Stochastic optimization methods for structure learning in Gaussian graphical models and the general log-determinant optimization

Wu, Songwei

2020

Wu, S. (2020). Stochastic optimization methods for structure learning in Gaussian graphical models and the general log-determinant optimization. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/142033>

<https://doi.org/10.32657/10356/142033>

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

Downloaded on 26 Apr 2023 21:38:33 SGT

Stochastic Optimization Methods for Structure Learning in Gaussian Graphical Models and the General Log-determinant Optimization

Wu Songwei

School of Electrical & Electronic Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

2020

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

June 12, 2020

.....

Date

.....

Wu Songwei

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

June 12, 2020

.....

Date

.....

Prof. Justin Dauwels

Authorship Attribution Statement

This thesis contains material from 2 papers published in the following peer-reviewed journal(s) / from papers accepted at conferences in which I am listed as an author.

Chapter 3 is published as Songwei Wu, Hang Yu, and Justin Dauwels. "Efficient Stochastic Subgradient Descent Algorithms for High-dimensional Semi-sparse Graphical Model Selection." ICASSP 2019-2019 IEEE *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.

and

Songwei Wu, Hang Yu, Justin Dauwels, GIBSON: Gibbs Sampling-based Stochastic Subgradient Descent Algorithms for Inverse Covariance Estimation, IEEE *Transactions on Signal Processing* (under peer review)

The contributions of the co-authors are as follows:

- Three authors formulated the initial objective during group meetings.
- I prepared the manuscript drafts. The manuscript was revised by Dr. Hang Yu and Prof. Justin Dauwels.
- I prepared the MATLAB code. Dr. Hang Yu helped with debugging and optimization.
- I co-designed the study with Dr. Hang Yu and performed all laboratory work at the School of Electrical and Electronic Engineering.

Chapter 4 is published as Songwei Wu, Hang Yu, and Justin Dauwels. "Primal-Dual Stochastic Subgradient Method for Log-determinant Optimization" ICASSP 2020-2020 IEEE *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020. (submitted)

and

Songwei Wu, Hang Yu, Justin Dauwels, "Primal-Dual Stochastic Subgradient Method for Log-determinant Optimization", IEEE *Transactions on Signal Processing* (drafted)

The contributions of the co-authors are as follows:

- Three authors formulated the initial objective during group meetings.
- I wrote the drafts of the manuscript. Prof. Justin Dauwels provided comments and I revised it accordingly.
- I designed the study of this work and prepared the MATLAB code accordingly. Dr. Hang Yu optimized the code.
- All laboratory work was conducted by me at the School of Electrical and Electronic Engineering.

June 12, 2020

.....

Date

.....

Wu Songwei

Acknowledgements

First of all, I am greatly indebted to my supervisor, Prof. Justin Dauwels, for his detailed instructions, professional guidance, and ceaseless encouragement throughout my Ph.D study. I began my Ph.D research right after finishing my undergraduate study. For lack of research experience, there were lots of challenges at the beginning. I am more than grateful for his support to attend the machine learning summer school in the first year. This trip substantially broadened my learning experience, expanded my research horizon, and developed my enthusiasm for my Ph.D study. Justin continuously enlightened my research with positive encouragement and beneficial pressure, which motivated me to overcome lots of difficulties. This dissertation would never have been possible without his invaluable support. Moreover, I have learned perseverance in research, optimism about life, and strategic mindset on work from his personality. I also thank him for the chance to meet and interact with other world-class researchers.

I want to express my gratitude to Hang Yu as my senior, my colleague, and my friend. His brilliance inspired many insightful ideas in my research. The initial idea of improving the efficiency in structure learning in Gaussian graphical models originated from him. I have learned various aspects of research from him including coding, writing, presenting, etc. I also benefitted significantly from his resourceful suggestions and experience through daily discussions. He has been particularly helpful in my Ph.D study and my life in Singapore.

I have also benefitted a lot from favorable interactions with my other colleagues and lab mates - Wang Mengmeng, Wang Yuanzhe, Yang Chule, Jiang Wentao, Zeng Yijie, Jiang Xiaoyue, Sun Chao, Yue Yufeng, Zhao Wei, Guo Fanghong, Yang Shuai, Tang Wenchang, Zhou Xiaosong, Zou Ying, Yang Huan, Jing Jin, Wang Xiaoyun. I would to thank them for the pleasant time spent together. Especially, I thank Wang Mengmeng for templates of my QE and thesis report, valuable advice on my research, and constructive suggestions on life.

I would like to take the opportunity to thank my close friends: Wang Peiyu, Li Li, Li Yuxin, Wang Zhongjiong, Wu Xuan, Dong Wenhao, Yang Jing, Chen Yating, Li Chen, Huang Yin, Zhang Chuqing, Zeng Junru, Long Huiwu, Chen Cheng, Huang Qiang, Wu Xuan, Chen Yang, Yu Mengni, Zhang Yin, Zhao Han. Thank them for the delightful time and enjoyable treatment. In particular, I am more than lucky to be a friend of Wang Peiyu. I would always remember the time when we discussed mathematics, traveled abroad, and shared exotic experiences with each other. Moreover, I deeply appreciate the friendship with Li Li. She has continuously been a wise, supportive, and trusted friend to me. I also appreciate the delightful and illuminating four years I spent with Li Yuxin, Wang Zhongjiong, Wu Xuan, my room mates in Sun-Yat Sen University. The experiences with them made me a better person.

Special thanks go to my girlfriend, Yang Huan. She has gave me endless, unconditional, and warming love. The time we spent together and will share in the future constitutes the most precious memory in my life. I thank her for her tolerance for my childishness, my stubbornness, and all my other shortcomings. May it be that we can share both bitter and sweet life in the future side by side, hand in hand, heart to heart.

Finally, I sincerely thank my brother and my parents for their encouragement, support and love. They bring happiness into my life, motivate my determination, and provide me unconditional shelter all the time.

Abstract

Graphical models compactly represent the most significant interactions of multivariate probability distributions, provide an efficient inference framework to answer challenging statistical queries, and incorporate both expert knowledge with data to extract information from complex systems. When the graphical model is assumed to be Gaussian, the resulting model features attractive properties and appears frequently in cutting-edge applications. In the application of Gaussian graphical models, it is fundamental and challenging to learn the graph structure from given multivariate data. The time complexity of existing methods is subject to the cumbersome operations such as matrix inversion, full spectral decomposition, and Cholesky decompositions, which generally take $\mathcal{O}(N^3)$ arithmetic operations. This computational bottleneck severely impedes the scalability of those methods into problems at large scale. Moreover, structure learning in Gaussian graphical model is a specialized case of the general log-determinant optimization, which entails general matrix constraints and covers much broader fields of applications. The bottleneck of the cubic time complexity also occurs in existing methods for the general log-determinant optimization. In this thesis, we propose a novel method to learn the structure of the Gaussian graphical model with the aid of stochastic approximation so that the time complexity can be reduced from $\mathcal{O}(N^3)$ to $\mathcal{O}(N^2)$. In addition, we generalize the proposed method within the framework of primal-dual optimization so that we can settle the general log-determinant optimization problem with similar efficiency.

We first address structure learning in Gaussian graphical models. Given multivari-

ate data, the problem can be formulated as estimating a sparse precision matrix, because of the equivalence between the graph structure of the graphical model, the conditional independence relationship of the underlying Gaussian distribution, and the zero pattern of the precision matrix. The resulting objective function leads to a non-smooth positive definite convex programming problem, which contains a complex log-determinant term. Generally, there are two important questions for all methods: how to update the precision matrix and how to ensure positive definiteness of the estimate. To address the first question, we construct an efficient stochastic approximation of the matrix inverse term, obtain the stochastic realization of the subgradient at much reduced cost, and iteratively update the precision matrix following the opposite direction of the subgradient. The resulting time complexity of the stochastic subgradient is only $\mathcal{O}(N^2)$. As for the second challenge, we repeatedly evaluate the smallest eigenvalue of the updated precision matrix by the implicitly restarted Lanczos method, search for a proper step size, and render the estimate in the positive definite cone. The implicitly restarted Lanczos method takes only $\mathcal{O}(N^2)$ arithmetic operations to estimate the extreme eigenvalue of a N -dimensional matrix. Furthermore, we propose an adaptive step size which mitigates the influence of noise, avoids overshooting, and accelerates convergence. We theoretically and empirically analyze the resulting method. Both theoretical and numerical results demonstrate the improved scalability, certain convergence, and favorable properties of the proposed method.

The general log-determinant problem imposes general linear matrix constraints on the objective function. The proposed method for structure learning in Gaussian graphical models can only solve the unconstrained optimization problem. Therefore, the matrix constraints render the proposed method inapplicable to the general log-determinant problem. In order to generalize the proposed method, we analyze the augmented Lagrangian function, establish the equivalence between the primal-dual pair and the saddle point, and convert the optimization problem into searching for a saddle point of the augmented Lagrangian function. Following this analysis,

the generalized method extends the Uzawa iterations to stochastic optimization, incorporates the techniques employed in structure learning in Gaussian graphical models, and obtains the optimum at the cost of only $\mathcal{O}(N^2)$ arithmetic operations. Furthermore, we investigate the theoretical performance and establish the guarantee of convergence. Besides the theoretical analysis, we also numerically analyze the performance on both synthetic and real data. Through numerical experiments, we compare various values of important parameters, suggest informative guidelines on parameter settings, validate the main theorem, and provide evidence for the reduction of the time complexity from cubic to quadratic.

In conclusion, we propose a novel method for structure learning in Gaussian graphical models with the aid of stochastic optimization and further extend the method into primal-dual optimization to settle the general log-determinant optimization problem. Such methods overcome the computational bottleneck of existing methods and improve the scalability for high-dimensional problems.

List of Abbreviations

iff	if and only if
i.i.d.	independent and identically distributed
a.k.a.	also known as
ES	Embedded Subgraphs
BP	Belief Propagation
SPD	Symmetric Positive Definite
MAP	Maximum A Posterior
GGM	Gaussian Graphical Model
GMRF	Gaussian Markov random fields
LASSO	Least Absolute Shrinkage and Selection Operator
GLASSO	Graphical LASSO
DP-GLASSO	Dual update in the Primal GLASSO
SINCO	a greedy coordinate ascent method for Sparse INverse COvari- ance
ANS	Adaptive Nesterov's Method
ISTA	Iterative Shrinkage Thresholding Algorithm
IIPM	Inexact primal-dual path-following Interior-Point Method
ADM	Alternating Direction Method
PGM	Projection Gradient Method
ASPG	Adaptive Spectral Projection Gradient
QUIC	QUadratic approximation of Inverse Covariance matrices
G-ISTA	Graphical ISTA

Logdet PPA	Log-determinant PPA
BIG&QUIC	BIG QUadratic approximation of Inverse Covariance matrices
Newton-CG	Newton Conjugate Gradient
PDSGD-ad	adaptive Primal-Dual Stochastic subGradient Descent
GIBSON	GIBbs-sampling-based StOchastic subgradient desceNt
BCD-IC	Block-Coordinate Descent approach for large-scale sparse Inverse Covariance estimation
StARS	Stability Approach to Regularization Selection
SGD	Stochastic subGradient Descent
PPA	Proximal Point Algorithm
SAA	Sample Average Approximation
SA	Stochastic Approximation

List of Symbols

$\mathbb{C}$ or $\mathcal{C}$	Name of sets (calligraphic uppercase letters) if without further specification
$ \mathcal{C} $	The cardinality of the set $\mathcal{C}$
$\mathbb{R}$	The set of real numbers
$\mathbb{R}_+$	The set of non-negative real numbers
$\mathbb{N}$	The set of natural numbers
$\mathbb{R}^n$	The set of n -dimensional real-valued column vectors
$\mathbb{R}^{n \times n}$	The set of n -dimensional real-valued matrices
$\mathbb{S}^n$	The set of n -dimensional real-valued symmetric matrices
$\mathbb{S}_+^n$	The set of n -dimensional real-valued symmetric positive definite matrices
$\mathcal{A}(\cdot)$	Linear matrix operator
$\mathcal{A}^T(\cdot)$	The adjoint of the linear matrix operator $\mathcal{A}(\cdot)$
$\mathcal{O}(\cdot)$	Notation for time complexity
$\mathcal{N}(\cdot, \cdot)$	Gaussian distribution
$\mathcal{W}_n(\cdot, \cdot)$	n -dimensional Wishart distribution
$\mathbb{P}(\cdot)$	Projection operator
$\mathbb{E}[\cdot]$	Expectation operator
$\mathbb{V}[\cdot]$	Variance operator
a, α	Scalar (lowercase letters or Greek letters) if without further specification
$[a]_+$	$\max[0, a]$

$ a $	The absolute value of scalar a
$\lfloor a \rfloor$	$\max\{b b \text{ is an integer and } b < a\}$
$\mathbf{a}$	Column vector in $\mathbb{R}^n$ (bold lowercase letters)
a_i	The i -th element in $\mathbf{a}$
$\mathbf{a}_i$ or $\mathbf{a}^{(i)}$	Column vector indexed for some purpose
I or I_n	n -dimensional identity matrix
A	Matrix (uppercase letters) if without further specification
$A^T, \mathbf{a}^T$	The transpose of A or $\mathbf{a}$
A^{-1} or $\{A\}^{-1}$	The inverse of A (if A is nonsingular)
$A_{:,j}$	The j -th column of A
$A_{i:j,m:n}$	The submatrix of A by extracting rows i to j and columns m to n
A_i or $A^{(i)}$	Matrix indexed for some purpose
A^i	Matrix indexed for some purpose or the i -th power of a square matrix
$A_{i,j}$	The element of A in i -th row and j -th column
$\langle \cdot, \cdot \rangle$	Inner product
$\ \cdot \ $	Norm (subscript if any denotes respective norm)
f, g, h, p	Function names
$\ell(\cdot)$	The function name of log-likelihood functions
$L(\cdot)$	The function name of Lagrangian functions
$\mathcal{L}(\cdot)$	The function name of augmented Lagrangian functions
∇f	The gradient of f (if f is differentiable)
$\partial f(x)$	The set of subgradients of f at the point x (if f is convex)
$\text{dom} f$	The domain of f

Contents

Acknowledgements	iii
Abstract	vi
List of Abbreviations	ix
List of Symbols	xi
List of Contents	xiv
List of Figures	xvi
List of Tables	xvii
1 Introduction	1
1.1 Motivation	1
1.2 Related Works	5
1.2.1 Structure Learning of GGM	5

1.2.2	General Log-determinant Optimization	12
1.3	Thesis Organization and Overview of Contributions	16
2	Preliminaries	19
2.1	Gaussian Graphical Model	19
2.1.1	Graph Theory	20
2.1.2	Graphical Model	21
2.1.3	Gaussian Graphical Model	22
2.1.4	Learning Gaussian Graphical Models	24
2.2	Convex Optimization	26
2.2.1	Convex Sets and Convex Functions	26
2.2.2	Hilbert Space and Important Inequalities	28
2.2.3	Basic Subgradient Method for Unconstrained Optimization . .	30
2.2.4	Lagrangian Duality Theory	33
2.3	Stochastic Optimization	35
3	Stochastic Optimization for Structure Learning in Gaussian Graphical Models	37
3.1	Stochastic Subgradient Descent	37
3.1.1	Updating X Using Stochastic Subgradient Descent	38
3.1.2	Guarantee of the Positive Definiteness of X	40

3.2	Theoretical Results	43
3.2.1	Convergence Analysis	43
3.3	Implementation	52
3.4	Numerical Results	54
3.4.1	Synthetic Dataset	55
3.4.2	Real Data	60
3.5	Conclusion	64
4	Stochastic Optimization for the General Log-determinant Opti- mization	67
4.1	Primal-Dual Stochastic Subgradient Method	67
4.1.1	Primal-Dual Subgradient Method	68
4.1.2	Stochastic Subgradient	71
4.1.3	Guarantee of the Positive Definiteness of X	73
4.2	Theoretical Results	73
4.2.1	Assumptions and Preliminaries	73
4.2.2	Bounds on the Stochastic Subgradient	75
4.2.3	Definitions and Lemmas	81
4.2.4	Convergence Analysis	85
4.3	Implementation	90
4.3.1	Initialization	90

4.3.2	Adaptive Step Size	91
4.3.3	Termination Criterion	92
4.4	Numerical Results	93
4.4.1	Synthetic Data	94
4.4.2	Real Data	102
4.5	Conclusion	104
5	Conclusions and Future Work	105
5.1	Summary of Contributions	105
5.2	Recommendations for Future Work	108
5.2.1	Log-determinant Optimization with Nonlinear Constraints . .	108
5.2.2	Alternative Sampling Method	111
5.2.3	Problem Partition	112
A	An Alternative Convergence Analysis of Algorithm 3	115
	Author's Publications	123
	Bibliography	134

List of Figures

2.1	The brown nodes, blue nodes, and red nodes denote $\mathcal{P}, \mathcal{Q}, \mathcal{S} \subset \mathcal{V}$ respectively. After removing the red nodes and red edges, there would be no path between any arbitrary brown node and blue node. In such way, we say $\mathcal{S}$ separates $\mathcal{P}$ and $\mathcal{Q}$	21
3.1	Time comparison across different μ for $N=10,000$ and $\tau = 1\%$	57
3.2	Time comparison across different density, experiments are executed on $N=10,000$ while μ is chosen to minimize the F_1 -score.	59
3.3	Computational time as a function of dimension N for different μ values.	61
3.4	Computational time as a function of dimension N for the best μ value.	62
3.5	Randomly selected precision matrix entries plot of the three algorithms. There are 5 non-zero elements and 5 zero elements in the global optimal solution. We show $N = 1000$ and $N=10,000$ separately. The graph density is 1% and the penalty parameter is 0.06. . .	63

-
- 4.1 The dimension is $N = 500$ and the number of constraints is $M = 55,917$. Under the same parameter setting, we test various number K of Gibbs samples. The objective value $f_0(X)$, the feasibility $\|\mathcal{A}(X) - \mathbf{b}\|/M$, and the mean square error $\|X - X^*\|/N^2$ converge. More details are listed in Table 4.1. 96
- 4.2 The dimension $N = 500$, the number of constraints $M = 55,917$, and sample size equals 400. The synthetic data set is the same as in Figure 4.1. Experiments are conducted for various penalty parameters $\rho \in [10^{-4}, 10^4]$. When ρ is chosen within a proper range, the estimate is convergent. 99
- 4.3 Comparison of convergence performance in terms of the objective value, the feasibility, and the MSE. 101
- 4.4 The dimension increases from 500 to 10,000 and $K = 400$. For varying dimensions, the feasibility $\|\mathcal{A}(X) - \mathbf{b}\|/M$ and the mean square error $\|X - X^*\|/N^2$ of the estimate of PDSGD-ad are presented in Table 4.3. 103

List of Tables

1.1	The time complexity for updating X or W and for guaranteeing the positive definiteness of X of existing methods [1–16].	8
3.1	MSE between GIBSON estimates and results from QUIC and G-ISTA for different dimensions with $\mu = 0.06$ and $\tau = 1\%$	60
3.2	Results for real gene data sets. The penalty parameter μ is chosen by StARS and listed in the third column. The computational time is listed and we also compare the GIBSON estimate to those of QUIC and G-ISTA in terms of MSE.	64
4.1	Results for different sample sizes corresponding to Figure 4.1	97
4.2	Results for different penalty parameters corresponding to Figure 4.2 .	100
4.3	The feasibility and MSE $\ X - X^*\ /N^2$ of each method corresponding to Figure 4.4. X^* is the averaged estimate of benchmark methods. .	102
4.4	Performance for four real gene-expression data with $K = 400$. We report the run time of each method and their relative error against X^* (the averaged estimate of benchmark methods).	104

Chapter 1

Introduction

1.1 Motivation

Recent technology advancements have considerably facilitated data collection. A significant challenge is to exploit high-dimensional data and gain insight into complicated systems. Probabilistic graphical models [17] provide a general framework for compact representations, efficient inference, and effective learning of complicated systems. Specifically, probabilistic graphical models can be characterized by a set of nodes and edges. Given a complicated system, probabilistic graphical models formulate a corresponding multivariate distribution where each random variable is represented by one node in the graphical model. An edge is absent between two nodes if and only if the two random variables are conditionally independent given all the other remaining variables. In such way, probabilistic graphical models display the most significant interactions among variables. Moreover, probabilistic graphical models are helpful to derive efficient probabilistic inference algorithms. For example, the maximum a posterior (MAP) is one common and challenging inference problem. Inference algorithms such as belief propagation (BP) and embedded subgraphs (ES) take advantage of the graph structure and are generally orders of magnitude more

efficient than methods that do not exploit the graph structure [18]. Third, probabilistic graphical models empower us to construct the models effectively, either by expert knowledge or by data-driven learning. Such probabilistic models constitute a reasonable approximation to the data. Therefore, applications of probabilistic graphical models are popular across a variety of fields such as signal processing [19], computer vision [20], modeling of gene regulatory network [21], and protein structure analysis [22].

It is attractive, reasonable, and often safe to assume Gaussianity in assorted applications of probabilistic graphical models. The resulting model is named undirected Gaussian graphical models (GGM) (a.k.a. Gaussian Markov random fields). First of all, properties of the Gaussian distribution have been extensively studied. For example, the structure of the Gaussian graphical model is fully specified by the precision matrix. This property simplifies convoluted problems such as inference and learning in GGM. Other than the mathematical simplicity, the Gaussian distribution also provides surprisingly satisfactory approximation for assorted real-world distributions. The central limit theorem implies that, under certain common conditions, the sum of a large number of independent and identically distributed (i.i.d.) instances often follows an approximately Gaussian distribution, no matter what distributions the instance follows. As a consequence, many physical quantities can be regarded as Gaussian distributed. Another reason to assume Gaussianity is from the rule of maximum entropy. Among all real-valued probability distributions with predetermined first and second-order moments, the Gaussian distribution maximizes the entropy. In other words, the assumption of Gaussianity imposes the least number of constraints on the distribution given the specific mean and covariance, which can be conveniently obtained from data. Therefore, it is safe to assume Gaussianity for the rule of noninformative assumption or maximum entropy.

The primary and fundamental challenge in the application of GGM is how to construct a reliable model efficiently. There is a large body of literature about structure learning of GGM. However, the time complexity of the majority of existing

methods is $\mathcal{O}(N^3)$, i.e., cubic in terms of the dimension N , as such methods typically entail computationally burdensome operations such as matrix inversion, full spectral decomposition, and Cholesky decomposition. Such cumbersome operations are inevitable among existing methods for two reasons. The resulting objective function involves a log-determinant term. Miscellaneous approaches to tackle the log-determinant term necessitate either full spectral decomposition or the gradient which requires the inverse of a $N \times N$ symmetric positive definite (SPD) matrix. The other reason is the constraint of positive definiteness. It is often nontrivial to ensure the updated matrix is positive definite. In existing methods, the Cholesky decomposition is frequently adopted to check the definiteness. Such operations impose a $\mathcal{O}(N^3)$ computational bottleneck on existing methods, which severely impedes the scalability of structure learning for large problems.

In order to address this issue and further improve the scalability to high-dimensional problems, we employ the stochastic optimization method and the implicitly restarted Lanczos algorithm. Stochastic optimization such as stochastic approximation has become attractively popular in the optimization at large scale [23]. Stochastic approximation constructs a noisy yet efficient subgradient as a replacement of the complicated subgradient, which is either impossible to evaluate or too expensive in practice. As mentioned above, the log-determinant term makes the matrix inverse or spectral decomposition unavoidable, in other words, it convolutes structure learning. We propose an alternative method with the aid of stochastic approximation so that we can bypass cumbersome operations. Specifically, instead of calculating the exact subgradient that involves matrix inverse as in the existing methods, we leverage stochastic subgradients to determine the update direction. In other words, we circumvent the computation of the matrix inverse by finding a computationally efficient estimate of it. After addressing the issue from the log-determinant term, we proceed to reduce the time complexity of ensuring positive definiteness. Following the idea of back-tracking line search, we find the largest step size along the opposite direction of the stochastic subgradient so that the update matrix remains positive

definite. We check the definiteness by evaluating its smallest eigenvalue with the implicitly restarted Lanczos algorithm [24,25]. The resulting time complexity of the proposed algorithm is only $\mathcal{O}(N^2)$.

