. Compressed sensing and redundant dictionaries. *IEEE Transactions on Information Theory*, May 2008.

[99] P. Ravikumar, M. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Advances in Neural Information Processing Systems*, 2008.

[100] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22(3):400–407, 1951.

[101] R. T. Rockafellar and R. J-B. Wets. *Variational Analysis*. Springer, 1998.

[102] A. Rohde and A. B. Tsybakov. Estimation of high-dimensional low-rank matrices. *Annals of Statistics*, 39(2):887–930, 2011.

[103] A. Rothman, P. Bickel, E. Levina, and J. Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

[104] A. Rothman, E. Levina, and J. Zhu. Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104(485):177–186, 2009.

[105] A. Rucci, S. Tebaldini, and F. Rocca. Skp-shrinkage estimator for sar multi-baselines applications. In *Proceedings of IEEE Radar Conference*, 2010.

[106] V. Saligrama, M. Alanyali, and O. Savas. Distributed detection in sensor networks with packet loss and finite capacity links. *IEEE Transactions on Signal Processing*, 54(11):4118–4132, November 2006.

[107] E. Seneta. *Non-negative Matrices and Markov Chains*. New York: Springer, 2 edition, 1981.

[108] B. Settles. Active learning literature survey. Technical report, Computer Sciences Technical Report 1648, University of Wisconsin-Madison, 2009.

[109] M. Stein. Space-time covariance functions. *Journal of the American Statistical Association (JASA)*, 100:310–321, 2005.

[110] S. Tebaldini. Algebraic synthesis of forest scenarios from multibaseline polinsar data. *IEEE Transactions on Geoscience and Remote Sensing*, 47(12), December 2009.

[111] T. Tsiligkaridis and A. O. Hero. Sparse covariance estimation under Kronecker product structure. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3633–3636, March 2012.

[112] T. Tsiligkaridis and A. O. Hero. Covariance Estimation in High Dimensions via Kronecker Product Expansions. *arXiv: 1302.2686*, February 2013.

[113] T. Tsiligkaridis and A. O. Hero. Covariance Estimation in High Dimensions via Kronecker Product Expansions. *IEEE Transactions on Signal Processing*, 61(21), November 2013.

[114] T. Tsiligkaridis, A. O. Hero, and S. Zhou. Convergence Properties of Kronecker Graphical Lasso algorithms. *arXiv:1204.0585*, July 2012.

[115] T. Tsiligkaridis, A. O. Hero, and S. Zhou. Kronecker Graphical Lasso. In *Proceedings of IEEE Statistical Signal Processing (SSP) Workshop*, pages 884–887, August 2012.

[116] T. Tsiligkaridis, A. O. Hero, and S. Zhou. On Convergence of Kronecker Graphical Lasso Algorithms. *IEEE Transactions on Signal Processing*, 61(7):1743–1755, April 2013.

[117] T. Tsiligkaridis, B. M. Sadler, and A. O. Hero. Blind Collaborative 20 Questions for Target Localization. In *Proceedings of IEEE GlobalSIP - Symposium on Controlled Sensing For Inference: Applications, Theory and Algorithms*, December 2013.

[118] T. Tsiligkaridis, B. M. Sadler, and A. O. Hero. Collaborative 20 Questions for Target Localization. *Preprint, arXiv: 1306.1922*, August 2013.

[119] T. Tsiligkaridis, B. M. Sadler, and A. O. Hero. A Collaborative 20 Questions model for target search with human-machine interaction. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013.

[120] T. Tsiligkaridis, B. M. Sadler, and A. O. Hero. On Decentralized Eestimation via Active Queries. *Preprint*, December 2013.

[121] J. Tsitsiklis. *Problems in decentralized decision making and computation*. PhD thesis, Massachussets Institute of Technology, Cambridge, MA, November 1984.

[122] J. Tsitsiklis, D. Bertsekas, and M. Athans. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE Transactions on Automatic Control*, 31(9):803–812, September 1986.

[123] E. Tyrtyshnikov. Kronecker-product approximations for some function-related matrices. *Linear Algebra and its Applications*, 379:423–437, 2004.

[124] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv:1011.3027v7*, November 2011.

[125] R. Waeber, P. I. Frazier, and S. G. Henderson. A bayesian approach to stochastic root finding. In *Winter Simulation Conference*, 2011.

[126] R. Waeber, P. I. Frazier, and S. G. Henderson. Bisection search with noisy responses. *SIAM Journal of Control and Optimization*, 53(3):2261–2279, 2013.

[127] H. Wainer. *Computerized Adaptive Testing: A Primer*. Routledge, 2 edition, 2000.

[128] X. Wang, J. O. Berger, and D. S. Burdick. Bayesian analysis of dynamic item response models in educational testing. *Annals of Applied Statistics*, 7(1):126–153, 2013.

[129] G. A. Watson. Characterization of the subdifferential of some matrix norms. *Linear Algebra and Applications*, 170:33–45, 1992.

[130] K. Werner, M. Jansson, and P. Stoica. On estimation of covariance matrices with Kronecker product structure. *IEEE Transactions on Signal Processing*, 56(2), February 2008.

[131] Karl Werner and Magnus Jansson. Estimation of kronecker structured channel covariances using training data. In *Proceedings of EUSIPCO*, 2007.

[132] G. B. Wetherill and K. D. Glazebrook. *Sequential Methods in Statistics, Monographs on Statistics and Applied Probability*. Chapman & Hall, London, third edition, 1986.

[133] J. Xie and P. M. Bentler. Covariance structure models for gene expression microarray data. *Structural Equation Modeling: A Multidisciplinary Journal*, 10(4):556–582, 2003.

[134] J. Yin and H. Li. Model selection and estimation in the matrix normal graphical model. *Journal of Multivariate Analysis*, 107:119–140, 2012.

[135] K. Yu, J. Lafferty, S. Zhu, and Y. Gong. Large-scale collaborative prediction using a non-parametric random effects model. *ICML*, pages 1185–1192, 2009.

[136] M. Yuan and Y. Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94:19–35, 2007.

[137] Y. Zhang and J. Schneider. Learning multiple tasks with a sparse matrix-normal penalty. *Advances in Neural Information Processing Systems*, 23:2550–2558, 2010.

[138] Y. Zhang, W. Xu, and J. Callan. Exploration and exploitation in adaptive filtering based on bayesian active learning. In *Proceedings of the Twentieath International Conference on Machine Learning (ICML)*, pages 896–903, 2003.

[139] S. Zhou, J. Lafferty, and L. Wasserman. Time varying undirected graphs. *Journal of Machine Learning Research*, 80:295–319, 2010.

[140] S. Zhou, P. Rutimann, M. Xu, and P. Buhlmann. High-dimensional covariance estimation based on gaussian graphical models. *Journal of Machine Learning Research*, pages 2975–3026, October 2011.