Moreover, we notice that structure learning of GGM is a specific instance of the general max-det optimization. Other than structure learning of GGM, log-determinant optimization problems arise in many other fields, such as computational geometry, statistics, and information and communication theory [26]. For example, given an ellipsoid denoted as $\{\mathcal{E} = \mathbf{x} \mid \|A\mathbf{x} + \mathbf{b}\| < 1\}$ with A as a SPD matrix, the volume of $\mathcal{E}$ is proportional to $\det(A^{-1})$. Thus finding the minimum volume ellipsoid with certain constraints often results in a log-determinant optimization problem. In the case of statistics, the log-likelihood function of exponential distributions includes the log-determinant term. Therefore, maximum likelihood estimation generally contains log-determinant optimization in plenty of statistical models. One particular example is sparse inverse covariance selection of Gaussian graphical model (GGM) [1] mentioned above. As for information theory, the entropy of a Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ involves $\log(\det(\Sigma))$. As a consequence, log-determinant optimization arises naturally in information theory and communications problems related to the computation of channel capacity [27]. In summary, the general log-determinant optimization is an important type of optimization covering assorted applications in various fields.

However, there is a crucial difference between the general log-determinant optimization problem and structure learning problem of GGM. Besides the challenging log-determinant term, the general log-determinant optimization commonly involves linear matrix constraints. The matrix constraints fundamentally convert the optimization problem from unconstrained optimization into constrained optimization. As a consequence, the aforementioned methods targeting structure learning problem cannot settle the general log-determinant optimization as such.

In order to adapt the proposed method to constrained optimization, we consider

the corresponding augmented Lagrangian form. In this way, we convert the optimization of the primal problem into searching for a saddle point of the augmented Lagrangian function. This technique has been introduced independently in [28, 29]. It incorporates the equality constraints into an unconstrained problem and provides stability in the optimization methods [30].

In the subsequent section, we review the literature for both structure learning of GGM and the general log-determinant optimization problem, compare their general mechanisms, and further motivate our research on the basis of their disadvantages.

1.2 Related Works

1.2.1 Structure Learning of GGM

Generally speaking, there are frequentist methods and Bayesian methods for structure learning of GGM [18]. Frequentist methods estimate the sparse factorization of the likelihood function of the input data. As a result, we formulate an explicit objective function which combines the likelihood function and an sparsity-inducing term. Bayesian methods perform MAP inference in the space of graphs. If the graph is decomposable, the marginal likelihood function can be computed in closed form. As a result, such methods are particularly efficient. However, Bayesian methods requires approximation techniques to solve non-decomposable cases. In this thesis, we focus on the frequentist methods, which leads to an optimization problem as follow.

1.2.1.1 Problem Formulation

We aim to learn the precision matrix given n N -variate Gaussian distributed observations $\mathbf{x}_{1:n}$. This problem is typically formulated as [1, 2, 13]:

$$\hat{X} \triangleq \underset{X \succ 0}{\operatorname{argmin}} \underbrace{\operatorname{tr}(SX) - \log \det X + \mu \|X\|_1}_{f(X)}, \quad (1.1)$$

where X is referred to as the precision matrix in multivariate Gaussian distribution, $f(X)$ is the objective function to be minimized, S is the empirical or sample covariance matrix defined as $S = 1/n \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T$, $\|\cdot\|_1$ is the ℓ_1 norm (i.e., the sum of the absolute value of all the elements in the matrix), and μ is the regularization or penalty parameter. In eq. (1.1), $\operatorname{tr}(SX) - \log \det X$ corresponds to the negative log-likelihood of the data $\mathbf{x}_{1:n}$ in terms of the precision matrix X . As an analogy to $\|\cdot\|$ in the classic least absolute shrinkage and selection operator (LASSO) [31], the ℓ_1 norm penalty on X promotes sparsity in the optimal precision matrix X . A properly chosen μ promotes sparsity in X while accurately describing the data. The objective function is non-smooth due to $\|X\|_1$.

Since eq. (1.1) is a non-smooth convex function with a positive definite constraint, and challenging to deal with, we hereby introduce two equivalent formulations that are frequently investigated in the literature. First, if we split the smooth and non-smooth part by introducing an auxiliary variable $Z = X$, eq. (1.1) can be equivalently rewritten as:

$$\begin{aligned} \min \quad & \operatorname{tr}(SX) - \log \det(X) + \mu \|Z\|_1 \\ \text{s.t.} \quad & X \succ 0, \quad X - Z = 0. \end{aligned} \quad (1.2)$$

The corresponding Lagrangian function is denoted by [9, 10]:

$$L(X, Z; W) \triangleq -\log \det(X) + \operatorname{tr}(SX) + \mu \|Z\|_1 + \operatorname{tr}(W(X - Z)), \quad (1.3)$$

where $W \in \mathbb{R}^{N \times N}$ is the multiplier of the linear constraint $X - Z = 0$. We can obtain the dual problem of eq. (1.1) by minimizing $L(X, Z; W)$ in eq. (1.3) w.r.t. the primal variables X and Z . We focus on the terms involving Z first. We have the following equality:

$$\inf_Z \mu \|Z\|_1 - \text{tr}(WZ) = \begin{cases} 0, & \text{if } |W_{i,j}| \leq \mu \\ \infty, & \text{otherwise.} \end{cases} \quad (1.4)$$

The terms associated with X are given by:

$$-\log \det(X) + \text{tr}((S + W)X), \quad (1.5)$$

whose infimum is attained at $X = (S + W)^{-1}$. Therefore, the dual form of eq. (1.1) can be written as [1, 3, 11]:

$$\hat{W} \triangleq \underset{\|W\|_\infty \leq \mu}{\text{argmax}} \log \det(S + W), \quad (1.6)$$

where $\|W\|_\infty \triangleq \max_{i,j} |W_{i,j}|$, and the positive definite constraint of $S + W$ is implicit in the objective function. Note that eq. (1.6) offers an alternative approach to tackle eq. (1.1) and is smooth [11]. Another important formulation is the corresponding augmented Lagrangian function [9, 10] of eq. (1.2):

$$\mathcal{L}(X, Z; W) \triangleq -\log \det(X) + \text{tr}(SX) + \mu \|Z\|_1 + \text{tr}(W(X - Z)) + \frac{\rho}{2} \|X - Z\|^2, \quad (1.7)$$

where $\|\cdot\|$ represent the Frobenius norm, and ρ is another penalty parameter controlling the violation of the linear constraint $X - Z = 0$.

Table 1.1: The time complexity for updating X or W and for guaranteeing the positive definiteness of X of existing methods [1–16].

Methods	Updating X or W	Time Complexity	Positive Definiteness	Time Complexity
GLASSO [1, 2]/DP-GLASSO [3]	row/column subproblem	$\mathcal{O}(N^4)$	structure property	NA
SINCO [4]	element-wise subproblem	$\mathcal{O}(N^4)$		
Nesterov's method [5]/ANS [6, 7]	smooth approximation and designed update	$\mathcal{O}(N^3)$	eigenvalue decomposition	$\mathcal{O}(N^3)$
IIPM [8]	Interior point method	$\mathcal{O}(N^4)$	Cholesky decomposition	$\mathcal{O}(N^4)$
ADM [9, 10]	alternating direction method	$\mathcal{O}(N^3)$	eigenvalue decomposition	$\mathcal{O}(N^3)$
PSM [11]	projection subgradient method	$\mathcal{O}(N^3)$	Cholesky decomposition	$\mathcal{O}(N^3)$
G-ISTA [12]	iterative shrinkage thresholding algorithm	$\mathcal{O}(NP)$	Cholesky decomposition	$\mathcal{O}(N^3)$
QUTC [13]		$\mathcal{O}(NP)$	Cholesky decomposition	$\mathcal{O}(N^3)$
BIG&QUTC [14, 15]		$\mathcal{O}(NP)$		
BCD-IC [16]	Newton's method	$\mathcal{O}(NP)$	Schur Complement	$\mathcal{O}(NP)$

1.2.1.2 Literature Review

We can divide existing methods into 7 categories: (block) coordinate descent methods, Nesterov’s methods, interior point methods, alternating direction methods, projection subgradient methods, proximal gradient methods, and Newton’s methods. Our discussion of these methods are concerned with two problems: first, how to update the precision matrix X or the dual variable W in order to optimize the objective function; second, how to guarantee the positive definiteness.

We first focus on the group of (block) coordinate descent methods. This group includes the popular GLASSO (Graphical LASSO) [1,2], DP-GLASSO (Dual update in the Primal GLASSO) [3], and SINCO (a greedy coordinate ascent method for Sparse INverse COvariance) [4]. These methods typically divide all variables into a certain number of blocks, formulate a subproblem of optimizing each block, and recursively solve these subproblems until convergence. The resulting subproblems are more tractable than the one in eq. (1.1). There are two partition strategies worth mentioning here: row or column-wise and elementwise. The first one partitions X or W into rows or columns. As an example, GLASSO [1,2] considers the dual problem in eq. (1.6) and decomposes W into rows or columns. Each row or column can be updated by solving a Lasso problem [31]. As an alternative, Banerjee *et al.* optimizes each row or column via the primal-dual interior point method [32]. On the other hand, DP-GLASSO targets at the primal problem and updates each row or column by quadratic programming. The second partition approach updates each element in X individually, such as SINCO. The time complexity of solving the subproblem is $\mathcal{O}(N^2)$ for SINCO and $\mathcal{O}(N^3)$ for GLASSO and DP-GLASSO. As a consequence, the overall time complexity of (block) coordinate methods is $\mathcal{O}(N^4)$. One appealing advantage of this group of methods is that the positive definiteness of X or $S + W$ is guaranteed automatically when solving the subproblems, given an initial positive definite guess.

Next, we turn our attention to the second set of methods, which is inspired by the

smoothing techniques proposed by Nesterov [33]. In [1, 5], the non-smooth objective function eq. (1.1) is approximated by a smooth function, therefore, a Lipschitz-continuous gradient can be computed. As a result, X can be updated following Nesterov's scheme. Along this line, an adaptive Nesterov smooth (ANS) method is further proposed in [6, 7] to tackle the dual problem eq. (1.6). These algorithms require eigendecomposition when updating X or W . Hence, the time complexity is $\mathcal{O}(N^3)$. By ensuring that the eigenvalues are positive at each step, X or $S + W$ is guaranteed to be positive definite.

The third category of methods leverages the interior point method. The standard interior point method costs $\mathcal{O}(N^6)$ arithmetic operations, as mentioned in [5, 8]. To mitigate the problem of high time complexity, an inexact primal-dual path-following interior-point method (IIPM) is proposed in [8]. By relying on inexact solutions in each iteration, the time complexity can be reduced to $\mathcal{O}(N^4)$. The positive definiteness of X is ensured in each iteration by performing Cholesky decomposition, whose time complexity is $\mathcal{O}(N^3)$.

The fourth group of methods, alternating direction methods (ADM) [9, 10], deals with the augmented objective function eq. (1.7). ADM updates X , Z , and W in eq. (1.7) separately in turn by solving the corresponding subproblem, until convergence. The subproblems involve eigendecomposition though [9, 10]. Thus, the time complexity of these approaches is $\mathcal{O}(N^3)$. Similar to Nesterov's method, X is guaranteed to be positive definite by making the eigenvalues positive in all iterations.

The fifth set employs projected gradient methods (PGM) [11]. It is obtained from the dual problem eq. (1.6). In each iteration, PGM updates W following the direction of the negative gradient and then projects the estimated W to the feasible domain $\|W\|_\infty \leq \mu$ by thresholding W . The gradient requires matrix inversion, and therefore, the time complexity of this method is $\mathcal{O}(N^3)$. Again, the definiteness of X is checked by performing Cholesky decomposition.

The sixth category adopts proximal gradient methods [12]. It separates the ob-

jective function eq. (1.1) into the smooth and non-smooth part, and replaces the smooth part by its first-order approximation in the proximal mapping. Interestingly, the proximal mapping amounts to a Lasso problem, which is solved by the iterative shrinkage thresholding algorithm (ISTA) [34]. The resulting algorithm is named G-ISTA (Graphical Iterative Shrinkage Thresholding). The computational bottleneck lies in calculating the gradient of the smooth part for the sake of finding the first-order approximation, which involves matrix inversion. However, if X is a sparse matrix with P non-zero elements, the time complexity of matrix inversion can be reduced to be $O(NP)$, where P is the number of non-zero elements in the intermediate estimation of X . This condition is met in G-ISTA, since the Lasso solver typically yields a sparse X . On the other hand, the definiteness of X is checked in every iteration via Cholesky decomposition.

Finally, let us concentrate on the last category of methods, i.e., proximal Newton methods [13–16]. This group consists of the popular QUIC (QUadratic approximation of Inverse Covariance matrices) method [13], BIG&QUIC [14, 15], and BCD-IC (Block-Coordinate Descent approach for large-scale sparse Inverse Covariance estimation) [16]. Instead of leveraging first-order approximation as in G-ISTA, second-order approximations are utilized in the proximal mapping. As a result, this set of methods typically converges faster than the aforementioned methods. The computational bottleneck with these methods still stems from the matrix inversion in the gradient and Hessian of the smooth part. To settle this problem, it has been proven in [13] that only a subset of entries in X needs to be updated that satisfies either of the following two conditions: (1) $X_{i,j} \neq 0$; (2) $|\partial_{X_{i,j}} f(X)| > \mu$, where $\partial_{X_{i,j}} f(X)$ is a subgradient of the objective function $f(X)$ w.r.t. $X_{i,j}$. Starting with a sparse X and given a sufficiently large μ , X can be sparse in all iterations. Therefore, the time complexity of updating X is $\mathcal{O}(NP)$. QUIC checks the definiteness by Cholesky decomposition, while BIG&QUIC and BCD-IC take full advantage of the Schur complement, reducing the time complexity to $\mathcal{O}(NP)$. Practically, however, QUIC is faster than BIG&QUIC and BCD-IC when $N < 30,000$, due to the small

pre-factors in the expression of the time complexity.

In Table 1.1, we summarize the aforementioned algorithms for updating X or W and for checking the definiteness, together with the corresponding time complexity. As demonstrated in the table, most existing methods suffer from the high time complexity of at least $\mathcal{O}(N^3)$. We also note from Table 1.1 that fruitful achievements have been achieved in [12–16] recently. They reduce the time complexity to $\mathcal{O}(NP)$ by exploiting the sparsity of the estimated matrix. These methods still suffer from the issue of high time complexity when the underlying true graph is relatively dense and so the number of non-zero entries P is large. In particular, we consider semi-sparse graphs as a specific case:

Definition 1.1. We call the graph family $\{\mathcal{G}_i\}$ as semi-sparse if:

$$\exists c_0 \in \mathbb{R}_+, \forall \mathcal{G}_i \in \{\mathcal{G}_i\} \text{ satisfies } \frac{|\mathcal{E}_i|}{|\mathcal{V}_i| \times (|\mathcal{V}_i| - 1)/2} = c_0,$$

where $\mathcal{E}_i$ represent the set of edges in $\mathcal{G}_i$, $\mathcal{V}_i$ represent the set of vertices in $\mathcal{G}_i$, and $\mathbb{R}_+$ denotes the set of positive real number.

In other words, the number of edges $|\mathcal{E}|$ grows quadratically with the number of nodes $|\mathcal{V}| = N$ in semi-sparse graphs. Under this scenario, the time complexity of the methods proposed in [12–16] becomes $\mathcal{O}(N^3)$. This issue impedes the application of such methods to high-dimensional semi-sparse graphs.

1.2.2 General Log-determinant Optimization

In this section, we analyze the objective function of the general log-determinant optimization, which is a generalization of the unconstrained optimization problem eq. (1.1). We formulate the standard optimization problem, provide an alternative formulation for later analysis, and review existing methods.

1.2.2.1 Problem formulation

The primal log-determinant optimization problem is formulated as [35, 36]:

$$\begin{aligned} \min \quad & f_0(X) \triangleq \text{tr}(CX) - \mu_1 \log \det(X) + \mu_2 \|X\|_1 \\ \text{s.t.} \quad & \mathcal{A}(X) - \mathbf{b} = \mathbf{0}, \quad X \succ 0, \end{aligned} \quad (1.8)$$

where $C \in \mathbb{S}_+^N$, $\mathbf{b} \in \mathbb{R}^M$, and $\mu_1, \mu_2 > 0$ are given parameters. The linear mapping $\mathcal{A} : \mathbb{R}^{N \times N} \rightarrow \mathbb{R}^M$ is defined as:

$$\mathcal{A}(X) = [\text{tr}(A_1 X^T), \text{tr}(A_2 X^T), \dots, \text{tr}(A_M X^T)]^T, \quad (1.9)$$

with $A_m \in \mathbb{R}^{N \times N}$, $m = 1, 2, \dots, M$. By introducing an auxiliary variable $Z = X$, an equivalent formulation of eq. (1.8) can be written as:

$$\begin{aligned} \min \quad & \text{tr}(CX) - \mu_1 \log \det(X) + \mu_2 \|Z\|_1 \\ \text{s.t.} \quad & \mathcal{A}(X) - \mathbf{b} = \mathbf{0}, \quad X - Z = 0, \quad X \succ 0. \end{aligned} \quad (1.10)$$

We can deduce the dual formulation from eq. (1.10). In detail, the Lagrangian form of eq. (1.10) is

$$\begin{aligned} L(X, Z, W, \mathbf{y}) \\ &= \text{tr}(CX) - \mu_1 \log \det(X) + \mu_2 \|Z\|_1 - (\mathcal{A}(X) - \mathbf{b})^T \mathbf{y} + \text{tr}(W(X - Z)) \\ &= \mathbf{b}^T \mathbf{y} + \text{tr}((C - \mathcal{A}^T(\mathbf{y}) + W)X) - \mu_1 \log \det(X) + \mu_2 \|Z\|_1 - \text{tr}(WZ), \end{aligned} \quad (1.11)$$

where there are two dual variables, i.e., $W \in \mathbb{R}^{N \times N}$ and $\mathbf{y} \in \mathbb{R}^M$. Minimizing the Lagrangian function L w.r.t. the primal variables Z and X yields the dual problem in terms of W and $\mathbf{y}$. Equation (1.11) enables us to exploit the separable structure of the optimization problem. Concretely, we have the following equation w.r.t. Z :

$$\inf_Z \mu_2 \|Z\|_1 - \text{tr}(WZ) = \begin{cases} 0, & \text{if } |W_{i,j}| \leq \mu_2 \\ \infty, & \text{otherwise.} \end{cases} \quad (1.12)$$

After minimizing L over Z in eq. (1.11), we proceed to analyze the following problem associated with X :

$$\inf_X \text{tr}((C - \mathcal{A}^T(\mathbf{y}) + W)X) - \mu_1 \log \det(X), \quad (1.13)$$

where $\mathcal{A}^T : \mathbb{R}^M \rightarrow \mathbb{R}^{N \times N}$ is defined as:

$$\mathcal{A}^T(\mathbf{y}) = \sum_{m=1}^M y_m A_m, \quad (1.14)$$

which is known as the adjoint of $\mathcal{A}$. The objective function in eq. (1.13) is differentiable. Therefore, the minimum of eq. (1.13) w.r.t. X is attained at $((C - \mathcal{A}^T(\mathbf{y}) + W)/\mu_1)^{-1}$. Minimizing eq. (1.11) over X and Z yields the dual problem of eq. (1.8):

$$\begin{aligned} \max_{\mathbf{y}, W} h(\mathbf{y}, W) &\triangleq \mathbf{b}^T \mathbf{y} + \mu_1 \text{tr}(I) - \mu_1 \log \det \left(\frac{C - \mathcal{A}^T(\mathbf{y}) + W}{\mu_1} \right)^{-1} \\ &= \mathbf{b}^T \mathbf{y} + \mu_1 \log \det(C + W - \mathcal{A}^T(\mathbf{y})) + N\mu_1(1 - \log \mu_1) \\ \text{s.t.} \quad &\|W\|_\infty \leq \mu_2, C + W - \mathcal{A}^T(\mathbf{y}) \succ 0, \end{aligned} \quad (1.15)$$

where the last term $N\mu_1(1 - \log \mu_1)$ is a constant w.r.t. the variables.

1.2.2.2 Related works

As mentioned in Section 1.1, there are a few methods for solving the log-determinant optimization problem eq. (1.8). We briefly review them in this section.

In [26], the proposed interior point method is the first attempt to handle a wide variety of log-determinant problems. It parameterizes the constrained objective into a dual central path problem and solves it with Newton's method. The algorithm updates the variable following the obtained Newton direction minimizing the duality gap. The exact time complexity is unknown. Due to computing the gradient and Hessian of the central path, the resulting time complexity per iteration is at least $\mathcal{O}(N^3)$, where N is the dimension of the objective variable X . The resulting

algorithm is impractical for moderate dimension, for example, $N > 200$ [36].

A Newton conjugate gradient (Newton-CG) based proximal point algorithm (PPA) is proposed in [35] aiming to solve the log-determinant optimization problems in large scale. An inexact Newton-CG method is implemented to solve the dual subproblem accelerating the standard PPA. The method is modified to handle block structure sparsity in [36]. The proximal point algorithm requires full spectrum eigenvalue decomposition in each iteration, which results in a computational bottleneck of $\mathcal{O}(N^3)$ as well.

The augmented Lagrangian method is exploited in [37] to solve the log-determinant optimization. It combines the proximal augmented Lagrangian and Newton-CG augmented Lagrangian as a hybrid approach. However, solving a linear system of size $N^2 \times N^2$ and computing full spectrum eigenvalue decomposition leads to a computational bottleneck of at least $\mathcal{O}(N^4)$.

A dual spectral projected gradient method was proposed in [38]. This method solve the dual problem eq. (1.15) via the spectral gradient method proposed in [39]. The computational bottleneck lies in Cholesky decomposition and evaluation of log-determinant function, which required $\mathcal{O}(N^3)$ operations.

Other than the methods aiming to solve the general log-determinant optimization with general linear constraints eq. (1.8), there is a specific type of constraint that is of particular interest in the field of statistics. In the case of GGM, the structure of the graphical model corresponds to the conditional independence structure. Moreover, the conditional independence relationship is fully specified by the zero pattern precision matrix X of the Gaussian distribution. In other words, if $X_{i,j} = 0$, then variable i and j are conditional independent given all the other variables. As a consequence, the edge (i, j) is absent in the GGM. In practice, if some entries in the variable X are assumed to be zero according to empirical analysis or prior knowledge, i.e. $\{X_{i,j} = 0 | (i, j) \in \Omega\}$ for $\Omega \subset \{(i, j) | i = 1, \dots, N; j = 1, \dots, N\}$, the

resulting constrained log-determinant optimization problem can be written as:

$$\begin{aligned} \min \quad & \text{tr}(CX) - \mu_1 \log \det(X) + \mu_2 \|X\|_1 \\ \text{s.t.} \quad & X_{i,j} = 0 \text{ for } (i, j) \in \Omega, \quad X \succ 0, \end{aligned} \tag{1.16}$$

where $\mu_1 = 1$. There are some specialized methods targeting this type of problems. In [7], the constrained log-determinant optimization is converted to a series of unconstrained subproblems with adaptively increasing ℓ_1 penalty on the entries in Ω . The subproblems are solved by spectral gradient projection [39] and Nesterov's smooth method [40]. The resulting methods are referred to as adaptive spectral projection gradient (ASPG) method and adaptive Nesterov's smooth (ANS) method respectively. Both methods require iteratively evaluating the primal objective value, the dual objective value, and matrix inversion. The resulting time complexity per iteration is therefore at least $\mathcal{O}(N^3)$.

In summary, the existing methods are subject to at least cubic time complexity, which stems from cumbersome operations such as matrix inversion, determinant evaluation, and full eigenvalue decomposition. This severely impedes the scalability of existing methods. In order to alleviate the impact of such cumbersome operations and achieve better scalability, we propose a primal-dual stochastic method in this thesis.

1.3 Thesis Organization and Overview of Contributions

This thesis is structured as follows. In Chapter 2, we review some background knowledge that is essential to understand the rest of this thesis. There are three important topics: Gaussian graphical models, convex optimization theory, and stochastic optimization theory. Specifically, we begin with the definition of graphical models in

the context of statistical theory. The graphical models feature attractive properties and have widespread applications when Gaussianity can be assumed. Therefore, we introduce the definition of Gaussian graphical model (GGM) and review its favorable attributes. Next, we recapitulate structure learning of GGM. In the section of convex optimization theory, we review some fundamental definitions and properties relevant to convex optimization. We briefly introduce the definition of Hilbert space, convex function, and duality theory. In addition, we go over some key points of stochastic optimization.

In Chapter 3, we briefly recapitulate structure learning problem of GGM and explain the reason for the cumbersome operations. With the identification of the computational bottleneck in existing algorithms, we propose a novel method named GIBSON (GIBbs-sampling-based StOchastic subgradient desceNt) for structure learning problem. GIBSON entails stochastic optimization and construct an alternative method to address the computational challenges. After specifying the details of GIBSON, we provide theoretical analysis of its performance. Specifically, we prove the strong duality and the bounded stochastic gradient in terms of its second-order moments. These two essential properties can lead us to the sublinear convergence rate of the proposed method. Next, we present the numerical results of GIBSON on both synthetic and real data. The numerical results validate the convergence guarantee from the theoretical results and certify the quadratic time complexity of GIBSON whereas the time complexity of existing methods is at least in the order of $\mathcal{O}(N^3)$.

Subsequent to the proposed method GIBSON, we tackle the general log-determinant optimization in Chapter 4. We notice that many existing methods are inapplicable in the general log-determinant optimization due to the general linear matrix constraints. Therefore, we tackle the general optimization problem in the context of primal-dual optimization, and propose the PDSGD-ad (adaptive Primal-Dual Stochastic subGradient Descent). The proposed method yields the optimal estimate and fulfil the constraints with $\mathcal{O}(N^2)$ arithmetic operations. Similar to the

analysis of GIBSON, we present both theoretical and empirical analysis of the proposed method. The theoretical analysis consists of assumptions, bounds on the optimal primal-dual pair, elementary analysis of consecutive iterations, and convergence guarantee of the estimate. We then numerically analyze the performance of the proposed method and give some remarks on the choice of some importance coefficients. Moreover, we compare the performance of the proposed method against the benchmark methods, validate the concluding theorem in the theoretical analysis, and confirm its quadratic time complexity in comparison with the cubic time complexity of benchmark methods.

Finally, we summarize the results of this thesis, offer conclusive remarks on the results, and suggest some ideas for future work in Chapter 5.

Chapter 2

Preliminaries

In this chapter, we present some background knowledge that is crucial for following chapters. We develop the concepts of graphical models, provide the definition of GGM, describe the fundamental properties of GGM, and formulate the optimization problem of learning GGM in Section 2.1. Next, we review some basic knowledge of convex optimization in Section 2.2. Finally we provide a brief introduction of stochastic optimization in Section 2.3.

2.1 Gaussian Graphical Model

Graphical models are a marriage of graph theory and probability theory. In order to clarify the definition of graphical model, we review the necessary concepts of graph theory in Section 2.1.1 and link the graph with multivariate distributions in Section 2.1.2. We present the definition of graphical models, and then focus on Gaussian graphical models in Section 2.1.3 and the associated structure learning problem in Section 2.1.4.

2.1.1 Graph Theory

A graph $\mathcal{G}$ can be characterized by a set of N nodes/vertices $\mathcal{V} = \{1, 2, \dots, N\}$ and a set of edges $\mathcal{E} = \{(i, j) : i, j \in \mathcal{V}\} \subset \mathcal{V} \times \mathcal{V}$. We can represent the edges in a more compact manner, i.e., by its adjacency matrix G . If $(i, j) \notin \mathcal{E}$ or $i = j$, we write $G_{i,j} = 0$. Otherwise we specify $G_{i,j} = 1$. When there is an edge between nodes i and j , we say they are neighbors. In summary, the following statements are equivalent:

- there is an edge between nodes i and j ;
- $(i, j) \in \mathcal{E}$;
- $G_{i,j} = 1$;
- i and j are neighbors.

If the associated adjacency matrix G is symmetric, then we regard the graph $\mathcal{G}$ to be undirected, otherwise it is directed. We focus on undirected graphs in this thesis. A clique $\mathcal{B}$ is a set of nodes that are neighbors of each other, i.e., $\forall i, j \in \mathcal{B}$, we have $(i, j) \in \mathcal{E}$. We say the clique is a maximal clique if adding any other node from $\mathcal{V}$ would undermine the clique property. In other words, if adding any node from $\{i | i \in \mathcal{V}, i \notin \mathcal{B}\}$ to $\mathcal{B}$ results that there $\exists i, j \in \mathcal{B}$ such that $(i, j) \notin \mathcal{E}$, the clique $\mathcal{B}$ is maximal. A path between i and j is a series of consecutive edges leading from i to j .

At last, we introduce the concept of graph separation, which is closely connected to graphical models. Assume we have the subsets of nodes $\mathcal{P}, \mathcal{Q}, \mathcal{S} \subset \mathcal{V}$. We claim that $\mathcal{S}$ separates $\mathcal{P}$ and $\mathcal{Q}$ if there is no path between any $i \in \mathcal{P}$ and $j \in \mathcal{Q}$ after removing the set $\mathcal{S}$ and all connected edges to $\mathcal{S}$ from the graph $\mathcal{G}$. We illustrate this concept with an exemplary graph in Figure 2.1. As indicated in Figure 2.1, if we remove all the red nodes ($\mathcal{S}$) and all the red edges, then there is no path between any brown node ($\mathcal{P}$) and any blue node ($\mathcal{Q}$). We denote the separation by $\mathcal{P} \perp \mathcal{Q} | \mathcal{S}$.

Figure 2.1: The brown nodes, blue nodes, and red nodes denote $\mathcal{P}, \mathcal{Q}, \mathcal{S} \subset \mathcal{V}$ respectively. After removing the red nodes and red edges, there would be no path between any arbitrary brown node and blue node. In such way, we say $\mathcal{S}$ separates $\mathcal{P}$ and $\mathcal{Q}$.

2.1.2 Graphical Model

With the definition and concepts on graphs, we construct a graphical model based on a multivariate probabilistic distribution. Specifically, given a N -variate distribution $p(\mathbf{x})$ with $\mathbf{x} = [x_1, x_2, \dots, x_N]^T$, the graphical model represents $p(\mathbf{x})$ by constructing a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Concretely, each node $i \in \mathcal{V}$ corresponds to a random variable x_i . The edge (i, j) is absent if x_i and x_j are conditionally independent given all the random variables excluding x_i and x_j , denoted by $\mathbf{x}_{-\{i,j\}}$. In other words,

$$(i, j) \notin \mathcal{E} \Leftrightarrow x_i \perp x_j | \mathbf{x}_{-\{i,j\}} \Leftrightarrow p(x_i, x_j | \mathbf{x}_{-\{i,j\}}) = p(x_i | \mathbf{x}_{-\{i,j\}}) p(x_j | \mathbf{x}_{-\{i,j\}}). \quad (2.1)$$

The resulting graph is an undirected graphical model, a.k.a. Markov random field (MRF). Equation (2.1) is called local Markov property for the undirected graphical model. In a more general case, we define the set $\mathbf{x}_{\mathcal{P}}$, $\mathbf{x}_{\mathcal{Q}}$ and $\mathbf{x}_{\mathcal{S}}$ as the sets of random variables corresponding to the set of nodes respectively. If $\mathcal{S}$ separates $\mathcal{P}$ and $\mathcal{Q}$, then $\mathbf{x}_{\mathcal{P}}$ and $\mathbf{x}_{\mathcal{Q}}$ is conditionally independent given $\mathbf{x}_{\mathcal{S}}$. Equivalently we can

write

$$\mathcal{P} \perp \mathcal{Q} | \mathcal{S} \Leftrightarrow \mathbf{x}_{\mathcal{P}} \perp \mathbf{x}_{\mathcal{Q}} | \mathbf{x}_{\mathcal{S}} \Leftrightarrow p(\mathbf{x}_{\mathcal{P}}, \mathbf{x}_{\mathcal{Q}} | \mathbf{x}_{\mathcal{S}}) = p(\mathbf{x}_{\mathcal{P}} | \mathbf{x}_{\mathcal{S}}) p(\mathbf{x}_{\mathcal{Q}} | \mathbf{x}_{\mathcal{S}}). \quad (2.2)$$

This is named global Markov property for the undirected graphical model. In this manner, the graphical model is fully specified by the conditional relationship of the multivariate distribution. Conversely, the pairwise and the global Markov properties of the graphical model characterize the conditional independence relationship among the random variables.

In order to efficiently construct the graphical model $\mathcal{G}$ given $p(\mathbf{x})$, we can factorize the probability distribution $p(\mathbf{x})$ into factors and relate one clique (fully connected subgraph) with each factor according to the following theorem.

Theorem 2.1. (Hammersley-Clifford Theorem [41]) *Suppose the distribution function $p(\mathbf{x})$ satisfies $p(\mathbf{x}) > 0$ for all $\mathbf{x}$, then the conditional relationship of $p(\mathbf{x})$ satisfies the Markov properties of $\mathcal{G}$ iff $p(\mathbf{x})$ can be factorized into a product of factors, one per maximal clique, i.e.,*

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{\mathcal{B} \in \mathcal{D}} \varphi_{\mathcal{B}}(\mathbf{x}_{\mathcal{B}}), \quad (2.3)$$

where $\mathcal{D}$ is the set of all maximal cliques of $\mathcal{G}$, $\varphi_{\mathcal{B}}(\mathbf{x}_{\mathcal{B}})$ is the potential function on the clique $\mathcal{B}$, and Z is the nominator which ensures the overall distribution $p(\mathbf{x})$ sums to 1.

2.1.3 Gaussian Graphical Model

The distribution function of a N -dimensional normal distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ is given by:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{N/2} \det(\Sigma)^{1/2}} \exp \left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right), \quad (2.4)$$

where $\boldsymbol{\mu}$ is the mean vector, $\Sigma \in \mathbb{S}^N$ is the covariance matrix, and $\det(\Sigma)$ represents the determinant of Σ . In order to represent the Gaussian distribution as a graphical model, we reparameterize eq. (2.4) by denoting the precision matrix by J , i.e., $J = \Sigma^{-1}$ and the potential vector by $\boldsymbol{\eta} = J\boldsymbol{\mu}$:

$$p(\mathbf{x}) = \frac{\det(J)^{1/2}}{(2\pi)^{N/2}} \exp\left(-\frac{1}{2}\mathbf{x}^T J \mathbf{x} + \boldsymbol{\eta}^T \mathbf{x}\right) \propto \exp\left(-\frac{1}{2}\mathbf{x}^T J \mathbf{x} + \boldsymbol{\eta}^T \mathbf{x}\right), \quad (2.5)$$

which is the information form of multivariate Gaussian distribution denoted by $\mathcal{N}(\boldsymbol{\eta}, J)$. J is also known as concentration matrix, information matrix, or precision matrix. We can factorize $p(\mathbf{x})$ according to eq. (2.5):

$$p(\mathbf{x}) \propto \prod_{i,j,i \neq j} \exp\left(-\frac{1}{2}x_i J_{i,j} x_j\right) \prod_i \exp\left(-\frac{1}{2}J_{i,i}x_i^2 + \boldsymbol{\eta}_i x_i\right). \quad (2.6)$$

Equation (2.6) is a pairwise factorization of $p(\mathbf{x})$. The term $\exp(-\frac{1}{2}x_i J_{i,j} x_j)$ is known as the edge potential of the edge (i, j) and $\exp(-\frac{1}{2}J_{i,i}x_i^2 + \boldsymbol{\eta}_i x_i)$ is the node potential of the node i . We can determine the pairwise relationship of the resulting graphical model following the formulation eq. (2.6). Concretely, if $J_{i,j} = 0$, the pairwise term between x_i and x_j vanishes. According to Hammersley-Clifford theorem, we can conclude:

$$x_i \perp x_j | \mathbf{x}_{-\{i,j\}} \Leftrightarrow J_{i,j} = 0. \quad (2.7)$$

In conclusion, the structure of the Gaussian graphical model is fully specified by the precision matrix of the multivariate distribution, i.e., the edge (i, j) is absent iff $J_{i,j} = 0$. In other words, learning the structure of Gaussian graphical models is equivalent to the estimation of the inverse covariance matrix (precision matrix) J . In the subsequent section, we provide more details related to this equivalence.

2.1.4 Learning Gaussian Graphical Models

Structure learning is one of the fundamental challenges in the application of graphical models. Fortunately, in the case of Gaussian graphical models, we can convert the problem of structure learning into inverse covariance estimation as explained in Section 2.1.3. Without loss of generality, we formulate the problem as learning the graph structure according to n instances $\mathbf{x}_{1:n}$, which are i.i.d. samples drawn from the N -dimensional Gaussian distribution $\mathcal{N}(\mathbf{0}, J)$. Otherwise, we normalize the samples by subtracting the sample mean vector.

The objective function consists of the log-likelihood term and an additional penalty term. The log-likelihood ensures data fidelity and the penalty term promotes sparsity in the graphical model. We begin with the log-likelihood term by introducing maximum likelihood estimation for the Gaussian distribution $\mathcal{N}(\mathbf{0}, J)$. By denoting the likelihood function of $\mathbf{x}_{1:n}$ with $q(J)$, we can write it following eq. (2.5):

$$\begin{aligned}
 q(J) &= \prod_{i=1}^n \frac{\det(J)^{1/2}}{(2\pi)^{N/2}} \exp\left(-\frac{1}{2} \mathbf{x}_i^T J \mathbf{x}_i\right) \\
 &= \frac{\det(J)^{n/2}}{(2\pi)^{nN/2}} \exp\left(-\frac{1}{2} \sum_{i=1}^n \mathbf{x}_i^T J \mathbf{x}_i\right) \\
 &= \frac{\det(J)^{n/2}}{(2\pi)^{nN/2}} \exp\left(-\frac{1}{2} \sum_{i=1}^n \text{tr}(\mathbf{x}_i^T J \mathbf{x}_i)\right) \\
 &= \frac{\det(J)^{n/2}}{(2\pi)^{nN/2}} \exp\left(-\frac{1}{2} \sum_{i=1}^n \text{tr}(J \mathbf{x}_i \mathbf{x}_i^T)\right) \\
 &= \frac{\det(J)^{n/2}}{(2\pi)^{nN/2}} \exp\left(-\frac{1}{2} \text{tr}\left(J \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T\right)\right).
 \end{aligned} \tag{2.8}$$

Recall that we define the sample covariance by $S = 1/n \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T$. We take the logarithm on both sides of eq. (2.8) and obtain:

$$\log q(J) = -\frac{nN}{2} \log(2\pi) + \frac{n}{2} \log \det(J) - \frac{n}{2} \text{tr}(JS). \tag{2.9}$$

In order to write the objective function eq. (2.9) in a more standard form, we replace notation J by X , remove the constant term $-nN/2 \log(2\pi)$, and divide the remaining terms by $-n/2$. As a consequence, maximizing the likelihood $q(J)$ becomes equivalent to minimizing the negative loglikelihood function $\ell(X)$:

$$\ell(X) = \text{tr}(SX) - \log \det(X). \quad (2.10)$$

A sparse precision matrix is preferred in practice in order to obtain a compact representation of the data [42] and to guard against data overfitting [1], especially in the case when the sample size n of the data is smaller than the dimension N . Since the zeros pattern of X corresponds to the absence of edges in the Gaussian graphical model, we intend to promote sparsity in X on the basis of eq. (2.10) by introducing an additional penalty term. Taken together, the resulting problem is formulated as maximizing the log-likelihood of X with the penalty on X , subject to the constraint that X is positive definite. The resulting objective function can be written as:

$$f(X) = \text{tr}(SX) - \log \det(X) + \mu \|X\|_1. \quad (2.11)$$

Similarly to LASSO, the ℓ_1 norm $\|X\|_1$ encourages sparsity in the off-diagonal entries [31]. The regularization or penalty parameter μ in front of the ℓ_1 -norm controls the trade-off between data fidelity and the sparsity of X . If μ is small, then the estimated X tends to be dense, otherwise X tends to be sparse. The corresponding objective function is convex but non-smooth. The objective function eq. (2.11) corresponds to the optimization problem eq. (1.1).

2.2 Convex Optimization

In this section, we specify some notations, terminologies, concepts, and knowledge in relation to convex optimization. Concretely, we present the definitions of convex sets and convex functions in Section 2.2.1. Next we formulate the standard convex optimization problem and introduce the Lagrangian duality theory in Section 2.2.4. Finally, we outline the foundations of the primal-dual optimization theory in Section 2.2.3.

2.2.1 Convex Sets and Convex Functions

If a set is closed under finite addition and scalar multiplication, we define it as a vector space. We state the formal definition and its properties here.

Definition 2.1. $\mathcal{V}$ is a vector space over the the real numbers $\mathbb{R}$ iff $\forall x_1, x_2, x_3 \in \mathcal{V}$ and $\forall r_1, r_2 \in \mathbb{R}$, we have the following properties:

- commutativity: $x_1 + x_2 = x_2 + x_1$;
- associativity of addition: $(x_1 + x_2) + x_3 = x_1 + (x_2 + x_3)$;
- additive identity: $0 + x_1 = x_1$;
- existence of additive inverse: there exists a $-x_1$ such that $x_1 + (-x_1) = 0$;
- associativity of scalar multiplication: $r_1(r_2x_1) = (r_1r_2)x_1$;
- distributivity of scalar sums: $(r_1 + r_2)x_1 = r_1x_1 + r_2x_1$;
- distributivity of vector sums: $r_1(x_1 + x_2) = r_1x_1 + r_1x_2$;
- scalar multiplication identity: $1x_1 = x_1$.

Note that we only consider the vector space on real numbers in this thesis for simplicity. In the more general case, the set of scalars can be specified as any field [43].

When a subset of the vector space is closed under line segmentation, we call the subset as a convex set. We present a formal definition of convex sets:

Definition 2.2. Given a subset $\mathcal{C}$ of a vector space $\mathcal{V}$, i.e., $\mathcal{C} \subset \mathcal{V}$, if $\forall x_1, x_2 \in \mathcal{C}$, the segment between x_1 and x_2 is contained in $\mathcal{C}$, that is, $\forall \theta \in [0, 1]$, we have $\theta x_1 + (1 - \theta)x_2 \in \mathcal{C}$, we call $\mathcal{C}$ a convex set.

Following the definition of a convex set, we can present the definition of convex functions. Generally, a function f is a mapping from a subset of $\mathcal{X}$ to another set $\mathcal{Y}$, denoted as $f : \mathcal{X} \rightarrow \mathcal{Y}$. Note that the notation does not imply that f is defined for every element from $\mathcal{X}$. In particular, we denote $\mathbf{dom} f$ as the set of all elements where f is properly defined, i.e., f maps x to exactly one element in $\mathcal{Y}$. In this thesis, we consider only real-valued function thus we restrict $\mathcal{Y} = \mathbb{R}$ which indicates that $\mathbf{dom} f = \{x | x \in \mathcal{X}, -\infty < f(x) < \infty\}$. For a real-value function, we define convexity as follows:

Definition 2.3. The function $f : \mathcal{X} \rightarrow \mathbb{R}$ is a convex function if:

- $\mathbf{dom} f$ is a convex set;
- $\forall x_1, x_2 \in \mathcal{X}$, and $\forall \theta \in [0, 1]$, we have:

$$f(\theta x_1 + (1 - \theta)x_2) \leq \theta f(x_1) + (1 - \theta)f(x_2). \quad (2.12)$$

Following the definition of convex sets and convex functions, we can introduce the general convex optimization problem:

$$\begin{aligned} \min \quad & f_0(x) \\ \text{s.t.} \quad & x \in \mathcal{C}, \end{aligned} \quad (2.13)$$

where $f_0(x)$ is a convex function and $\mathcal{C}$ is a convex subset of $\mathbf{dom}f$. If $f_0(x^*) = \inf\{f_0(x)|x \in \mathcal{C}\}$ and $x^* \in \mathcal{C}$, we call x^* the optimal point and we define $f_0^* \triangleq f(x^*)$ as the optimal value. All elements in $\mathcal{C}$ are named feasible points. When $\mathcal{C}$ is the $\mathbf{dom}f$, the constraint is trivial and we regard the optimization problem as unconstrained. Otherwise, we call eq. (2.13) a constrained optimization problem.

We are particularly interested in a specific type of spaces where the convex functions are defined. In the following, we define the Hilbert space, review a basic method for unconstrained optimization, and introduce an essential theory about constrained optimization.

2.2.2 Hilbert Space and Important Inequalities

If the distance between two elements in $\mathcal{X}$ can be measured, we call $\mathcal{X}$ a metric space, and can analyze the convergence properties of the function defined on it. In the following, we give the definition of a complete metric space, i.e., Hilbert space $\mathcal{H}$.

Definition 2.4. [43] A Hilbert space $\mathcal{H}$ is a vector space associated with an inner product that is also a complete metric space with respect to the metric induced by the inner product.

In order to explain the definition of Hilbert space, we clarify the inner product and completeness. We have the following properties w.r.t. the inner product operation:

- $\langle r_1x_1 + r_2x_2, x_3 \rangle = r_1\langle x_1, x_3 \rangle + r_2\langle x_2, x_3 \rangle;$
- $\langle x_1, x_2 \rangle = \langle x_2, x_1 \rangle;$
- $\langle x_1, x_1 \rangle \geq 0;$
- $\langle x_1, x_1 \rangle = 0$ iff $x_1 = 0,$

where $r_1, r_2 \in \mathbb{R}$. We can define a norm on $\mathcal{H}$ according to the inner product: $\|x_1\| = \sqrt{\langle x_1, x_1 \rangle}$, a.k.a. the induced norm of the inner product. The metric or distance function $\|x_1 - x_2\|$ follows naturally from the definition of the norm.

Following the definition of the distance function, we can explain the completeness. First, we define a Cauchy sequence $\{x_1, x_2, x_3, \dots\} \subset \mathcal{H}$, that is, $\forall \varepsilon > 0$, there $\exists n_0 > 0$ being a positive integer such that $\forall n_1, n_2 > n_0$, the distance between x_{n_1} and x_{n_2} satisfies $\|x_{n_1} - x_{n_2}\| < \varepsilon$. Then the sequence converge to a point x_0 , i.e., $\lim_{i \rightarrow \infty} \|x_i - x_0\| = 0$. The completeness of $\mathcal{H}$ assures that $x_0 \in \mathcal{H}$ for any Cauchy sequence defined on $\mathcal{H}$.

There are three Hilbert spaces that are of particular interest throughout this thesis. We specify the basic definitions and terminologies in this paragraph. A more complete analysis can be found in [44]. The space $\mathbb{R}^n$ contains all the n -dimensional real-numbered column vectors. Given $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$, we define the inner product as $\langle \mathbf{a}, \mathbf{b} \rangle = \sum_{i=1}^n a_i b_i$, whose induced norm is $\|\mathbf{a}\| = \sqrt{\sum_{i=1}^n a_i^2}$, a.k.a. Euclidean norm. $\mathbb{R}^n$ is known as Euclidean space. We can define other norms for the Euclidean space. For examples, ℓ_1 norm $\|\mathbf{a}\|_1 = \sum_{i=1}^n |a_i|$, ℓ_∞ norm $\|x\|_\infty = \max(|a_1|, \dots, |a_n|)$, where $|\cdot|$ represents the absolute value operator on scalars. The other Hilbert space is the set of all n -dimensional real-valued matrices $\mathbb{R}^{n \times n}$, where the inner product is defined as $\langle A, B \rangle = \text{tr}(AB^T) = \sum_{i,j} A_{i,j} B_{i,j}$ for $A, B \in \mathbb{R}^{n \times n}$. As an analogy to Euclidean space, if i and j are positive integers satisfying $1 \leq i, j \leq n$, we can define norms including Frobenius norm (the induced norm of the inner product) $\|A\| = \sqrt{\text{tr}(AA^T)} = \sqrt{\sum_{i,j} A_{i,j}^2}$, ℓ_1 norm $\|A\|_1 = \sum_{i,j} |A_{i,j}|$, ℓ_2 norm $\|A\|_2 = \lambda_{\max}(A)$ as the maximum eigenvalue of A , infinity norm $\|A\|_\infty = \max(|A_{i,j}|)$. The last Hilbert space is the set of all real-valued symmetric matrices, denoted by $\mathbb{S}^n$. The inner product and norms are defined similarly as in $\mathbb{R}^{n \times n}$. Furthermore, we denote the set of real-valued symmetric positive definite matrices as $\mathbb{S}_+^n$. We have the following relationship: $\mathbb{S}_+^n \subset \mathbb{S}^n \subset \mathbb{R}^{n \times n}$.

Moreover, we introduce some important inequalities that facilitate the theoretical

analysis in this thesis. Given a Hilbert space $\mathcal{H}$ associated with an inner product $\langle \cdot, \cdot \rangle$. The Cauchy-Schwarz inequality is given by [45]:

$$|\langle x, y \rangle| \leq \|x\| \|y\|. \quad (2.14)$$

Young's inequality [46] is as follows: given $a, b \in \mathbb{R}_+$, $p, q \in \mathbb{R}$ satisfying $p \geq 1, q \geq 1, 1/p + 1/q = 1$:

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}. \quad (2.15)$$

In the special case $p = q = 2$, we have:

$$ab \leq \frac{a^2}{2} + \frac{b^2}{2}. \quad (2.16)$$

Moreover, given a constant $\epsilon > 0$, we have the following Young's inequality with ϵ :

$$ab \leq \frac{a^2}{2\epsilon} + \frac{\epsilon b^2}{2}. \quad (2.17)$$

Given a vector $\mathbf{a} \in \mathbb{R}^n$, we have:

$$\|\mathbf{a}\| \leq \|\mathbf{a}\|_1 \leq \sqrt{M} \|\mathbf{a}\|. \quad (2.18)$$

In the case of n -dimensional matrix $A \in \mathbb{R}^{n \times n}$, the following inequality is valid [44]:

$$\|A\|_2 \leq \|A\| \leq \sqrt{N} \|A\|_2. \quad (2.19)$$

2.2.3 Basic Subgradient Method for Unconstrained Optimization

We focus on convex functions defined on $\mathcal{X} \subset \mathcal{H}$, where $\mathcal{H}$ is a Hilbert space. If the function f is differentiable, we denote its gradient by ∇f . The necessary and

sufficient condition for a differentiable function to be convex is that:

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \quad (2.20)$$

holds $\forall x, y \in \mathbf{dom} f$ [47]. If a function is convex but not differentiable, we can define an alternative concept as subgradient.

Definition 2.5. We say g is a subgradient of the convex function f at $x \in \mathbf{dom} f$ if $\forall y \in \mathbf{dom} f$, we have

$$f(y) \geq f(x) + \langle g, y - x \rangle. \quad (2.21)$$

We denote the set of all subgradients of f at x by $\partial f(x)$. The notation $g \in \partial f(x)$ indicates that g is a subgradient of the convex function f at x . Moreover, when f is differentiable, $\nabla f(x)$ is the only subgradient of $f(x)$ at the point x . In this sense, the subgradient is a generalization of the gradient in the context of convex optimization. By combining the definition of Hilbert space and the subgradient, we can proceed to define strong convexity as follow:

Definition 2.6. We say $f(x)$ is ξ -strongly convex for some $\xi > 0$ in a convex domain $\mathcal{X}$ that is a subset of some Hilbert space $\mathcal{H}$ with an inner product $\langle \cdot, \cdot \rangle$ and an induced norm $\| \cdot \|$, if $\forall x, y \in \mathcal{X}$ and any subgradient g of f :

$$f(x) - f(y) \geq \langle g(y), x - y \rangle + \frac{\xi}{2} \|x - y\|^2. \quad (2.22)$$

In the remaining part of this section, we assume $f_0 : \mathbb{R}^n \rightarrow \mathbb{R}$ for simplicity, and analyze the unconstrained optimization and constrained optimization respectively. When the optimization problem eq. (2.13) is unconstrained, our goal is to minimized $f_0(\mathbf{x})$ w.r.t. $\mathbf{x} \in \mathbb{R}^n$. In order to achieve this goal, the subgradient method iterates the following step:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t g(\mathbf{x}_t), \quad (2.23)$$

where t is the iteration number, $\mathbf{x}_t$ is the estimate at the t -th iteration, $g(\mathbf{x}_t)$ is

the subgradient of f at $\mathbf{x}_t$ (see Definition 2.5), and $\eta_t > 0$ is the step size (a.k.a. learning rate) in the t -th iteration.

Intuitively, the iterations eq. (2.23) yield a series of instances following the direction of $-g(\mathbf{x})$. With the subgradient $g(\mathbf{x})$ specified, the fundamental challenge is to choose a proper step size η_t . We briefly list five deterministic step size strategies:

- Constant step size: $\eta_t = c$ with $c > 0$.
- Constant step length: $\eta_t = c/\|g(\mathbf{x}_t)\|$ with $c > 0$.
- Square summable but not summable:

$$\eta_t \geq 0, \quad \sum_{t=0}^{\infty} \eta_t^2 < \infty, \quad \sum_{t=0}^{\infty} \eta_t = \infty. \quad (2.24)$$

One typical step size strategy is $\eta_t = c/(t+1)$ with $c > 0$.

- Nonsummable diminishing step size satisfying:

$$\eta_t \geq 0, \quad \lim_{t \rightarrow \infty} \eta_t = 0, \quad \sum_{t=0}^{\infty} \eta_t = \infty. \quad (2.25)$$

Typically, we can set $\eta = c/\sqrt{t+1}$ with $c > 0$.

- Nonsummable diminishing step length: $\eta_t = c_t/\|g(\mathbf{x}_t)\|$ with:

$$c_t \geq 0, \quad \lim_{t \rightarrow \infty} c_t = 0, \quad \sum_{t=0}^{\infty} c_t = \infty. \quad (2.26)$$

Such deterministic step size strategies prove effective for many simple applications. A thorough analysis of their convergence performance can be found in [48].

2.2.4 Lagrangian Duality Theory

In this section, we focus on the constrained version of eq. (2.13), i.e., $\mathcal{C}$ is a non-trivial constraint. In order to conveniently present relevant definitions, we assume $\mathcal{C}$ can be defined explicitly with constraints. In this way, we can write the standard convex optimization problem:

$$\begin{aligned} \min \quad & f_0(\mathbf{x}) \\ \text{s.t.} \quad & f_i(\mathbf{x}) \leq 0, i = 1, \dots, m \\ & l_i(\mathbf{x}) = 0, i = 1, \dots, p, \end{aligned} \tag{2.27}$$

where $\mathbf{x} \in \mathbb{R}^n$, $f_0(\mathbf{x}), f_1(\mathbf{x}), \dots, f_p(\mathbf{x}), l_1(\mathbf{x}), \dots, l_p(\mathbf{x}) : \mathbb{R}^n \rightarrow \mathbb{R}$ are convex functions, and we assume $\mathcal{D} = \cap_{i=0}^m \text{dom} f_i \cap_{i=1}^p \text{dom} l_i$ is non-empty. The Lagrangian function of eq. (2.27) is:

$$L(\mathbf{x}, \boldsymbol{\mu}, \mathbf{y}) = f_0(\mathbf{x}) + \sum_{i=1}^m \mu_i f_i(\mathbf{x}) + \sum_{i=1}^p y_i l_i(\mathbf{x}), \tag{2.28}$$

where $\boldsymbol{\mu} \in \mathbb{R}_+^m$ with $\mu_i \geq 0$ are the multipliers of inequality constraints, and $\mathbf{y} \in \mathbb{R}^p$ are the multipliers of the equality constraints. The Lagrangian dual function (a.k.a. dual function) of eq. (2.27) is obtained by minimizing eq. (2.28) w.r.t. $\mathbf{x}$:

$$h(\boldsymbol{\mu}, \mathbf{y}) = \inf_{\mathbf{x} \in \mathcal{D}} L(\mathbf{x}, \boldsymbol{\mu}, \mathbf{y}), \tag{2.29}$$

where $\boldsymbol{\mu}$ and $\mathbf{y}$ are dual variables. Suppose $\mathbf{x}_0$ is a feasible point of eq. (2.27), then $\sum_{i=1}^m \mu_i f_i(\mathbf{x}_0) + \sum_{i=1}^p y_i l_i(\mathbf{x}_0) \leq 0$ and we have the following inequality:

$$L(\mathbf{x}_0, \boldsymbol{\mu}, \mathbf{y}) \leq f_0(\mathbf{x}_0). \tag{2.30}$$

Equations (2.29) and (2.30) imply that

$$h(\boldsymbol{\mu}, \mathbf{y}) \leq f_0(\mathbf{x}_0) \tag{2.31}$$

holds for every feasible $\mathbf{x}_0$, thus $h(\boldsymbol{\mu}, \mathbf{y}) \leq f^*$ holds according to the definition of f^* . In other words, $h(\boldsymbol{\mu}, \mathbf{y})$ establishes a lower bound on the optimal value of eq. (2.27). The lower bound is non-trivial when $(\boldsymbol{\mu}, \mathbf{y}) \in \text{dom}h$ and $\mu_i \geq 0$. Maximizing the lower bound $h(\boldsymbol{\mu}, \mathbf{y})$ constitutes an alternative approach to tackle eq. (2.27) when the original problem is difficult to solve. The resulting optimization problem is the Lagrangian dual problem (a.k.a. dual problem):

$$\begin{aligned} \max \quad & h(\boldsymbol{\mu}, \mathbf{y}) \\ \text{s.t.} \quad & \mu_i \geq 0. \end{aligned} \tag{2.32}$$

The optimal value of eq. (2.32) is denoted by d^* . The optimization problem eq. (2.27) is called the primal problem. The inequality $d^* \leq f^*$, a.k.a. weak duality, holds according to eq. (2.30). If $d^* = f^*$ holds, we say that strong duality holds. Strong duality suggests that the lower bound obtained by optimizing the dual problem is a tight bound for the primal problem. However, the strong duality generally does not hold.

Fortunately, there is a sufficient condition for strong duality. When the eq. (2.27) can be written as:

$$\begin{aligned} \min \quad & f_0(\mathbf{x}) \\ \text{s.t.} \quad & f_i(\mathbf{x}) \leq 0, i = 1, \dots, m, \\ & \langle \mathbf{a}_i, \mathbf{x} \rangle = b_i, i = 1, \dots, p, \end{aligned} \tag{2.33}$$

if there exist a strictly feasible point $\mathbf{x}$, i.e, $f_i(\mathbf{x}) < 0$, for $i = 1, \dots, m$, and $\langle \mathbf{a}_i, \mathbf{x} \rangle = b_i$, for $i = 1, \dots, p$, then the strong duality holds. When the equality constraints are linear, the strong duality holds if there exists a point $\mathbf{x}$ satisfying all constraints while the nonlinear constraints hold with strict inequalities.

2.3 Stochastic Optimization

Stochastic optimization refers to a collection of methods that generate and exploit random variables for the purpose of optimizing an objective function. Generally, it targets the following random optimization problem:

$$\begin{aligned} \min_{\mathbf{x} \in \mathcal{C}} f_0(\mathbf{x}) &= \mathbb{E}_{\boldsymbol{\xi}}[f(\mathbf{x}, \boldsymbol{\xi})] \\ \text{s.t. } \mathbf{x} &\in \mathcal{C}, \end{aligned} \tag{2.34}$$

where $\mathcal{C} \subset \mathbb{R}^n$ is a convex set, $\boldsymbol{\xi}$ is a random variable whose probability distribution is supported on the set $\Xi \subset \mathbb{R}^m$, $f : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}$ is a well-defined function on $\mathcal{C} \times \Xi$ and is convex w.r.t. $\mathbf{x}$, and the expectation is well-defined and finite for every $\mathbf{x} \in \mathcal{C}$. Therefore, $f_0 : \mathbb{R}^n \rightarrow \mathbb{R}$ is a well-defined and convex function on $\mathcal{C}$. On inspection, eq. (2.34) is a specialized case of eq. (2.27). There are two important strategies solving eq. (2.34), which we will review in the following: sample average approximation (SAA) and stochastic approximation (SA).

Sample Average Approximation (SAA) [49, 50] consists of two steps. The first step invokes sampling. Specifically, it generates i.i.d. samples of $\boldsymbol{\xi}_{1:k}$, where k is a predetermined positive integer. It approximates $f_0(x)$ by averaging the cost function realization $f(\mathbf{x}, \boldsymbol{\xi}_i)$:

$$f_0(\mathbf{x}) = \mathbb{E}_{\boldsymbol{\xi}}[f(\mathbf{x}, \boldsymbol{\xi})] \approx \frac{1}{k} \sum_{i=1}^k f(\mathbf{x}, \boldsymbol{\xi}_i). \tag{2.35}$$

The sampling step converts the random optimization problem into a deterministic problem by removing the random variables. Following the sampling step we solve the deterministic problem and obtain the approximate solution. The performance of the optimizer relies on proper sampling of $\boldsymbol{\xi}$. The theoretical guarantee originates from the law of large numbers and the central limit theorem.

The other optimization strategy is stochastic approximation (SA) [51], which is

also referred to as stochastic subgradient descent (SGD) in the field of optimization. Stochastic subgradient descent has become an increasingly important topic for large-scale optimization [23]. If the problem eq. (2.34) is unconstrained, SA iteratively updates $\mathbf{x}$ following the opposite direction of a stochastic realization of $\nabla f_0(\mathbf{x})$. In the case of differentiable closed-form objective function, it samples one realization $\boldsymbol{\xi}_t$ of $\boldsymbol{\xi}$, computes $g_0(\mathbf{x}_t, \boldsymbol{\xi}_t)$, the noisy estimate of the subgradient $\partial f_0(\mathbf{x})$ in iteration t according to $\boldsymbol{\xi}_t$, and obtains the estimate of iteration $t + 1$ as:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t g_0(\mathbf{x}_t, \boldsymbol{\xi}_t), \quad (2.36)$$

where $\eta_t > 0$ is the step size (a.k.a. learning rate). If there is no closed-expression of f_0 , finite difference [52] and variant methods can be exploited to approximate the subgradient. The choice of η_t significantly affects the convergence performance and is crucial for stochastic optimization. In general, the following condition guarantees the convergence of SA:

$$\eta_i > 0, \sum_{i=1}^{\infty} \eta_i = \infty, \sum_{i=1}^{\infty} \eta_i^2 < \infty. \quad (2.37)$$

A simple and common choice is $\eta_t = c/t + 1$, where c is a positive constant.

Chapter 3

Stochastic Optimization for Structure Learning in Gaussian Graphical Models

In this chapter, we introduce the proposed stochastic method GIBSON (GIBbs-sampling-based StOchastic subgradient desceNt). In Section 3.1, we introduce the methodology which covers the updating procedure and the guarantee of positive definiteness. We provide theoretical analysis of the proposed method in Section 3.2. Some implementation details are specified in Section 3.3. Then we present the numerical results of GIBSON on the basis of synthetic and real data in Section 3.4. Finally, we offer some concluding remarks in Section 3.5.

3.1 Stochastic Subgradient Descent

We follow the problem formulation in Section 1.2.1. As mentioned in Section 1.2.1, the log-determinant term and the positive definite constraint render the computational bottleneck $\mathcal{O}(N^3)$ unavoidable in existing methods. The proposed method

GIBSON reduces the time complexity by addressing those two challenges respectively.

3.1.1 Updating X Using Stochastic Subgradient Descent

Recall that we define the objective function $f(X)$ in eq. (1.1):

$$\hat{X} \triangleq \operatorname{argmin}_{X \succ 0} \underbrace{\operatorname{tr}(SX) - \log \det X + \mu \|X\|_1}_{f(X)},$$

where X denotes the precision matrix, S represents the sample covariance matrix of the input data $\mathbf{x}_{1:n}$, and μ is the sparse penalty parameter. One intuitive approach to find the precision matrix X that minimizes $f(X)$ is to update X along the opposite direction of the subgradient of $f(X)$ w.r.t. X . The exact subgradient can be expressed as:

$$f_X(X) = S - X^{-1} + \Lambda, \quad (3.1)$$

where

$$\Lambda_{ij} = \begin{cases} \mu, & i \neq j \text{ and } X_{ij} > 0 \\ -\mu, & i \neq j \text{ and } X_{ij} < 0 \\ 0, & i = j \text{ or } X_{ij} = 0. \end{cases} \quad (3.2)$$

Note that the computational bottleneck in eq. (3.1) lies in the matrix inversion X^{-1} . As mentioned in Section 1.2, exact gradient or subgradients are employed in the aforementioned methods for updating X [1–4, 8, 11–16]. Due to the matrix inversion term in the exact subgradient, the time complexity of these methods is $\mathcal{O}(N^3)$ for semi-sparse graphs (see Definition 1.1).

To settle the problem of the high time complexity, we propose to use stochastic subgradients when updating X . Stochastic approximation [53] iteratively updates X

along the opposite direction of an unbiased stochastic estimate of the exact subgradient. More precisely, we seek an unbiased estimate $g(X)$ of the exact subgradient $f_X(X)$, that is, $\mathbb{E}[g(X)] = f_X(X)$, where $\mathbb{E}[\cdot]$ denotes expectation operation, and we refer to $g(X)$ as the stochastic subgradient. It is often computationally cheaper to evaluate the stochastic subgradient $g(X)$ instead of the exact subgradient $f_X(X)$. The stochastic subgradient descent algorithm then proceeds by iteratively following realizations of $-g(X)$ with the step size η_t in iteration t :

$$X_t = X_{t-1} - \eta_t g(X_{t-1}). \quad (3.3)$$

When the step size $\eta_t > 0$ is set properly, this algorithm is guaranteed to converge to the global minimum of $f(X)$, which will be analyzed in Section 3.2.

Since the high computational complexity of computing $f_X(X)$ in eq. (3.1) arises from the matrix inverse X^{-1} , we intend to find an unbiased estimate of X^{-1} that can be efficiently calculated. To this end, we consider X_t as the precision matrix of a zero-mean Gaussian distribution and draw samples from this Gaussian distribution $\mathcal{N}(0, X_t^{-1})$. It follows that the covariance matrix of the samples constitutes an unbiased estimate of X_t^{-1} [54].

In this section, we employ Gibbs sampling [55] to draw samples from $\mathcal{N}(0, X^{-1})$. In general, given a multivariate distribution, Gibbs sampling periodically draws samples of each variable, conditioned on the most recent values of other variables. In our case, Gibbs sampling constructs a Markov chain with the stationary distribution being the targeted distribution $\mathcal{N}(0, X^{-1})$. Particularly, given the arbitrary initial sample $\mathbf{y}^{(0)}$, it samples the i -th component $y_i^{(k)}$ of the k -th sample ($1 \leq i \leq N$ and $1 \leq k \leq K$), from the following conditional scalar distribution:

$$p(y_i^{(k)} | y_1 = y_1^{(k)}, \dots, y_{i-1} = y_{i-1}^{(k)}, y_{i+1} = y_{i+1}^{(k-1)}, \dots, y_n = y_n^{(k-1)}) = \mathcal{N}\left(\frac{\sum_{j=1}^{i-1} X_{i,j} y_j^{(k)} + \sum_{j=i+1}^N X_{i,j} y_j^{(k-1)}}{X_{i,i}}, \frac{1}{X_{i,i}}\right), \quad (3.4)$$

where $(\sum_{j=1}^{i-1} X_{i,j} y_j^{(k)} + \sum_{j=i+1}^N X_{i,j} y_j^{(k-1)})/X_{i,i}$ is the conditional mean and $1/X_{i,i}$ is the conditional variance. In order to analyze the convergence of the sampling procedure, we introduce another interpretation of eq. (3.4): we obtain one sample $\mathbf{y}^{(k)}$ by cyclically repeating $i = 1, \dots, N$:

$$y_i^{(k)} \leftarrow -\frac{1}{X_{i,i}} \left(\sum_{j=1}^{i-1} X_{i,j} y_j^{(k)} + \sum_{j=i+1}^N X_{i,j} y_j^{(k-1)} \right) + \frac{v_i}{\sqrt{X_{i,i}}}, \quad (3.5)$$

where $v_i \sim \mathcal{N}(0, 1)$. After repeating the cycle K times, we attain the desired Gibbs samples $\mathbf{y}^{(1:K)}$.

The estimate of X_t^{-1} can then be expressed as the sample covariance of the Gibbs samples:

$$\Sigma^t \triangleq \frac{1}{K} \sum_{k=1}^K \mathbf{y}_t^{(k)} \mathbf{y}_t^{(k)T}, \quad (3.6)$$

and so the stochastic subgradient can be written as:

$$g(X_t) \triangleq S - \Sigma^t + \Lambda. \quad (3.7)$$

When setting the number of Gibbs samples K to be invariant w.r.t. the dimension N , the computational complexity of evaluating $g(X_t)$ eq. (3.7) is only $\mathcal{O}(N^2)$. Note that the above procedure spares us from computing X_t^{-1} deterministically, an operation that has time complexity $\mathcal{O}(N^3)$.

3.1.2 Guarantee of the Positive Definiteness of X

It has been proven in [1, 6] that the optimal solution X^* to eq. (1.1) has bounded eigenvalues.

Lemma 3.1. *For every $\mu > 0$, the solution of problem eq. (1.1) X^* is unique and*

is contained in a convex cone, i.e., $X^* \in \mathcal{M} \triangleq \{X \in \mathbb{S}^N \mid \alpha I \preceq X \preceq \beta I\}$, where

$$\alpha = \frac{1}{\|S\|_2 + N\mu}, \quad \beta = \frac{N}{\mu}. \quad (3.8)$$

Therefore, to safeguard against the non-positive definiteness of X_t , we can ensure that the smallest eigenvalue $\lambda_{\min}(X_t) \geq \alpha$ and the largest eigenvalue $\lambda_{\max}(X_t) \leq \beta$ for all t . To this end, we first compute the smallest eigenvalue $\lambda_{\min}(X_t)$ and the largest eigenvalue $\lambda_{\max}(X_t)$ of X_t . We then proceed to the next iteration $t + 1$ if $\lambda_{\min}(X_t) \geq \alpha$ and $\lambda_{\max}(X_t) \leq \beta$; otherwise, we halve the step size η_t until the conditions are satisfied. This line search method can be formally defined as follows:

Definition 3.1. Given $E = X - g\eta$ according to eq. (3.3), $\mathbb{P}_{\mathcal{M}}(V) \triangleq \operatorname{argmin}_F \{\|E - F\| \mid F \in \mathcal{M}, F = K - g\eta/2^i, i \in \mathbb{N}\}$.

We employ the implicitly restarted Lanczos (IRL) algorithm [25, 44] for the purpose of efficiently evaluating the extreme (smallest and largest) eigenvalues of X . To describe the IRL, we begin by reviewing the standard Lanczos iteration, which aims to tri-diagonalize X . Specifically it decomposes X as follows:

$$X = QTQ^T, \quad (3.9)$$

where T is $N \times N$ tridiagonal matrix and $Q = [\mathbf{q}_1, \dots, \mathbf{q}_N]$ is a $N \times N$ matrix consisting of orthonormal columns. To accomplish this decomposition, we initialize $\mathbf{r}_1$ as an arbitrary normalized vector. The Lanczos algorithm proceeds as [24]:

$$\mathbf{q}_1 = \frac{\mathbf{r}_1}{\|\mathbf{r}_1\|}, \quad (3.10a)$$

$$T_{1,1} = \mathbf{q}_1^T X \mathbf{q}_1, \quad (3.10b)$$

$$\mathbf{r}_2 = (X - T_{1,1}I)\mathbf{q}_1 - \|\mathbf{r}_1\|\mathbf{q}_1, \quad (3.10c)$$

$$T_{2,1} = T_{1,2} = \|\mathbf{r}_2\|. \quad (3.10d)$$

The first iteration of the Lanczos iterations in eq. (3.10) produces the first tri-

Algorithm 1 Lanczos Iterations

Input: Matrix X .**Output:** Decomposition $X = QTQ^T$ with $Q = [\mathbf{q}_1, \dots, \mathbf{q}_N]$ and T being tridiagonal.

- 1: Initialize an arbitrary $\mathbf{r}_1, T_{0,1} \leftarrow 1, \mathbf{q}_0 \leftarrow \mathbf{0}, t \leftarrow 0$;
 - 2: **while** $\|\mathbf{T}_{t,t+1}\| \neq 0$ **do**
 - 3: $\mathbf{q}_t \leftarrow \mathbf{r}_t / \|\mathbf{r}_t\|$;
 - 4: $t \leftarrow t + 1$;
 - 5: $T_{t,t} \leftarrow \mathbf{q}_t^T A \mathbf{q}_t$;
 - 6: $\mathbf{r}_t \leftarrow (A - T_{t,t}I)\mathbf{q}_t - T_{t-1,t}\mathbf{q}_{t-1}$;
 - 7: $T_{t,t+1} \leftarrow \|\mathbf{r}_t\|_2, T_{t,t-1} \leftarrow \|\mathbf{r}_t\|$;
 - 8: **end while**
-

diagonal elements $\{T_{1,1}, T_{1,2}, T_{2,1}\}$, the first column $\mathbf{q}_1$ in Q , and the initialization vector for next column $\mathbf{r}_2$. In a similar manner, the Lanczos algorithm yields $\{T_{t,t}, T_{t,t+1}, T_{t+1,t}\}$ and $\mathbf{q}_t$ for $t = 2, \dots, N$. The resulting algorithm is summarized in Algorithm 1. Notice that the decomposition $X = QTQ^T$ would be different for different initialization vector $\mathbf{r}_1$. Since the columns in Q are orthonormal, the eigenvalues of X and T are identical to each other. Eigendecomposition of the tri-diagonal matrix T is substantially simplified compared with that of X . The resulting time complexity is reduced from $\mathcal{O}(N^3)$ to $\mathcal{O}(N^2)$ or even $\mathcal{O}(N \log(N))$ [56]. As a result, Lanczos iteration dominates the time complexity to obtain the smallest eigenvalue. The time complexity of eq. (3.10) is $\mathcal{O}(N^2)$ stemming from the matrix vector product, and consequently, it takes $\mathcal{O}(N^3)$ operations for the entire Lanczos iteration.

To address this issue, we can terminate the Lanczos iterations early, i.e., set a maximum number of Lanczos iterations m such that $m \ll N$ ($m = 20$ for example) [44]. As indicated in the Kaniel-Paige convergence theory [44], the extreme (largest and smallest) eigenvalues of $T_{1:m,1:m}$ yield a reliable estimate of the extreme eigenvalues of X . Furthermore, we can employ the IRL method to shrink the estimation error. Specifically, after terminating the Lanczos iterations and obtaining $T_{1:m,1:m}$ and $Q_{:,1:m}$ such that $X \approx Q_{:,1:m} T_{1:m,1:m} Q_{:,1:m}^T$, we can evaluate the residual of eq. (3.9) via $R_0 \triangleq XQ_{:,1:m} - Q_{:,1:m} T_{1:m,1:m}$, where $R_0 \in \mathbb{R}^{N \times m}$. The residual R_0 is then

used to construct an improved starting vector $\tilde{\mathbf{r}}_1$, with which we restart the Lanczos iterations and attain a new decomposition pair $X \approx \tilde{Q}_{:,1:m} \tilde{T}_{1:m,1:m} \tilde{Q}_{:,1:m}^T$. IRL recursively terminates the Lanczos iteration early and restarts it according to the residual R_0 . Assorted IRL schemes [57] have been proposed to force $\|R_0\|$ to 0. As a consequence, the extreme eigenvalues of $T_{1:m,1:m}$ converge to that of X as proven in [58]. In practice, IRL converges extremely fast [59] and the maximum times of restarting the Lanczos iterations can be fixed regardless of the dimension N [60]. Therefore, the overall time complexity of the IRL algorithm is $\mathcal{O}(N^2)$.

With determined step size, we can summarize the update rules as below:

$$\Sigma^t \triangleq \frac{1}{K} \sum_{k=1}^K \mathbf{y}_t^{(k)} \mathbf{y}_t^{(k)T}, \quad (3.11a)$$

$$g(X_t) \triangleq S - \Sigma^t + \Lambda, \quad (3.11b)$$

$$X_t = \mathbb{P}_{\mathcal{M}}(X_t - \eta_t g(X_{t-1})). \quad (3.11c)$$

We proceed to analyze the convergence performance of the proposed method.

3.2 Theoretical Results

After presenting the algorithm in Section 3.1, we proceed to analyze the theoretical properties, including the strong convexity of the objective function, the bound of the stochastic subgradient in terms of its second-order moment, and the convergence rate of the proposed method.

3.2.1 Convergence Analysis

In this section, we analyze the convergence of the proposed algorithm. We will show that the algorithm is guaranteed to converge to the global optimum with a

convergence rate of $\mathcal{O}(\ln(t)/t)$. As a first step, we show that our objective function eq. (1.1) is strongly convex. The definition of strongly convex functions is given in Definition 2.6.

In our case, $\mathbb{R}^{N \times N}$ is a finite linear space featuring well-defined inner product as explained in Section 2.2.2. Without loss of generality, we use the norm $\|X\|_\infty = \max_{i,j}(|X_{i,j}|)$ defined in Section 2.2.2 in the following analysis. Given Definition 2.6, it is easy to derive that the negative log-determinant term in eq. (1.1) is strongly convex on the compact convex set $\mathcal{M} \subset \mathbb{R}^{N \times N}$. Moreover, the other two terms $\text{tr}(SX)$ and $\mu\|X\|_1$ in eq. (1.1) are convex. As such, we have the following conclusion w.r.t. the objective function eq. (1.1).

Lemma 3.2. *The objective function $f(X)$ in eq. (1.1) is ξ -strongly convex in $\mathcal{M}$.*

Next, we prove that $\|g(X)\|_\infty^2$, i.e., the second-order moment of the stochastic subgradient $g(X)$, is finite, which further implies a bound on $\|g(X)\|^2$. In every iteration t , the stochastic subgradient (cf. eqs. (3.6) and (3.7)) consists of three terms: the empirical covariance matrix S of the input data $\mathbf{x}_{1:n}$, the sample covariance Σ^t of Gibbs samples $\mathbf{y}^{(1:K)}$, and the subgradient Λ of the penalty term $\mu\|X\|_1$. S is deterministic and $\|\Lambda\|_\infty^2 \leq \mu^2$. Thus, we focus on the remaining term $\Sigma^t = 1/K \sum_{k=1}^K \mathbf{y}_t^{(k)} \mathbf{y}_t^{(k)T}$. The distribution of $\mathbf{y}_t^{(k)} \mathbf{y}_t^{(k)T}$ is a N -variate singular Wishart distribution $\mathcal{W}_N(C_t^{(k)}, 1)$, where $C_t^{(k)}$ is a N -dimensional positive semidefinite scale matrix satisfying

$$C_t^{(k)} = \mathbb{E}[\mathbf{y}_t^{(k)} \mathbf{y}_t^{(k)T}], \quad (3.12)$$

and 1 is the degree of freedom. As a result, we can find an upper bound of the elements in $\mathbf{y}_t^{(k)} \mathbf{y}_t^{(k)T}$ as in the following lemma.

Lemma 3.3. *If A follows an N -dimensional singular Wishart distribution $\mathcal{W}_N(C, 1)$ with degree of freedom 1 and positive-semidefinite scale matrix C , then $\mathbb{E}[A_{ij}^2] = 2C_{ij}^2 + C_{ii}C_{jj}$, and therefore, $\mathbb{E}[\|A\|_\infty^2] \leq 3\|C\|_\infty^2$.*

Proof. Suppose that $A \sim \mathcal{W}_N(C, 1)$ and $B \sim \mathcal{W}_N(C, d)$, where $d \geq N$ is the degree of freedom for the Wishart distribution of B . It follows that $A + B \sim \mathcal{W}_N(C, 1 + d)$ (see Theorem 7 in [61]). According to the properties of Wishart distributions, we can obtain:

$$\mathbb{E}[(A + B)_{ij}^2] = (d + 1)(C_{ij}^2 + C_{ii}C_{jj}) + (d + 1)^2C_{ij}^2, \quad (3.13)$$

for $i, j \in \{1, 2, \dots, N\}$. On the other hand, we can rewrite the left hand side of the above expression as:

$$\begin{aligned} \mathbb{E}[(A + B)_{ij}^2] &= \mathbb{E}[A_{ij}^2 + 2A_{ij}B_{ij} + B_{ij}^2] \\ &= \mathbb{E}[A_{ij}^2] + 2\mathbb{E}[A_{ij}]\mathbb{E}[B_{ij}] + \mathbb{E}[B_{ij}^2], \end{aligned} \quad (3.14)$$

where $\mathbb{E}[A_{ij}] = C_{ij}$, $\mathbb{E}[B_{ij}] = dC_{ij}$ and $\mathbb{E}[B_{ij}^2] = d(C_{ij}^2 + C_{ii}C_{jj}) + d^2C_{ij}^2$. By canceling the common terms on both sides of eq. (3.13) and eq. (3.14), we can have:

$$\mathbb{E}[A_{ij}^2] = 2C_{ij}^2 + C_{ii}C_{jj}. \quad (3.15)$$

Furthermore, it follows from the definition of the norm that:

$$\begin{aligned} \mathbb{E}[\|A\|_\infty^2] &= \max_{i,j} (2C_{ij}^2 + C_{ii}C_{jj}) \\ &\leq \max_{i,j} 2C_{ij}^2 + \max_{i,j} C_{ii}C_{jj} \\ &\leq 3\|C\|_\infty^2. \end{aligned} \quad (3.16)$$

□

Based on lemma 3.3, we can present the following lemma, which upper bounds the second-order moments of $g(X_t)$ defined in eq. (3.6) eq. (3.7).

Lemma 3.4. *For an arbitrary symmetric positive definite matrix $X_t \in \mathcal{M}$ and $C_t^{(1)}$ with bounded $\|C_t^{(1)}\|_2$, $\exists r > 0$ such that $\forall t$, $E[\|g(X_t)\|^2] < r$.*

Proof. We can analyze the infinity norm $\mathbb{E}[\|g(X_t)\|_\infty^2]$ of $g(X_t)$ as:

$$\begin{aligned}\mathbb{E}[\|g(X_t)\|_\infty^2] &= \mathbb{E}[\|S - \Sigma^t + \Lambda\|_\infty^2] \\ &\leq 3(\mathbb{E}[\|S\|_\infty^2] + \mathbb{E}[\|\Sigma^t\|_\infty^2] + \mathbb{E}[\|\Lambda\|_\infty^2]).\end{aligned}\tag{3.17}$$

The above expression can be derived from the Cauchy-Schwarz inequality. As mentioned before, S is deterministic and bounded, while $\|\Lambda\|_\infty^2 \leq \mu^2$. The remaining term $\mathbb{E}[\|\Sigma^t\|_\infty^2]$ can be upper bounded as:

$$\begin{aligned}\mathbb{E}[\|\Sigma^t\|_\infty] &= \mathbb{E}\left[\left\|\frac{1}{K} \sum_{k=1}^K \mathbf{y}_t^{(k)} \mathbf{y}_t^{(k)T}\right\|_\infty\right] \\ &\leq \mathbb{E}\left[\frac{1}{K} \sum_{k=1}^K \|\mathbf{y}_t^{(k)} \mathbf{y}_t^{(k)T}\|_\infty\right] \\ &\leq \mathbb{E}\left[\max_k \|\mathbf{y}_t^{(k)} \mathbf{y}_t^{(k)T}\|_\infty\right] \\ &= \max_k \mathbb{E}\left[\|\mathbf{y}_t^{(k)} \mathbf{y}_t^{(k)T}\|_\infty\right].\end{aligned}\tag{3.18}$$

In light of Lemma 4.5, we can have:

$$\mathbb{E}\left[\|\mathbf{y}_t^{(k)} \mathbf{y}_t^{(k)T}\|_\infty\right] \leq \sqrt{3} \|C_t^{(k)}\|_\infty.\tag{3.19}$$

Therefore,

$$\mathbb{E}[\|\Sigma^t\|_\infty] \leq \sqrt{3} \max_k \|C_t^{(k)}\|_\infty.\tag{3.20}$$

Now, we target at seeking an upper bound of $\max_k \|C_t^{(k)}\|_\infty$. Note that the Gibbs sampling update rule in eq. (3.5) can be regarded as a linear combination of other variables with additional Gaussian noise. Equivalently, the update rule can be written as [62]:

$$\mathbf{y}_t^{(k)} = -D_t^{-1} U_t^T \mathbf{y}_t^{(k)} - D_t^{-1} U_t \mathbf{y}_t^{(k-1)} + D_t^{-\frac{1}{2}} \mathbf{v}^{(k)},\tag{3.21}$$

where $\mathbf{v}^{(k)} \sim \mathcal{N}(\mathbf{0}, I)$, D_t is the diagonal of X_t , and U_t is the strictly upper triangular

part. The above expression can be further simplified as eq. (3.22):

$$\mathbf{y}_t^{(k)} = - (D_t + U_t^T)^{-1} U_t \mathbf{y}_t^{(k-1)} + (D_t + U_t^T)^{-1} D_t^{\frac{1}{2}} \mathbf{v}^{(k)}. \quad (3.22)$$

As a result, the following equality holds [63]:

$$C_t^{(k)} - X_t^{-1} = B_t (C_t^{(k-1)} - X_t^{-1}) B_t^T, \quad (3.23)$$

where $B_t \triangleq -(D_t + U_t^T)^{-1} U_t$. It has been proven in [64] that $\|B_t\|_2 < 1$ for symmetric positive definite X_t , where $\|B_t\|_2$ is the spectral norm (i.e., the maximum eigenvalue of B_t). As such, we can tell from eq. (3.23) that the difference between $C_t^{(k)}$ and X_t^{-1} becomes smaller as k increases. In other words, $\{C_t^{(k)} | k = 1, \dots, \infty\}$ converges to X_t^{-1} . By recursion, we can further express $C_t^{(k)}$ as a function of $C_t^{(1)}$ as:

$$\begin{aligned} C_t^{(k)} &= X_t^{-1} + B_t^{k-1} (C_t^{(1)} - X_t^{-1}) B_t^{k-1T} \\ &= X_t^{-1} - B_t^{k-1} X_t^{-1} B_t^{k-1T} + B_t^{k-1} C_t^{(1)} B_t^{k-1T}. \end{aligned} \quad (3.24)$$

We have the the following relationship between the infinity norm and the spectral norm [44, 65]:

$$\|C_t^{(k)}\|_\infty \leq \|C_t^{(k)}\|_2 \leq N \|C_t^{(k)}\|_\infty. \quad (3.25)$$

As a result, we can show that $\|C_t^{(k)}\|_\infty$ is upper bounded by proving that $\|C_t^{(k)}\|_2$ is upper bounded. To this end, we can relax $\|C_t^{(k)}\|_2$ according to eq. (3.24) as:

$$\|C_t^{(k)}\|_2 \leq \|X_t^{-1}\|_2 + \|(B_t^{k-1} X_t^{-1}) B_t^{k-1T}\|_2 + \|B_t^{k-1} C_t^{(1)} B_t^{k-1T}\|_2 \quad (3.26a)$$

$$\leq 2\|X_t^{-1}\|_2 + \|C_t^{(1)}\|_2, \quad (3.26b)$$

where eq. (3.26) holds due to the norm property $\|E + F\|_2 \leq \|E\|_2 + \|F\|_2$ for all $E, F \in \mathcal{M}$, and eq. (3.26b) holds since $\|EF\|_2 \leq \|E\|_2 \|F\|_2$ and $\|B_t\|_2 < 1$. As $C_t^{(1)}$ denotes the covariance of the initial Gibbs sample $\mathbf{y}_t^{(1)}$ in iteration t , it can be user-specified, and there, can be set to have bounded eigenvalues. We can then conclude that $\|C_t^{(k)}\|_2$ is upper bounded for all k and so is $\|C_t^{(k)}\|_\infty$. Therefore, we

establish the bounded on eq. (3.17). According to the equivalence of norms, this concludes the proof. $\square$

In order to analyze the convergence performance, we state the following two lemmas. The first one analyzes the line search procedure defined in Definition 3.1:

Lemma 3.5. *For the convex set $\mathcal{M} \subset \mathbb{R}^{N \times N}$, $\forall E \in \mathbb{R}^{N \times N}$ defined as in Definition 3.1 and $\forall E' \in \mathcal{M}$, we have*

$$\|\mathbb{P}_{\mathcal{M}}(E) - E'\| \leq \|E - E'\|. \quad (3.27)$$

This is a simple corollary of the definition of the line search procedure and convexity. This property assures that the line search procedure defined in Definition 3.1 shrinks the distance between E and $\mathcal{M}$. In addition, we can analyze any initial guess $X_0 \in \mathcal{M}$ that has the following property:

Lemma 3.6. *Given the parameter ξ in Lemma 3.2 and the inequality $\mathbb{E}[\|g(X_t)\|^2] < r$ in Lemma 3.4, $\forall X \in \mathcal{M}$, $\mathbb{E}[\|X - X^*\|^2] \leq 4r/\xi^2$.*

Proof. From Definition 2.6 we have:

$$f(X^*) - f(X) \geq \langle f_X(X), X^* - X \rangle + \frac{\xi}{2} \|X^* - X\|^2, \quad (3.28)$$

and consequently:

$$\begin{aligned} 0 &\leq f(X) - f(X^*) \\ &\leq \langle f_X(X), X - X^* \rangle - \frac{\xi}{2} \|X^* - X\|^2. \end{aligned} \quad (3.29)$$

The above inequality combined with the Cauchy-Schwarz inequality yields:

$$\begin{aligned} \|f_X(X)\|^2 \|X^* - X\|^2 &\geq \frac{\xi^2}{4} \|X^* - X\|^4, \\ \|f_X(X)\|^2 &\geq \frac{\xi^2}{4} \|X^* - X\|^2. \end{aligned} \quad (3.30)$$

Moreover,

$$\begin{aligned}
\mathbb{E}[\|g(X)\|^2] &= \mathbb{E}\|g(X) - f_X(X) + f_X(X)\|^2 \\
&= \mathbb{E}[\|\nabla_X f(X)\|^2] + \mathbb{E}[\|g(X) - \nabla_X f(X)\|^2] + 2\mathbb{E}[\langle g(X) - f_X(X), f_X(X) \rangle] \\
&\geq \mathbb{E}[\|f_X(X)\|^2].
\end{aligned} \tag{3.31}$$

According to Lemma 3.4 it follows that $\|X^* - X\|^2 \leq 4r/\xi^2$. $\square$

After obtaining the strong convexity of the objective function (cf. Lemma 3.2) and the bound of the stochastic subgradients (cf. Lemma 3.4), we can proceed to analyze how the algorithm converges according to the following theorem.

Theorem 3.1. [66] *Suppose $f(X)$ is ξ -strongly convex, and that $\mathbb{E}[\|g_t\|^2] \leq r$, $\forall t$. Consider stochastic subgradient descent with step size $\eta_t = c/t$. Then $\forall t > 1$, it holds that*

$$\mathbb{E}[f(X_t) - f(X^*)] \leq \frac{10cr}{t} + \frac{4zr}{ct} \left(1 + \ln\left(\frac{t}{2}\right)\right) + \frac{cr}{2} \frac{1 + \ln(t)}{t}. \tag{3.32}$$

Proof. For simplicity we denote here $g \triangleq g(X_t)$. According to Definition 2.6,

$$\langle g, X_t - X^* \rangle \geq f(X_t) - f(X^*) + \frac{\xi}{2} \|X_t - X^*\|^2. \tag{3.33}$$

On the basis of X^* being the optimal value of $f(X)$:

$$f(X_t) - f(X^*) \geq \frac{\xi}{2} \|X_t - X^*\|^2. \tag{3.34}$$

Together with the convexity property, we have $\forall X_{t+1}, X \in \mathcal{M}$ based on Lemma 3.5:

$$\begin{aligned}
\mathbb{E}[\|X_{t+1} - X\|^2] &= \mathbb{E}[\|\mathbb{P}_{\mathcal{M}}(X_t - \eta_t g_t) - X\|^2] \\
&\leq \mathbb{E}[\|X_t - \eta_t g_t - X\|^2] \\
&= \mathbb{E}[\|X_t - X\|^2] - 2\eta_t \mathbb{E}[\langle g_t, X_t - X \rangle] + \eta_t^2 r.
\end{aligned} \tag{3.35}$$

We select $i \in \{1, \dots, \lfloor t/2 \rfloor\}$. Similarly as in [67], we rearrange the summation of inner-product in eq. (3.35) over $\{t-i, \dots, t\}$:

$$\begin{aligned} \sum_{j=t-i}^t \mathbb{E}[\langle g_j, X_j - X \rangle] &\leq \frac{1}{2\eta_{\{t-i\}}} \mathbb{E}[\|X_{t-i} - X\|^2] \\ &\quad + \sum_{j=t-i+1}^t \frac{\mathbb{E}[\|X_j - X\|^2]}{2} \left(\frac{1}{\eta_j} - \frac{1}{\eta_{j-1}} \right) + \frac{r}{2} \sum_{j=t-i}^t \eta_j. \end{aligned} \quad (3.36)$$

By the convexity of $f(X)$ in eq. (1.1) we know:

$$f(X_i) - f(X) \leq \langle g_j, X_j - X \rangle. \quad (3.37)$$

We substitute eq. (3.37) and $\eta_t = c/t$ into eq. (3.36):

$$\begin{aligned} \mathbb{E} \left[\sum_{j=t-i}^t (f(X_{\{i\}}) - f(X)) \right] &\leq \frac{t-i}{2c} \mathbb{E}[\|X_{t-i} - X\|^2] + \frac{1}{2c} \sum_{j=t-i+1}^t \mathbb{E}[\|X_j - X\|^2] \\ &\quad + c \frac{r}{2} \sum_{j=t-i}^t \frac{1}{j}. \end{aligned} \quad (3.38)$$

When we set $X = X_{t-i}$, based on Lemma 3.6 we obtain $\forall j \geq t-i$:

$$\begin{aligned} \mathbb{E}[\|X_j - X_{t-i}\|^2] &\leq 2\mathbb{E}[\|X_j - X^*\|^2 + \|X_{t-i} - X^*\|^2] \\ &\leq \frac{8zr}{\xi^2} \left(\frac{1}{j} + \frac{1}{t-i} \right) \\ &\leq \frac{8zr}{t}. \end{aligned} \quad (3.39)$$

From eq. (3.38) and eq. (3.39) :

$$\mathbb{E} \left[\sum_{j=t-i}^t (f(X_i) - f(X_{t-i})) \right] \leq \frac{4zir}{ct} + c \frac{r}{2} \sum_{j=t-i}^t \frac{1}{j}. \quad (3.40)$$

By defining $h_{\{i\}} \triangleq 1/(i+1) \sum_{j=t-i}^t \mathbb{E}[f(X_j)]$ as the expected average of the last $i+1$

iteration, we have:

$$-\mathbb{E}[f(X_{\{t-i\}})] \leq -\mathbb{E}[h_{\{i\}}] + r \left(\frac{4zi}{ct} + \frac{c}{2} \sum_{t=t-i}^t \frac{1}{(i+1)t} \right), \quad (3.41)$$

and

$$\begin{aligned} i\mathbb{E}[h_{\{i-1\}}] &= (i+1)\mathbb{E}[h_{\{i\}}] - \mathbb{E}[f(X_{t-i})] \\ &\leq (i+1)\mathbb{E}[h_{\{i\}}] - \mathbb{E}[h_{\{i\}}] + r \left(\frac{4zi}{ct} + \frac{c}{2} \sum_{j=t-i}^t \frac{1}{(i+1)j} \right). \end{aligned} \quad (3.42)$$

From eq. (3.41) and eq. (3.42) we can write:

$$\mathbb{E}[h_{\{i-1\}}] \leq \mathbb{E}[h_{\{i\}}] + \frac{r}{i} \left(\frac{4zi}{ct} + \frac{c}{2} \sum_{t=t-i}^t \frac{1}{(i+1)t} \right). \quad (3.43)$$

Using this inequality repeatedly and summing from $i = 1$ to $i = \lfloor t/2 \rfloor$, we obtain:

$$\mathbb{E}[f(X_t)] = \mathbb{E}[h_{\{0\}}] \leq \mathbb{E}[h_{\{\lfloor t/2 \rfloor\}}] + \frac{4zr}{ct} \sum_{i=1}^{\lfloor t/2 \rfloor} \frac{1}{i} + \frac{c^2 r}{2} \sum_{i=1}^{\lfloor t/2 \rfloor} \sum_{t=t-i}^t \frac{1}{i(i+1)t}. \quad (3.44)$$

From Theorem 5 in [66], we have:

$$\mathbb{E}[h_{\lfloor t/2 \rfloor}] \leq f(X^*) + \frac{10cr}{t} \quad (3.45)$$

for $t > 1$. Moreover, we can write:

$$\sum_{i=1}^{\lfloor t/2 \rfloor} \sum_{t=t-i}^t \frac{1}{(i)(i+1)t} \leq \sum_{k=1}^{\lfloor t/2 \rfloor} \frac{1}{i(t-i)} = \frac{1}{t} \sum_{i=1}^{\lfloor t/2 \rfloor} \left(\frac{1}{i} + \frac{1}{t-i} \right) \leq \frac{1 + \ln t}{t}. \quad (3.46)$$

By combining eq. (3.44), eq. (3.45), and eq. (3.46), we can obtain:

$$\mathbb{E}[f(X_t) - f(X^*)] \leq \frac{10cr}{t} + \frac{4zr}{ct} \left(1 + \ln\left(\frac{t}{2}\right) \right) + \frac{cr}{2} \frac{1 + \ln(t)}{t}, \quad (3.47)$$

which is in the order of $\ln(t)/t$. □

Taken together, the convergence theorem of GIBSON can be stated as follows:

Theorem 3.2. *Given the objective function $f(X)$ in eq. (1.1), an initial positive definite guess of the precision matrix $X_0 \in \mathcal{M}$, and $C_t^{(1)}$ with bounded eigenvalues for all t , if we set step size $\eta_t = c/t$ in the GIBSON updates eq. (3.11), where c is a positive constant, then GIBSON converges to the global optimum of $f(X)$ with convergence rate $\mathcal{O}(\ln(t)/t)$.*

Proof. Given $X_0 \in \mathcal{M}$, it is straightforward to derive that $X_t \in \mathcal{M}$, $\forall t$ following the update rule in eq. (3.11). Following Lemma 3.4, we can show that the second-order moments of the stochastic subgradient $\mathbb{E}[\|g(X_t)\|^2] < r$ are bounded. On the other hand, it has been proven in Lemma 3.2 that the objective function eq. (1.1) is a ξ -strongly convex function. Consequently, the convergence rate of GIBSON is $\mathcal{O}(\ln(t)/t)$ resulting from Theorem 3.1. $\square$

3.3 Implementation

We elaborate on several implementation matters, such as the initialization, the default setting of parameters, the step size selection scheme, and the convergence criterion in this section.

We first discuss how we initialize the precision matrix X_0 . Equating the subgradient eq. (3.1) to zero yields the optimality conditions for the solution to eq. (1.1):

$$S - X^{-1} + \Lambda = 0, \quad (3.48)$$

where Λ is defined as eq. (3.2). In light of the optimality condition eq. (3.48), a reasonable initialization of X is a diagonal matrix X_0 with $[X_0]_{i,i} = (S_{i,i} + \mu)^{-1}$:

$$[X_0]_{i,j} = \begin{cases} 0, & i \neq j \\ \frac{1}{S_{i,i} + \mu}, & i = j. \end{cases} \quad (3.49)$$

Second, we notice that $g(X_t)$ resulting from the Gibbs sampler is only asymptotically unbiased when $K \rightarrow \infty$. In practice, however, the algorithm still converges with a finite K . We obtain empirically that a finite K is sufficient to convergence.

Third, we specify how to initialize the Gibbs sampler in each iteration. In the first iteration $t = 0$, we begin with a random vector drawn from the diagonal matrix X_0 . It is therefore easy to obtain that $\|C_0^{(1)}\|_2$ is upper bounded. As derived in Lemma 3.4, $\|C_0^{(2:K)}\|_2$ are all upper bounded. When $t \geq 1$, we set:

$$\mathbf{y}_t^{(1)} \triangleq \mathbf{y}_{t-1}^{(K)}. \quad (3.50)$$

As a consequence, $C_t^{(1)} = C_{t-1}^{(K)}$ and its spectral norm is bounded. According to Lemma 3.4, the second-order moments of the stochastic subgradients in all iterations are upper bounded, and therefore, the algorithm is guaranteed to converge.

Fourth, instead of using the fixed step size scheme c/t as in Theorem 3.2, we utilize a dynamic step size scheme, since such scheme is shown to speed up the convergence both theoretically and practically [68, 69]. Concretely, we determine the step size as:

$$\begin{aligned} \eta_t &= \eta^t \frac{\|\mathbb{E}[g]^T \mathbb{E}[g]\|^2}{\|\mathbb{E}[g^2]\|^2}, \\ \mathbb{E}[g^2]_t &= \theta \mathbb{E}[g^2]_{t-1} + (1 - \theta) g_t(X_t)^T g_t(X_t), \\ \mathbb{E}[g]_t &= \theta \mathbb{E}[g]_{t-1} + (1 - \theta) g_t(X_t), \end{aligned} \quad (3.51)$$

where $\|\cdot\|$ denotes the Frobenius norm, $\mathbb{E}[g]_t$ and $\mathbb{E}[g^2]_t$ respectively represent the first and second-order moment of the stochastic subgradient in iteration t , and η is a shrinking parameter. The moment terms $\mathbb{E}[g]_t$ and $\mathbb{E}[g^2]_t$ are approximated by the corresponding moving average with $\theta < 1$ being a weight parameter. Note that the second-order moment can be decomposed as $\mathbb{E}[g^2] = \mathbb{E}[g]^T \mathbb{E}[g] + \mathbb{V}[g]$, where $\mathbb{V}[g]$ denotes the variance of the stochastic subgradient. If the variance $\mathbb{V}[g]$ is large, the above scheme shrinks the step size, thus mitigating the risk of taking a large

step in a wrong direction, which enables the stochastic subgradient descent to avoid overshooting in a certain degree. Otherwise, the step size becomes large, and the convergence is accelerated.

Fifth, the termination condition is defined based on the run difference of the estimates. Specifically, we average the estimates of X every t_0 iterations as:

$$\tilde{X}_j \triangleq \frac{1}{t_0} \sum_{i=1}^{t_0} X_{jt_0+i}, \quad (3.52)$$

where $j \in \mathbb{N}$. We stop the algorithm with X_t as the estimate $\hat{X}$ if the following condition is satisfied:

$$\|\tilde{X}_n - \tilde{X}_{n-1}\| < \epsilon, \quad (3.53)$$

where ϵ is a small constant controlling the termination. We set $t_0 = 20$ and $\epsilon = 0.001$ in our experiment.

Altogether, we name the resulting algorithm GIBSON (GIBbs-sampling-based StOchastic subgradient desceNt), which is summarized in Algorithm 2.

3.4 Numerical Results

In this section, we compare our model with the state-of-the-art methods, including QUIC [13], G-ISTA [12], and BIG&QUIC (see Figure 3.1). We will test on both synthetic data and real data. All the computations are performed on a 64-bit OS Windows server 2012 with two Intel Xeon(R) CPU E5-2690 v2 @3.00GHz processors with 32.0 GB RAM. Note that the code of QUIC, G-ISTA, and BIG&QUIC provided by the developers is compiled in C++, whereas we wrote the code for the GIBSON method in Matlab R2017b. In particular, we are interested in how the performance of the methods changes with the penalty parameter, the graph density, and the

Algorithm 2 GIBbs sampling-based Stochastic subGradient Descent (GIBSON)

Input: Data $\mathbf{x}^{(1:N)}$, penalty parameter λ , number of samples $K = 100$, $\theta = 0.9$, $\eta = 0.99$, $t_0 = 20$, $\epsilon = 0.001$, and $\gamma = 0.9$.

Output: X , the sparse precision matrix.

- 1: $t \leftarrow 0$, X_0 as in eq. (3.49);
- 2: **while** $t < \text{Maxiter}$ **do**
- 3: Generate $\mathbf{y}^{(1:K)}$ via eq. (3.5);
- 4: Compute the stochastic subgradient according to eq. (3.6) and eq. (3.7) with the step size as in eq. (3.51);
- 5: **while** TRUE **do**
- 6: Update X as in eq. (3.3);
- 7: Compute $\lambda_{\min}$ and $\lambda_{\max}$ of the updated X_{new} via the IRL method;
- 8: **if** $\lambda_{\min} \geq \alpha$ and $\lambda_{\max} \leq \beta$ **then**
- 9: $X_{t+1} \leftarrow X_{\text{new}}$;
- 10: **break**;
- 11: **else**
- 12: $\eta_t \leftarrow \eta_t/2$;
- 13: **end if**
- 14: **if** Converged **then**
- 15: **break**;
- 16: **else**
- 17: $t \leftarrow t + 1$;
- 18: **end if**
- 19: **end while**
- 20: **end while**

dimension.

3.4.1 Synthetic Dataset

We generate the random precision matrix as follows. Given the density $\tau < 1$, we determine the nonzero pattern of X by randomly selecting $n_e \triangleq \tau N(N-1)/2$ entries from the lower triangular part X_L , fill the nonzero elements with random numbers uniformly distributed in $[-1, 1]$, and compute $X = X_L + X_L^T$. The value of the diagonal entries equals the sum of the absolute value of the corresponding off-diagonal entries plus 0.1. The resulting matrix is then diagonally dominant and positive definite. We further rescale X such that the variance equals 1. Finally, we draw 1000 samples from the Gaussian distribution with zero mean and precision ma-

trix X . After obtaining the artificial data, we proceed to analyze the computational time w.r.t. three factors: penalty parameter, graph density, and dimension.

We first investigate how the computational time of all methods changes with μ (see Figure 3.1). We fix the density $\tau = 1\%$ and dimension $N=10,000$. We sweep the penalty parameter from 0.02 to 0.2 with interval 0.02. A smooth curve is fitted to the runtime in the log domain in order to better describe the trend. The resulting computational time of QUIC, G-ISTA, BIG&QUIC, and GIBSON as a function of μ is displayed in Figure 3.1. We can observe that the computational time of QUIC, G-ISTA, and BIG&QUIC is more sensitive to the value of μ . As mentioned in Section 1.2.1, these methods only update a subset of the entries in the precision matrix in every iteration and their time complexity of updating X is $\mathcal{O}(NP)$, where P is the number of non-zero entries in the final estimate of X . As a result, when μ becomes smaller, the estimated X is denser, and hence, the computational time of QUIC, G-ISTA, and BIG&QUIC increases dramatically. As opposed to these methods, the computational time of the proposed GIBSON algorithm is robust to μ , since the number of entries updated in all iterations is invariant to μ . In the following experiments, we show that for semi-sparse graphs, a relatively small μ is favored, and consequently, GIBSON outperforms other benchmark methods. Note that the key steps in QUIC and BIG&QUIC are almost the same, but BIG&QUIC is always slower than QUIC in our experiments. Therefore, we only present the results of QUIC in the following.

In addition, we show how the performance of the algorithms vary with the graph density. Specifically, we fix the dimension as $N=10,000$ and test different graph densities from $\tau \in \{10^n \mid n = -9, -8, \dots, -2\}$. For each experiment, we select the best μ from the candidate set $\{0.02i \mid i = 1, 2, \dots, 10\}$ that minimizes the F_1 -score. F_1 -score can be regarded as a weighted average of precision and recall, which can

Figure 3.1: Time comparison across different μ for $N=10,000$ and $\tau = 1\%$.

be defined as:

$$\begin{aligned} \text{precision} &\triangleq \frac{|\mathcal{E} \cap \mathcal{E}_\mu|}{|\mathcal{E}_\mu|}, \\ \text{recall} &\triangleq \frac{|\mathcal{E} \cap \mathcal{E}_\mu|}{|\mathcal{E}|}, \\ \text{F}_1\text{-score} &\triangleq \frac{2}{1/\text{precision} + 1/\text{recall}}, \end{aligned} \tag{3.54}$$

where $\mathcal{E}$ and $\mathcal{E}_\mu$ denotes the true and estimated set of edges, and $|\cdot|$ denotes the cardinality of the set. We then show the computational time with the best μ as a function of the graph density in Figure 3.2. A straight line is fitted for each method in the log domain. As expected, QUIC and G-ISTA are more time-consuming with the increase of the density, since their time complexity of updating X is $\mathcal{O}(NP)$. On the other hand, the computational time of GIBSON is flat w.r.t. the density, as its time complexity is $\mathcal{O}(N^2)$, and does not depend on the graph density. In addition, we can tell from Figure 3.2 that QUIC and G-ISTA performs better than GIBSON only when the density $\tau < 10^{-4}$. In practice, the underlying true graph may be denser, and then GIBSON is more efficient than QUIC and G-ISTA for these problems.

Next, we present the performance of these methods as a function of the dimension N . Here, we first show the results in the case where μ is fixed in Figure 3.3 and then in the case where the best μ is used that minimizes the F₁-score in Figure 3.4. Specifically, the graph density τ is set to be $\tau = 1\%$ and the dimension increases from 1000 to 10,000 with interval 1000. In Figure 3.3, we fix the penalty parameter as $\mu = \{0.02, 0.04, 0.06\}$. For the data of Figure 3.3 and Figure 3.4, we fit a straight line in the log domain and further compute the slope. As expected, the slope resulting from GIBSON is approximately 2 regardless of μ . However, the slope of QUIC and G-ISTA is larger than 3, and increases for decreasing μ values. Such observations are consistent with the theoretical time complexity of the three methods. On the other hand, in Figure 3.4, we choose the best μ from $\{0.02n \mid n = 1, 2, \dots, 10\}$. The computational time of GIBSON is still approximately a quadratic function of N ,

Figure 3.2: Time comparison across different density, experiments are executed on $N=10,000$ while μ is chosen to minimize the F_1 -score.

Table 3.1: MSE between GIBSON estimates and results from QUIC and G-ISTA for different dimensions with $\mu = 0.06$ and $\tau = 1\%$.

MSE	$N = 1000$	$N = 5000$	$N=10,000$
QUIC	2.9e-6	3.7e-7	2.7e-7
G-ISTA	2.9e-6	3.7e-7	2.7e-7

whereas that of QUIC and G-ISTA is at least a cubic function. Therefore, GIBSON provides an efficient tool for estimating the precision matrix of semi-sparse graphical models, especially for high-dimensional problems.

Finally, we check the accuracy of GIBSON by comparing it with QUIC and G-ISTA both quantitatively and qualitatively. Concretely, we show the mean square error (MSE) between the estimated precision matrix resulting from GIBSON and from the other two methods in Table 3.1. The small MSE indicates that GIBSON achieves the same accuracy as QUIC and G-ISTA. We further select 5 nonzero and 5 zero elements randomly from the underlying true precision matrix and depict the estimates yielded by the three methods in Figure 3.5. It can be observed that the estimates of QUIC and G-ISTA are almost identical and the estimates of GIBSON are very similar. The small difference can be explained by the stochastic nature of GIBSON.

3.4.2 Real Data

We consider three real gene data sets, including Lymph node status data [70], Arabidopsis thaliana data [71], and Lukemia data [72, 73]. The Lymph data is commonly analyzed to understand the long-term outlook of breast cancer treatment. It contains 148 samples and 587 variables. The second one, Arabidopsis, contains genetic information regarding plant traits. There are 118 samples and 834 variables. The third data set was developed for cancer classification originally, and it contains 1255 genes monitored by microarrays from 72 samples.

Figure 3.3: Computational time as a function of dimension N for different μ values.

Figure 3.4: Computational time as a function of dimension N for the best μ value.

For real data sets, the ground truth of the penalty μ is unknown. Therefore, we adopt StARS (Stability Approach to Regularization Selection) [74] to select an appropriate μ based on the observed data, which is a popular method for regularization selection. Note that we prefer a graph that is slightly denser than the ground truth in practice [74], since the false positives can be further removed in the future analysis whereas the false negatives are buried by the massive number of true negatives. As a result, StARS chooses the smallest penalty parameter so that the graph is sparse and replicable under random sampling. The selected μ is listed in Table 3.2. We also show the computational time for all methods and compute the MSE between

Figure 3.5: Randomly selected precision matrix entries plot of the three algorithms. There are 5 non-zero elements and 5 zero elements in the global optimal solution. We show $N = 1000$ and $N=10,000$ separately. The graph density is 1% and the penalty parameter is 0.06.

Table 3.2: Results for real gene data sets. The penalty parameter μ is chosen by StARS and listed in the third column. The computational time is listed and we also compare the GIBSON estimate to those of QUIC and G-ISTA in terms of MSE.

Data Set	N	μ		G-ISTA	QUIC	GIBSON
Lymph	587	0.1003	Time	235.9683	40.5381	72.8
			MSE	7.9e-5	7.9e-5	NA
Arabidopsis	834	0.1003	Time	316.7	131.5	109.9
			MSE	7.7e-5	7.7e-5	NA
Leukemia	1255	0.1225	Time	1022.1	273.5	167.6
			MSE	2.6e-5	2.6e-5	NA

the estimates of GIBSON and those of QUIC and G-ISTA. As shown in Table 3.2, GIBSON achieves almost the same accuracy as QUIC and G-ISTA. Furthermore, it is faster than the other two methods for Arabidopsis and Lukemia data, when the dimension is large. On the other hand, for Lymph data with a lower dimension, QUIC is the fastest. Again, these observations validate the usefulness of GIBSON when coupled with StARS, especially for data with a relatively large dimension.

3.5 Conclusion

In this chapter, we present the GIBSON method for estimating the inverse covariance of semi-sparse Gaussian graphical models. The problem can be formulated as maximizing the ℓ_1 -norm penalized log-likelihood of the precision matrix. Instead of deriving deterministic update rules as in the existing works [1–16], we resort to stochastic subgradients. In particular, the stochastic subgradient is computed via Gibbs sampling. On the other hand, we apply the implicitly restarted Lanczos algorithm to check the definiteness of the precision matrix during the updates. The resulting computational complexity is only $\mathcal{O}(N^2)$, whereas it is at least $\mathcal{O}(N^3)$ for existing methods. We further analyze the convergence properties of the proposed GIBSON method. Numerical results show that the empirical computational time of

GIBSON is approximately a quadratic function of the dimension N . Moreover, it achieves almost the same accuracy as that of QUIC and G-ISTA while the computational time is substantially less compared to existing methods for high-dimensional problems.

Chapter 4

Stochastic Optimization for the General Log-determinant Optimization

In this chapter, we tackle the general log-determinant problem following a similar approach as that in GIBSON. In Section 4.1, we introduce the methodology of the proposed method. Next we analyze the theoretical performance in Section 4.2. Some implementation matters are specified in Section 4.3. In Section 4.4, we present numerical results on synthetic and real data, which analyze the choice of parameters, validate the theoretical convergence results, and substantiate the advantageous properties in comparison with benchmark methods. Finally, we summarize the main results in Section 4.5.

4.1 Primal-Dual Stochastic Subgradient Method

In this section, we construct the augmented Lagrangian formulation of eq. (1.8) and propose an efficient method tackling the resulting problem.

4.1.1 Primal-Dual Subgradient Method

The augmented form of eq. (1.8) is defined as in [30]:

$$\begin{aligned} \min \quad & \text{tr}(CX) - \mu_1 \log \det(X) + \mu_2 \|X\|_1 + \frac{\rho}{2} \|\mathcal{A}(X) - \mathbf{b}\|^2 \\ \text{s.t.} \quad & \mathcal{A}(X) - \mathbf{b} = \mathbf{0}, \quad X \succ \mathbf{0}, \end{aligned} \quad (4.1)$$

where $\rho > 0$ is a penalty parameter. It is worth emphasizing the equivalence between eq. (1.8) and eq. (4.1) because the added term $(\rho/2)\|\mathcal{A}(X) - \mathbf{b}\|^2$ does not change the optimal value or the set of optimal points. We choose a predetermined ρ and fix it within our method. Consequently, we can tackle the augmented problem eq. (4.1) instead, whose Lagrangian form is:

$$\mathcal{L}(X, \mathbf{y}) = f_0(X) + f_1(X, \mathbf{y}), \quad (4.2)$$

where $f_0(X)$ is defined in eq. (1.8), $\text{dom } f_0 = \mathbb{S}_+^N$, $f_1(X, \mathbf{y}) \triangleq -(\mathcal{A}(X) - \mathbf{b})^T \mathbf{y} + \rho/2 \|\mathcal{A}(X) - \mathbf{b}\|^2$, and $\mathbf{y}$ is the dual variable. Equation (4.2) is also known as the augmented Lagrangian function of eq. (1.8). The function $\mathcal{L}(X, \mathbf{y})$ combines the objective function $f_0(X)$ and the influence of the linear constraints $f_1(X, \mathbf{y})$. It constitutes a hybrid function with which we can optimize the objective function while fulfilling the constraints. The penalty parameter ρ balances the weight of both parts. On the one hand, if ρ is too small, $\mathcal{L}(X, \mathbf{y})$ degenerates to the Lagrangian function of eq. (1.8) and the result can be infeasible, i.e., $\mathcal{A}(X) - \mathbf{b} = \mathbf{0}$ is not satisfied. On the other hand, a large ρ can also be problematic. The condition number of $\mathcal{L}(X, \mathbf{y})$ depends on the second-order gradient of differentiable terms in $\mathcal{L}(X, \mathbf{y})$ w.r.t. X . If ρ is too large, the condition number can be large and the resulting problem is not stable. In such case, the proposed method will diverge. Therefore, a value of ρ within a suitable range is important in this method. Fortunately, as we will show by numerical analysis in Section 4.4, a ρ of reasonable scale will lead us towards the primal-dual optimal pair in practice.

Since both the objective function and the equality constraints are convex in eq. (4.1), we can convert the problem of solving the log-determinant optimization problem eq. (1.8) into finding a saddle point of its augmented Lagrangian form eq. (4.2) according to the following saddle point theorem.

Theorem 4.1. [75, Proposition 5.1.6] *The pair $(X^*, \mathbf{y}^*)$ with $X^* \in \mathbb{S}_+^n$ is a primal-dual optimal solution pair if and only if the following relation holds:*

$$\mathcal{L}(X^*, \mathbf{y}) \leq \mathcal{L}(X^*, \mathbf{y}^*) \leq \mathcal{L}(X, \mathbf{y}^*) \quad \forall X \in \mathbb{S}_+^n, \mathbf{y} \in \mathbb{R}^M,$$

i.e., $(X^*, \mathbf{y}^*)$ is a saddle point of the augmented Lagrangian function $\mathcal{L}(X, \mathbf{y})$.

One simple and reliable approach is a subgradient method derived from the Uzawa iterations [76]. First we can find a subgradient of $\mathcal{L}(X, \mathbf{y})$ with respect to X and $\mathbf{y}$ satisfying:

$$\mathcal{L}_X(X, \mathbf{y}) \in \partial_X \mathcal{L}(X, \mathbf{y}), \quad \mathcal{L}_\mathbf{y}(X, \mathbf{y}) \in \partial_\mathbf{y} \mathcal{L}(X, \mathbf{y}). \quad (4.3)$$

One instance of the subgradient can be defined as:

$$\mathcal{L}_X(X, \mathbf{y}) = \underbrace{C - \mu_1 X^{-1} + \Lambda}_{\in \partial f_0(X)} - \underbrace{\mathcal{A}^T(\mathbf{y}) + \rho \mathcal{A}^T(\mathcal{A}(X) - \mathbf{b})}_{\in \partial_X f_1(X, \mathbf{y})}, \quad (4.4a)$$

$$\mathcal{L}_\mathbf{y}(X, \mathbf{y}) = \mathbf{b} - \mathcal{A}(X), \quad (4.4b)$$

where $\Lambda_{i,j} = \text{sgn}(X_{i,j})\mu_2$, $\text{sgn}(\cdot)$ is the sign operator. It is easy to verify that the definition eq. (4.4) satisfies eq. (4.3). The subgradient eq. (4.4b) is also the gradient of $\mathcal{L}(X, \mathbf{y})$ with respect to $\mathbf{y}$. We refer to it as the subgradient for the simplicity and uniformity of notation. Next, starting with an arbitrary feasible pair $(X^0, \mathbf{y}^0)$, the primal dual pair is updated as:

$$X^{t+1} = \mathbb{P}_{\mathbb{S}_+^n}(X^t - \eta^t \mathcal{L}_X(X^t, \mathbf{y}^t)), \quad (4.5a)$$

$$\mathbf{y}^{t+1} = \mathbf{y}^t + \eta^t \mathcal{L}_\mathbf{y}(X^t, \mathbf{y}^t), \quad (4.5b)$$

where $t \in \mathbb{R}_+$ is the iteration number, $\mathbb{P}_{\mathbb{S}_+^n}(\cdot)$ denotes the projection onto sets $\mathbb{S}_+^n$, $\eta^t > 0$ is the step size in iteration t , $\mathcal{L}_X(X^t, \mathbf{y}^t)$ and $\mathcal{L}_\mathbf{y}(X^t, \mathbf{y}^t)$ are the subgradients at point $(X^t, \mathbf{y}^t)$ defined as in eq. (4.4) respectively. By combining eq. (4.4) and eq. (4.5), it can be shown that the resulting subgradient method is guaranteed to converge towards a saddle point of $\mathcal{L}(X, \mathbf{y})$ [77].

However, there are two challenges in the subgradient algorithm. The subgradient in eq. (4.4a) requires matrix inverse of X . The associated time complexity is typically $\mathcal{O}(N^3)$. Moreover, it is not trivial to project a general N -dimensional symmetric matrix onto the positive definite cone $\mathbb{S}_+^N$ in eq. (4.5a). One possible solution is to replace the non-positive eigenvalues with a small positive constant. This requires full eigenvalue decompositions whose time complexity is also $\mathcal{O}(N^3)$. Such challenges hinder the scalability of the subgradient method in eq. (4.4) and eq. (4.5).

In order to bypass the cubic time complexity of computing $\mathcal{L}_X(X, \mathbf{y})$, we adopt stochastic subgradients when updating the primal variable X . Stochastic optimization [53] yields unbiased but efficient estimates of the exact subgradient, and iteratively updates X according to a realization of the estimate. In other words, we aim to find a stochastic function $g(X, \mathbf{y})$ satisfying the unbiasedness, i.e., $\mathbb{E}[g(X, \mathbf{y})] = \mathcal{L}_X(X, \mathbf{y})$, where $\mathbb{E}[\cdot]$ denotes expectation operation. It often takes much less time to evaluate one instance of the stochastic function $g(X, \mathbf{y})$ than computing the exact subgradient $\mathcal{L}_X(X, \mathbf{y})$. The primal variable is iteratively updated following realizations of $-g(X, \mathbf{y})$ with step size η_1^t in iteration t . The stochastic update rule can be written as:

$$X^{t+1} = \mathbb{P}_{\mathbb{S}_+^n}(X^t - \eta_1^t g(X^t, \mathbf{y}^t)), \quad (4.6a)$$

$$\mathbf{y}^{t+1} = \mathbf{y}^t + \eta_2^t \mathcal{L}_\mathbf{y}(X^t, \mathbf{y}^t), \quad (4.6b)$$

where η_1^t and η_2^t are the step size of X and $\mathbf{y}$ in iteration t . A similar update rule can be developed based on stochastic mirror-prox method [78]. The stochastic mirror-prox requires $\eta_1^t = \eta_2^t$ for its analysis. By contrast, we allow different step

size for the primal and dual variables so that we can adopt adaptive step sizes in our method. The resulting method is guaranteed to converge to a saddle point of $\mathcal{L}(X, \mathbf{y})$, as we will show in Section 4.2. The challenge lies in finding an efficient and unbiased stochastic function, which we will address in Section 4.1.2

Next, we propose an alternative for the projection step. The necessity of the projection arises from the positive definiteness of the X . If we can guarantee the positive definiteness of X in each update and initialize it with a positive definite matrix, the above projection step is no longer necessary. Therefore, we intend to guarantee the positive definiteness efficiently during the update, as we will explain in Section 4.1.3.

4.1.2 Stochastic Subgradient

As mentioned earlier, since the high computational complexity of computing $\mathcal{L}_X(X_t, \mathbf{y}_t)$ in eq. (4.4a) arises from the matrix inverse term $\{X^t\}^{-1}$, the problem boils down to constructing an efficient and unbiased estimate of $\{X^t\}^{-1}$. In order to achieve that, we regard X (with the superscript omitted for simplicity in this subsection) as the precision matrix of a zero-mean Gaussian distribution $\mathcal{N}(\mathbf{0}, X^{-1})$ and draw samples from this Gaussian distribution. If the samples are independent, the covariance matrix of the samples forms an unbiased estimate of the matrix inverse term [54].

In this section, we employ Gibbs sampling [55] to draw samples from $\mathcal{N}(\mathbf{0}, X^{-1})$. In general, given a multivariate distribution, in Gibbs sampling one periodically draws samples of each variable, conditioned on the most recent values of other variables. In our case, Gibbs sampling constructs a Markov chain with the stationary distribution being the targeted $\mathcal{N}(\mathbf{0}, X^{-1})$. Particularly, given the arbitrary initial sample $\mathbf{x}^{(0)}$, it samples the i -th component $x_i^{(k)}$ of the k -th sample ($1 \leq i \leq N$ and $1 \leq k \leq K$),

from the following conditional distribution:

$$p(x_i^{(k)} | x_1 = x_1^{(k)}, \dots, x_{i-1} = x_{i-1}^{(k)}, x_{i+1} = x_{i+1}^{(k-1)}, \dots, x_n = x_n^{(k-1)}) = \mathcal{N}\left(\frac{\sum_{j=1}^{i-1} X_{i,j} x_j^{(k)} + \sum_{j=i+1}^N X_{i,j} x_j^{(k-1)}}{X_{i,i}}, \frac{1}{X_{i,i}}\right), \quad (4.7)$$

where $(\sum_{j=1}^{i-1} X_{i,j} x_j^{(k)} + \sum_{j=i+1}^N X_{i,j} x_j^{(k-1)})/X_{i,i}$ is the conditional mean and $1/X_{i,i}$ is the conditional variance. In order to analyze the convergence of the sampling procedure, we introduce another interpretation of eq. (4.7): we obtain one sample $\mathbf{x}^{(k)}$ by cyclically repeating $i = 1, \dots, N$:

$$x_i^{(k)} \leftarrow -\frac{1}{X_{i,i}} \left(\sum_{j=1}^{i-1} X_{i,j} x_j^{(k)} + \sum_{j=i+1}^N X_{i,j} x_j^{(k-1)} \right) + \frac{v_i}{\sqrt{X_{i,i}}}, \quad (4.8)$$

where $v_i \sim \mathcal{N}(0, 1)$. After repeating the cycle K times, we attain the desired Gibbs samples $\mathbf{x}^{(1:K)}$.

If the Markov chain is stationary, the estimate of $\{X^t\}^{-1}$ can then be expressed as the sample covariance of the Gibbs samples:

$$\Sigma^t \triangleq \frac{1}{K} \sum_{k=1}^K \mathbf{x}_t^{(k)} \mathbf{x}_t^{(k)T}. \quad (4.9)$$

The stochastic subgradient w.r.t. X in iteration t can be written as:

$$g(X^t, \mathbf{y}^t) = \underbrace{C - \mu_1 \Sigma^t + \Lambda^t}_{g_0^t} - \underbrace{\mathcal{A}^T(\mathbf{y}^t) + \rho \mathcal{A}^T(\mathcal{A}(X^t) - \mathbf{b})}_{g_1^t}. \quad (4.10)$$

Together with the update eq. (4.6), we obtain the stochastic primal-dual method. Furthermore, we set the number of Gibbs samples K to be invariant w.r.t the dimension N or we choose $K \ll N$. In this way, it only takes $\mathcal{O}(N^2)$ arithmetic operations to draw the Gibbs samples and compute the stochastic subgradient eq. (4.10). By contrast, the time complexity of computing the exact subgradient is $\mathcal{O}(N^3)$.

4.1.3 Guarantee of the Positive Definiteness of X

In order to verify the definiteness of X^t in eq. (4.6a), we evaluate the smallest eigenvalue $\lambda_{\min}(X)$ of X^t via the Lanczos process [24] (see Section 3.1.2). The time complexity of this approach is only $\mathcal{O}(N^2)$. If the smallest eigenvalue is positive, we proceed to the next iteration; otherwise, we halve the step size η_1^t until the value is positive. Consequently, the projection step in eq. (4.6a) is no longer required.

We can summarize the update rules as:

$$X^{t+1} = X^t - \eta_1^t g(X^t, \mathbf{y}^t), \quad (4.11a)$$

$$\mathbf{y}^{t+1} = \mathbf{y}^t + \eta_2^t \mathcal{L}_{\mathbf{y}}(X^t, \mathbf{y}^t). \quad (4.11b)$$

Moreover, for simplicity we define a linearly decaying step size:

$$\eta_1^t = \frac{\varepsilon}{t+1} \text{ and } \eta_2^t = \frac{\rho}{G}, \quad (4.12)$$

where $\varepsilon > 0$ and $G > 1$ are constants to be specified, and ρ is the penalty parameter specified in eq. (4.1). In the next three sections, we analyze theoretical properties, discuss some implementation details, and provide numerical results for the proposed algorithm.

4.2 Theoretical Results

4.2.1 Assumptions and Preliminaries

In this section, we make several assumptions and briefly review some critical inequalities that are useful throughout our analysis.

Initially, we can normalize $\mathcal{A}(X) - \mathbf{b} = 0$ as:

$$A_m \leftarrow \frac{A_m}{\|A_m\|}, b_m \leftarrow \frac{b_m}{\|A_m\|},$$

where $m = 1, \dots, M$. Moreover, if there exist $\{\lambda_m \in \mathbb{R} | m = 1, 2, \dots, M\}$ such that $\sum_{m=1}^M \lambda_m A_m = 0$ holds nontrivially ($\lambda_m \neq 0$ at least for one m), we can analytically substitute the equality constraints to reduce M , the total number of constraints, until $\sum_{m=1}^M \lambda_m A_m = 0$ holds only trivially for the reduced M . Therefore, the following assumption is reasonable.

Assumption 4.1. *We assume the following properties w.r.t. $\mathcal{A}$:*

- $\|A_m\|_2 = 1, m = 1, \dots, M$,
- $\{A_m | m = 1, \dots, M\}$ are linearly independent.

Moreover, below we introduce two standard assumptions for the log-determinant problem [37] and Gibbs sampler [54] respectively.

Assumption 4.2. *There exists at least one primal-dual optimal pair $(X^*, \mathbf{y}^*)$ satisfying the Karush-Kuhn-Tucker (KKT) conditions:*

$$X^* \in \mathbb{S}_+^n, \tag{4.13a}$$

$$\mathcal{A}(X^*) - \mathbf{b} = \mathbf{0}, \tag{4.13b}$$

$$\mathbf{0} \in \partial f_0(X^*) - \mathcal{A}^T(\mathbf{y}^*), \tag{4.13c}$$

where $\partial f_0(X^*)$ denotes the set of subgradients of $f_0(X)$ at point X^* .

Assumption 4.3. *The Markov chain constructed by Gibbs sampling eq. (4.8) is stationary.*

The Gibbs sampling update rule in eq. (4.8) can be regarded as a linear combination of other variables with additional Gaussian noise. The update rule can be

equivalently expressed as [62]:

$$\mathbf{x}_t^{(k)} = -D_t^{-1}U_t^T \mathbf{x}_t^{(k)} - D_t^{-1}U_t \mathbf{x}_t^{(k-1)} + D_t^{-\frac{1}{2}} \mathbf{v}^{(k)}, \quad (4.14)$$

where $\mathbf{v}^{(k)} \sim \mathcal{N}(\mathbf{0}, I)$, D_t is the diagonal of X^t , and U_t is the strictly upper triangular part. It has been shown in [63] that the covariance of $\mathbf{x}_t^k$ converges to $\{X^t\}^{-1}$ at the rate of $\{\rho(B_t)\}^2$, which is the square of the largest eigenvalue of $B_t \triangleq -(D_t + U_t^T)^{-1}U_t$. Moreover, it has also been proven in [64] that $\rho(B_t) < 1$ for arbitrary symmetric positive definite X^t . In other words, the distribution of $\mathbf{y}_t^{(k)}$ converges to $\mathcal{N}(\mathbf{0}, X^t)$ at the rate of $\{\rho(B_t)\}^2$ as k increases. On the other hand, $\lim_{t \rightarrow \infty} \eta_1^t = 0$ implies that the difference between $\mathcal{N}(\mathbf{0}, X^t)$ and $\mathcal{N}(\mathbf{0}, X^{t+1})$ shrinks as t increases. If we adopt the last sample in the t -th iteration as the initial sample of the $t+1$ -th iteration, the distribution of the sample $\mathbf{y}_{t+1}^0 \triangleq \mathbf{y}_t^K$ becomes increasingly similar to $\mathcal{N}(\mathbf{0}, X^{t+1})$. Therefore, Assumption 4.3 holds asymptotically as t increases.

4.2.2 Bounds on the Stochastic Subgradient

In this section, we establish bounds on the primal-dual optimal pair $(X^*, \mathbf{y}^*)$ and the second-order moments of the stochastic subgradient. Next we analyze the bound of the primal dual pair $(X^t, \mathbf{y}^t)$ generated by eq. (4.11).

We define the set of Lagrangian multipliers and state the properties of the Lagrangian multipliers. By Theorem 4.1, we know there is at least one primal optimal X^* for eq. (1.8). We define the set of Lagrange multipliers associated with X^* by $\mathcal{M}(X^*) \triangleq \{\mathbf{y} \in \mathbb{R}^M | \mathbf{0} \in \partial f_0(X^*) - \mathcal{A}^T(\mathbf{y})\}$. We state the following lemma about the existence of $\mathcal{M}(X^*)$.

Lemma 4.1. [79, Theorem 2.3.2] *Let X^* be a given primal optimum point in eq. (1.8). If*

- *the matrices A_m , $m = 1, \dots, M$ are linearly independent,*

- $\exists X \in \mathbb{S}_+^N$ such that $\mathcal{A}(X) - \mathbf{b} = \mathbf{0}$,

then $\mathcal{M}(X^*)$ is a nonempty, compact, and bounded convex set.

The conditions in Lemma 4.1 are fulfilled by Assumption 4.1 and Assumption 4.2. Moreover, we have the following property of the set of Lagrangian multipliers.

Lemma 4.2. [79, Proposition 3.1.1] *Let X^* and $\tilde{X}^*$ be two different solutions of the constrained optimization problem eq. (1.8), and let $\mathcal{M}(X^*)$ and $\mathcal{M}(\tilde{X}^*)$ be their associated set of Lagrange multipliers, then $\mathcal{M}(X^*) = \mathcal{M}(\tilde{X}^*)$.*

In other words, the set $\mathcal{M}(X^*)$ is actually identical for all primal optimal points of the problem eq. (1.8) under Assumptions 4.1 and 4.2. We denote the set of Lagrangian multipliers of eq. (1.8) by $\mathcal{M}$. Combining Lemmas 4.1 and 4.2, we obtain the following lemma about the boundedness of $\mathcal{M}$.

Lemma 4.3. *Let $\mathbf{y}^* \in \mathbb{R}^M$ such that there $\exists X^* \in \mathbb{S}_+^N$ together satisfying the KKT conditions eq. (4.13). Then $\mathbf{y}^*$ is bounded in a compact convex set, i.e., $\exists G_{\mathbf{y}} > 0$ s.t. $\|\mathbf{y}^*\| \leq \|\mathbf{y}^*\|_1 < G_{\mathbf{y}}$.*

Remark. *We can derive the following inequality regarding $\|\mathcal{A}^T(\mathbf{y})\|$ and $\|\mathbf{y}\|_1$:*

$$\begin{aligned} \|\mathcal{A}^T(\mathbf{y})\| &= \left\| \sum_{m=1}^M A_m y_m \right\| \leq \sum_m \|A_m y_m\| \\ &\leq \sum_m \|A_m\| |y_m| = \sum_m |y_m| = \|\mathbf{y}\|_1. \end{aligned} \tag{4.15}$$

Consequently, Lemma 4.3 suggests an upper bound for $\|\mathcal{A}^T(\mathbf{y}^*)\| \leq G_{\mathbf{y}}$, i.e., it yields a bound for the primal optimal point X^* .

Lemma 4.4. *For every $\mu_1, \mu_2 > 0$, the solution of problem eq. (1.8) X^* is unique*

and is contained in a convex cone, i.e., $X^* \in \mathcal{X}^+ \triangleq \{X \mid \alpha I \preceq X \preceq \beta I\}$, where:

$$\alpha = \frac{\mu_1}{\|C\|_2 + G_{\mathbf{y}} + N\mu_2}, \quad (4.16a)$$

$$\beta = \frac{1}{\mu_2} (\|\mathbf{b}\|G_{\mathbf{y}} + N\mu_1). \quad (4.16b)$$

Proof. By eq. (1.15), we can express X^* as $((C - \mathcal{A}(\mathbf{y}^*) + W^*)/\mu_1)^{-1}$, where $\|W^*\|_\infty < \mu_2$, and consequently it follows:

$$\begin{aligned} \|C - \mathcal{A}(\mathbf{y}^*) + W\|_2 &\leq \|C\|_2 + \|\mathcal{A}(\mathbf{y}^*)\|_2 + \|W^*\|_2 \\ &\leq \|C\|_2 + \sum_{m=1}^M \|y_m^* A_m\|_2 + N\|W^*\|_\infty \\ &\leq \|C\|_2 + \sum_{m=1}^M |y_m^*| + N\mu_2 \\ &\leq \|C\|_2 + G_{\mathbf{y}} + N\mu_2. \end{aligned} \quad (4.17)$$

Therefore, we have the lower bound on the eigenvalue of X^* :

$$\|X^*\|_2 \geq \frac{\mu_1}{\|C\|_2 + G_{\mathbf{y}} + N\mu_2}. \quad (4.18)$$

At the optimum, the primal dual gap of eq. (1.8) and eq. (1.15) is zero:

$$\text{tr}(CX^*) - \mu_1 \log \det(X^*) + \mu_2 \|X^*\|_1 - \mathbf{b}^T \mathbf{y}^* + \mu_1 \log \det(X^*) - N\mu_1 = 0, \quad (4.19)$$

from which we have:

$$\|X^*\|_1 + \frac{\text{tr}(CX^*)}{\mu_2} \leq \frac{1}{\mu_2} (\|\mathbf{b}\| \|\mathbf{y}^*\| + N\mu_1). \quad (4.20)$$

Since $\|X^*\|_2 \leq \|X^*\| \leq \|X^*\|_1$ and $\text{tr}(CX^*) \geq 0$ due to $C \succeq 0$ and $X^* \succeq 0$, we have the upper bound for $\|X^*\|_2$. $\square$

Lemma 4.4 implicitly gives a bound on the primal optimal point X^* in terms of the Frobenius norm considering the following relationship between ℓ_2 norm and

Frobenius norm:

$$\|X\|_2 \leq \|X\| \leq \sqrt{N}\|X\|_2. \quad (4.21)$$

Therefore, for $X \in \mathcal{X}^+$, $\|X\| \leq \sqrt{N}\beta$. From Lemmas 4.3 and 4.4 we can derive a bound on the primal-dual optimal pair $(X^*, \mathbf{y}^*)$. We proceed to analyze the sequential primal-dual pair $(X^t, \mathbf{y}^t)$. In view of Lemma 4.4, we know eq. (1.8) is equivalent to the following problem:

$$\begin{aligned} \min \quad & f_0(X) \triangleq \text{tr}(CX) - \mu_1 \log \det(X) + \mu_2 \|X\|_1 \\ \text{s.t.} \quad & \mathcal{A}(X) - \mathbf{b} = \mathbf{0}, \quad X \in \mathcal{X}^+, \end{aligned} \quad (4.22)$$

where $\mathcal{X}^+$ is a bounded compact convex set defined in Lemma 4.4. The extreme eigenvalues $\lambda_{\min}(X)$ and $\lambda_{\max}(X)$ can be evaluated via the implicitly restarted Lanczos method reviewed in Section 3.1. We can determine proper step sizes by backtrack line searching so that the updated X is contained in $\mathcal{X}^+$. After specifying the bound on X^t , we establish the bound of $\{\|\mathbf{y}^t\| | t \in \mathcal{N}\}$ with the following assumption.

Assumption 4.4. *We assume the $\{\|\mathbf{y}^t\| | t \in \mathcal{N}\}$ generated in eq. (4.11) is upper bounded, i.e., $\forall t \in \mathcal{N}$,*

$$\mathbb{E}[\|\mathbf{y}^t\|^2] \leq v^2. \quad (4.23)$$

In other words, we require that the dual variable is bounded in our method. Similar assumptions are made in other gradient-based primal-dual method such as in [78, 80]. Following the scheme in [81], we will elaborate in Appendix A on a different approach to convergence analysis without Assumption 4.4, where we restrict the step size of the dual variable.

Next, we prove that the second-order moments $\|g(X, \mathbf{y})\|^2$ of the stochastic sub-gradient eq. (4.10) w.r.t. X are finite. First we recap the definition of $g(X, \mathbf{y})$ in iteration t (cf. eqs. (4.9) and (4.10)):

$$\Sigma^t \triangleq \frac{1}{K} \sum_{k=1}^K \mathbf{x}_t^{(k)} \mathbf{x}_t^{(k)T},$$

$$g(X^t, \mathbf{y}^t) \triangleq C - \mu_1 \Sigma^t + \Lambda^t - \mathcal{A}^T(\mathbf{y}^t) + \rho \mathcal{A}^T(\mathcal{A}(X^t) - \mathbf{b}).$$

The term Σ^t dominates the variance of the subgradient. We proceed to analyze the second-order moment of the stochastic term Σ^t . Given Assumption 4.3, the distribution of $\sum_{k=1}^K \mathbf{x}_t^{(k)} \mathbf{x}_t^{(k)T}$ is an N -variate singular Wishart distribution denoted by $\mathcal{W}_N(B^t, K)$, where B^t is an N -dimensional positive definite matrix satisfying $B^t X^t = I$, and K is the degree of freedom. To analyze its second-order moments, we state the following lemma on singular Wishart distributions:

Lemma 4.5. *If E follows an N -dimensional singular Wishart distribution $\mathcal{W}_N(B, K)$ with degree of freedom $K \ll N$ and positive definite scale matrix B , then $\mathbb{E}[\|E\|^2] \leq K(2 + N)\|B\|^2$.*

Proof. Suppose that $E \sim \mathcal{W}_N(B, K)$ and $F \sim \mathcal{W}_N(B, N)$. It follows that $E + F \sim \mathcal{W}_N(B, N + K)$ (see Theorem 7 in [61]). According to the properties of Wishart distributions, we can obtain:

$$\begin{aligned} \mathbb{E}[(E + F)_{i,j}^2] &= \mathbb{V}(E + F)_{i,j} + \mathbb{E}[E + F]_{i,j}^2 \\ &= (N + K)(B_{i,j}^2 + B_{i,i}B_{j,j}) + (N + K)^2 B_{i,j}^2, \end{aligned} \tag{4.24}$$

for $i, j \in \{1, 2, \dots, N\}$. On the other hand, we can rewrite the left hand side of the above expression as:

$$\begin{aligned} \mathbb{E}[(E + F)_{i,j}^2] &= \mathbb{E}[E_{i,j}^2 + 2E_{i,j}F_{i,j} + F_{i,j}^2] \\ &= \mathbb{E}[E_{i,j}^2] + 2\mathbb{E}[E_{i,j}]\mathbb{E}[F_{i,j}] + \mathbb{E}[F_{i,j}^2], \end{aligned} \tag{4.25}$$

where $\mathbb{E}[E_{i,j}] = KB_{i,j}$, $\mathbb{E}[F_{i,j}] = NB_{i,j}$ and $\mathbb{E}[F_{i,j}^2] = N(B_{i,j}^2 + B_{i,i}B_{j,j}) + N^2 B_{i,j}^2$. By canceling the common terms on both sides of eq. (4.24) and eq. (4.25), we can have:

$$\mathbb{E}[E_{i,j}^2] = K(2B_{i,j}^2 + B_{i,i}B_{j,j}). \tag{4.26}$$

Furthermore, it follows from the definition of the norm that:

$$\begin{aligned}
\mathbb{E}[\|E\|^2] &= K \sum_{i,j} (2B_{i,j}^2 + B_{i,i}B_{j,j}) \\
&\leq K(2\|B\|^2 + \sum_i B_{i,i} \sum_j B_{j,j}) \\
&\leq K(2 + N)\|B\|^2.
\end{aligned} \tag{4.27}$$

□

If we restrict that $X^t \in \mathcal{X}^+$, we can find a uniform bound for $\|B^t\| \leq \sqrt{N}/\alpha, \forall t \in \mathcal{N}$. By combining the bounds of the set of multipliers in Lemma 4.3, primal variable in Lemma 4.4, and the bounds on $\mathbb{E}[\|\Sigma^t\|^2]$ in Lemma 4.5, we obtain the following lemma:

Lemma 4.6. $\exists \sigma, \varrho \in \mathbb{R}_+, \forall t$, given $\mathbf{y}^t \in \mathbb{R}^M$ and $X^t \in \mathcal{X}^+$, it holds that $E[\|g_0^t\|^2] < \sigma^2$, and $\mathbb{E}[\|g_1^t\|^2] \leq \varrho_{\mathbf{y}}^2$ with

$$\sigma \triangleq \|C\| + \frac{\mu_1}{\alpha} \sqrt{\frac{N(N+2)}{K}} + \mu_2 N, \tag{4.28a}$$

$$\varrho_{\mathbf{y}} \triangleq \sqrt{2\mathbb{E}[\|\mathbf{y}^t\|_1^2] + 4\rho^2(M\sqrt{N}\beta + \|\mathbf{b}\|_1)^2}. \tag{4.28b}$$

Proof. Since $\|\mathbf{y}\| \leq G_{\mathbf{y}}$ and $X \in \mathcal{X}^+$, we can write:

$$\begin{aligned}
E[\|g_0^t\|] &= \mathbb{E}[\|C - \mu_1 \Sigma^t + \Lambda^t\|] \\
&\leq \|C\| + \mu_1 \mathbb{E}[\|\Sigma^t\|] + \mathbb{E}[\|\Lambda^t\|] \\
&\leq \|C\| + \frac{\mu_1}{\alpha} \sqrt{\frac{N(N+2)}{K}} + \mu_2 N \triangleq \sigma.
\end{aligned} \tag{4.29}$$

Moreover, it follows:

$$\begin{aligned}
\mathbb{E}[\|g_1^t\|^2] &\leq 2\mathbb{E}[\|\mathcal{A}^T(\mathbf{y}^t)\|^2] + 2\mathbb{E}[\rho^2\|\mathcal{A}^T(\mathcal{A}(X^t) - \mathbf{b})\|^2] \\
&\leq 2\mathbb{E}[\|\mathbf{y}^t\|_1^2] + 2\rho^2\mathbb{E}[\|\mathcal{A}(X^t) - \mathbf{b}\|_1^2] \\
&\leq 2\mathbb{E}[\|\mathbf{y}^t\|_1^2] + 4\rho^2(\mathbb{E}[\|\mathcal{A}(X^t)\|_1] + \mathbb{E}[\|\mathbf{b}\|_1])^2 \\
&\leq 2\mathbb{E}[\|\mathbf{y}^t\|_1^2] + 4\rho^2(M\sqrt{N}\beta + \|\mathbf{b}\|_1)^2,
\end{aligned} \tag{4.30}$$

where we utilized the inequality $\|\mathcal{A}(X^t)\|_1 = \sum_{i=1}^M \langle A_i, X^t \rangle \leq \sum_{i=1}^M \|A_i\| \|X^t\| \leq M\sqrt{N}\beta$. $\square$

4.2.3 Definitions and Lemmas

In this section, we introduce definitions and establish useful theoretical properties that we will rely on in the convergence analysis in Section 4.2.4. Note that the term $\log \det(X)$ in eq. (1.8) is strongly convex in $\mathcal{X}^+$ and the other terms are convex in $\mathcal{X}^+$. Therefore, we have the following property regarding the optimization problem.

Theorem 4.2. *The objective function $f_0(X)$ in eq. (4.22) is ξ -strongly convex on $\mathcal{X}^+$ defined in Lemma 4.4 for a constant $\xi > 0$.*

Furthermore, we establish two lemmas:

Lemma 4.7. *Let $\mathbf{y} \in \mathbb{R}^M$ be an arbitrary vector, the following equality holds:*

$$\begin{aligned}
&-f_1(X^t, \mathbf{y}^t) - \langle \mathbf{y}, \mathcal{A}(X^t) - \mathbf{b} \rangle + \frac{1}{2\eta_2^t} \|\mathbf{y}^{t+1} - \mathbf{y}\|^2 \\
&= \frac{1}{2\eta_2^t} \|\mathbf{y}^t - \mathbf{y}\|^2 - \frac{1}{2\eta_2^t} \left(\frac{\rho}{\eta_2^t} - 1 \right) \|\mathbf{y}^{t+1} - \mathbf{y}^t\|^2.
\end{aligned} \tag{4.31}$$

Proof. Note that:

$$-f_1(X^t, \mathbf{y}^t) - \langle \mathbf{y}, \mathcal{A}(X^t) - \mathbf{b} \rangle = \langle \mathbf{y}^t - \mathbf{y}, \mathcal{A}(X^t) - \mathbf{b} \rangle - \frac{\rho}{2} \|\mathcal{A}(X^t) - \mathbf{b}\|^2. \tag{4.32}$$

Consequently eq. (4.31) is equivalent with:

$$\begin{aligned} & \langle \mathbf{y}^t - \mathbf{y}, \mathcal{A}(X^t) - \mathbf{b} \rangle - \frac{\rho}{2} \|\mathcal{A}(X^t) - \mathbf{b}\|^2 + \frac{1}{2\eta_2^t} \|\mathbf{y}^{t+1} - \mathbf{y}\|^2 \\ &= \frac{1}{2\eta_2^t} \|\mathbf{y}^t - \mathbf{y}\|^2 - \frac{1}{2\eta_2^t} \left(\frac{\rho}{\eta_2^t} - 1 \right) \|\mathbf{y}^{t+1} - \mathbf{y}^t\|^2. \end{aligned} \quad (4.33)$$

Because $\mathbf{y}^{t+1} = \mathbf{y}^t + \eta_2^t(\mathbf{b} - \mathcal{A}(X^t))$, we have:

$$\begin{aligned} \langle \mathbf{y}^t - \mathbf{y}, \mathcal{A}(X^t) - \mathbf{b} \rangle &= -\frac{1}{\eta_2^t} \langle \mathbf{y}^t - \mathbf{y}, \mathbf{y}^{t+1} - \mathbf{y}^t \rangle \\ &= -\frac{1}{2\eta_2^t} (\|\mathbf{y}^{t+1} - \mathbf{y}\|^2 - \|\mathbf{y}^t - \mathbf{y}\|^2 - \|\mathbf{y}^{t+1} - \mathbf{y}^t\|^2). \end{aligned} \quad (4.34)$$

Moreover, we can write the following two equalities:

$$\langle \mathbf{y}^t - \mathbf{y}, \mathcal{A}(X^t) - \mathbf{b} \rangle + \frac{1}{2\eta_2^t} \|\mathbf{y}^{t+1} - \mathbf{y}\|^2 = \frac{1}{2\eta_2^t} \|\mathbf{y}^t - \mathbf{y}\|^2 + \frac{1}{2\eta_2^t} \|\mathbf{y}^{t+1} - \mathbf{y}^t\|^2, \quad (4.35)$$

and:

$$-\frac{\rho}{2} \|\mathcal{A}(X^t) - \mathbf{b}\|^2 = -\frac{\rho}{2\{\eta_2^t\}^2} \|\mathbf{y}^{t+1} - \mathbf{y}^t\|^2. \quad (4.36)$$

By summing up eqs. (4.35) and (4.36), we obtain eq. (4.33). $\square$

By combining Lemmas 4.3, 4.4 and 4.6, and Theorem 4.2, we can present the following lemma concerning the performance of eq. (4.11) in one iteration:

Lemma 4.8. *Let $X \in \mathcal{X}^+$ and $\mathbf{y} \in \mathbb{R}^M$, then $\forall t \in \mathbb{N}$:*

$$\begin{aligned} & \mathbb{E}[f_0(X^t) - \langle \mathbf{y}, \mathcal{A}(X^t) - \mathbf{b} \rangle] + \frac{1}{2\eta_1^t} \mathbb{E}\|X^{t+1} - X\|^2 + \frac{1}{2\eta_2^t} \mathbb{E}\|\mathbf{y}^{t+1} - \mathbf{y}\|^2 \\ & \leq \mathbb{E}[f_0(X) + f_1(X, \mathbf{y}^t)] + \frac{1}{2} \left(\frac{1}{\eta_1^t} - \xi \right) \mathbb{E}\|X^t - X\|^2 + \frac{1}{2\eta_2^t} \mathbb{E}\|\mathbf{y}^t - \mathbf{y}\|^2 + \eta_1^t(\sigma^2 + \varrho_{\mathbf{y}}^2) \\ & \quad - \frac{1}{2\eta_2^t} \left(\frac{\rho}{\eta_2^t} - 1 \right) \|\mathbf{y}^{t+1} - \mathbf{y}^t\|^2 - \mathbb{E}[\langle X^t - X, g_0^t - \mathbb{E}[g_0^t] \rangle]. \end{aligned} \quad (4.37)$$

Proof. Combining eq. (4.10) and eq. (4.11a), we know:

$$\langle X^{t+1} - X, g_0^t + g_1^t + \frac{1}{\eta_1^t}(X^{t+1} - X^t) \rangle = 0. \quad (4.38)$$

We analyze the terms in eq. (4.38) one by one:

$$\begin{aligned} \langle X^{t+1} - X, g_0^t \rangle &= \langle X^{t+1} - X^t, g_0^t \rangle + \langle X^t - X, g_0^t \rangle \\ &= \langle X^{t+1} - X^t, g_0^t \rangle + \langle X^t - X, \mathbb{E}[g_0^t] \rangle + \langle X^t - X, g_0^t - \mathbb{E}[g_0^t] \rangle, \end{aligned} \quad (4.39)$$

where $\mathbb{E}[g_0^t] = C - \mu_1\{X^t\}^{-1} + \Lambda^t$. By the perturbed Young's inequality:

$$\langle X^{t+1} - X^t, g_0^t \rangle \geq -\frac{1}{4\eta_1^t} \|X^{t+1} - X^t\|^2 - \eta_1^t \|g_0^t\|^2. \quad (4.40)$$

From Definition 2.6, it follows:

$$f_0(X) - f_0(X^t) - \frac{\xi}{2} \|X^t - X\|^2 \geq \langle \mathbb{E}[g_0^t], X - X^t \rangle. \quad (4.41)$$

By substituting eq. (4.40) and eq. (4.41) into eq. (4.39), we obtain:

$$\begin{aligned} \langle X^{t+1} - X, g_0^t \rangle &\geq -\frac{1}{4\eta_1^t} \|X^{t+1} - X^t\|^2 - \eta_1^t \|g_0^t\|^2 \\ &\quad - f_0(X) + f_0(X^t) + \frac{\xi}{2} \|X^t - X\|^2 + \langle X^t - X, g_0^t - \mathbb{E}[g_0^t] \rangle. \end{aligned} \quad (4.42)$$

Taking expectation on both sides, we have:

$$\begin{aligned} \mathbb{E}[\langle X^{t+1} - X, g_0^t \rangle] &\geq -\frac{1}{4\eta_1^t} \mathbb{E}[\|X^{t+1} - X^t\|^2] - \eta_1^t \sigma^2 + \mathbb{E}[f_0(X^t) - f_0(X)] \\ &\quad + \frac{\xi}{2} \mathbb{E}[\|X^t - X\|^2] + \mathbb{E}[\langle X^t - X, g_0^t - \mathbb{E}[g_0^t] \rangle]. \end{aligned} \quad (4.43)$$

Similar results about $\mathbb{E}[\langle X^{t+1} - X, g_1^t \rangle]$ can be obtained by combining the inequality similar with eq. (4.39), by perturbed Young's inequality, and by convexity of $f_1(X, \mathbf{y})$ w.r.t. X :

$$\langle X^{t+1} - X, g_1^t \rangle = \langle X^{t+1} - X^t, g_1^t \rangle + \langle X^t - X, g_1^t \rangle, \quad (4.44)$$

$$\langle X^{t+1} - X^t, g_1^t \rangle \geq -\frac{1}{4\eta_1^t} \|X^{t+1} - X^t\|^2 - \eta_1^t \|g_1^t\|^2, \quad (4.45)$$

$$f_1(X, \mathbf{y}^t) - f_1(X^t, \mathbf{y}^t) \geq \langle \mathbb{E}[g_1^t], X - X^t \rangle. \quad (4.46)$$

Therefore we have:

$$\begin{aligned} \mathbb{E}[\langle X^{t+1} - X, g_1^t \rangle] &\geq -\frac{1}{4\eta_1^t} \mathbb{E}[\|X^{t+1} - X^t\|^2] - \eta_1^t \mathbb{E}[\|g_1^t\|^2] + \mathbb{E}[f_1(X^t, \mathbf{y}^t) - f_1(X, \mathbf{y}^t)] \\ &\geq -\frac{1}{4\eta_1^t} \mathbb{E}[\|X^{t+1} - X^t\|^2] - \eta_1^t \varrho_{\mathbf{y}}^2 + \mathbb{E}[f_1(X^t, \mathbf{y}^t) - f_1(X, \mathbf{y}^t)]. \end{aligned} \quad (4.47)$$

Moreover, we have the following equation:

$$\begin{aligned} \mathbb{E}[\langle X^{t+1} - X, \frac{1}{\eta_1^t} X^{t+1} - X^t \rangle] &= \frac{1}{2\eta_1^t} (\mathbb{E}[\|X^{t+1} - X\|^2] - \mathbb{E}[\|X^t - X\|^2] \\ &\quad + \mathbb{E}[\|X^{t+1} - X^t\|^2]). \end{aligned} \quad (4.48)$$

Summing up eqs. (4.43), (4.47) and (4.48), in view of eq. (4.38), we can obtain:

$$\begin{aligned} &\mathbb{E}[f_0(X^t) - f_0(X) + f_1(X^t, \mathbf{y}^t) - f_1(X, \mathbf{y}^t)] + \frac{1}{2\eta_1^t} \mathbb{E}[\|X^{t+1} - X\|^2] \\ &\leq \frac{1}{2} \left(\frac{1}{\eta_1^t} - \xi \right) \mathbb{E}[\|X^t - X\|^2] + \eta_1^t (\sigma^2 + \varrho_{\mathbf{y}}^2) - \mathbb{E}[\langle X^t - X, g_0^t - \mathbb{E}[g_0^t] \rangle]. \end{aligned} \quad (4.49)$$

Taking the expectation of eq. (4.31), we obtain:

$$\begin{aligned} &\mathbb{E}[-f_1(X^t, \mathbf{y}^t) - \langle \mathbf{y}, \mathcal{A}(X^t) - \mathbf{b} \rangle] + \frac{1}{2\eta_2^t} \mathbb{E}[\|\mathbf{y}^{t+1} - \mathbf{y}^t\|^2] \\ &= \frac{1}{2\eta_2^t} \mathbb{E}[\|\mathbf{y}^t - \mathbf{y}\|^2] - \frac{1}{2\eta_2^t} \left(\frac{\rho}{\eta_2^t} - 1 \right) \mathbb{E}[\|\mathbf{y}^{t+1} - \mathbf{y}^t\|^2]. \end{aligned} \quad (4.50)$$

Adding up eqs. (4.49) and (4.50) yields eq. (4.37). $\square$

4.2.4 Convergence Analysis

In this section, we analyze the convergence performance of the proposed method similarly as in [81]. We first define an auxiliary function as:

$$h(X; \tilde{X}, \tilde{\mathbf{y}}) \triangleq f_0(X) - f_0(\tilde{X}) - \langle \tilde{\mathbf{y}}, \mathcal{A}(X) - \mathbf{b} \rangle. \quad (4.51)$$

When $(\tilde{X}, \tilde{\mathbf{y}})$ is specified as the primal-dual optimal pair $(X^*, \mathbf{y}^*)$, we have the following lemma.

Lemma 4.9. *If $h(X; \tilde{X}, \tilde{\mathbf{y}})$ is defined as in eq. (4.51), when $(\tilde{X}, \tilde{\mathbf{y}}) = (X^*, \mathbf{y}^*)$, we have the following inequality*

$$h(X; X^*, \mathbf{y}^*) = f_0(X) - f_0(X^*) - \langle \mathbf{y}^*, \mathcal{A}(X) - \mathbf{b} \rangle \geq 0. \quad (4.52)$$

Proof. According to eq. (4.13c), we have the following relationship:

$$\mathcal{A}^T(\mathbf{y}^*) \in \partial f_0(X^*). \quad (4.53)$$

Therefore, the following inequality holds from the convexity of $f(X)$:

$$f_0(X) \geq f_0(X^*) + \langle \mathcal{A}^T(\mathbf{y}^*), X - X^* \rangle, \forall X \in \mathcal{X}^+. \quad (4.54)$$

Moreover, we can obtain the following relationship by the definition of $\mathcal{A}(\cdot)$ and $\mathcal{A}^T(\cdot)$:

$$\langle \mathcal{A}^T(\mathbf{y}^*), X - X^* \rangle = \langle \mathbf{y}^*, \mathcal{A}(X) - \mathcal{A}(X^*) \rangle. \quad (4.55)$$

By combining eq. (4.54), eq. (4.55), and $\mathcal{A}(X^*) - \mathbf{b} = \mathbf{0}$, we have validated eq. (4.52). $\square$

With the aid of definition eq. (4.51) and its property on the primal-dual pair eq. (4.52), we introduce the following lemma which will greatly facilitate the convergence anal-

ysis.

Lemma 4.10. *Let $\tilde{X} \in \mathcal{X}^+$, $\tilde{\mathbf{y}}$ be random chosen dual variable, and $\varepsilon_1, \varepsilon_2 > 0$ be scalars. If $\forall \tilde{X} \in \mathcal{X}^+$ and $\mathbf{y}$, it holds that*

$$\mathbb{E}[h(\tilde{X}; X^*, \mathbf{y})] \leq \varepsilon_1 + \varepsilon_2 \mathbb{E}\|\mathbf{y}\|^2, \quad (4.56)$$

then for any $(X^*, \mathbf{y}^*)$ satisfying eq. (4.13), we have the following associated properties:

$$\mathbb{E}[\|\mathcal{A}(\tilde{X}) - \mathbf{b}\|_1] \leq \varepsilon_1 + 2M\varepsilon_2 + 2\varepsilon_2 \|\mathbf{y}^*\|^2, \quad (4.57)$$

$$\mathbb{E}[|f_0(\tilde{X}) - f_0(X^*)|] \leq \max\left(\varepsilon_1, \frac{1}{2}(\varepsilon_1 + 9\varepsilon_2 \mathbb{E}\|\mathbf{y}^*\|^2)\right). \quad (4.58)$$

Proof. By the inequality eq. (4.52):

$$f_0(X) - f_0(X^*) \geq \langle \mathbf{y}^*, \mathcal{A}(X) - \mathbf{b} \rangle. \quad (4.59)$$

Setting $\mathbf{y} = -\text{sgn}(\mathcal{A}(X) - \mathbf{b}) + \mathbf{y}^*$ in eq. (4.56), we have:

$$\begin{aligned} & \mathbb{E}[h(X; X^*, -\text{sgn}(\mathcal{A}(X) - \mathbf{b}) + \mathbf{y}^*)] \\ &= \mathbb{E}[f_0(X) - f_0(X^*) - \langle -\text{sgn}(\mathcal{A}(X) - \mathbf{b}) + \mathbf{y}^*, \mathcal{A}(X) - \mathbf{b} \rangle] \\ &\geq \mathbb{E}[\langle \mathbf{y}^*, \mathcal{A}(X) - \mathbf{b} \rangle - \langle -\text{sgn}(\mathcal{A}(X) - \mathbf{b}) + \mathbf{y}^*, \mathcal{A}(X) - \mathbf{b} \rangle] \\ &= \mathbb{E}[\langle \text{sgn}(\mathcal{A}(X) - \mathbf{b}), \mathcal{A}(\tilde{X}) - \mathbf{b} \rangle] = \mathbb{E}[\|\mathcal{A}(X) - \mathbf{b}\|_1]. \end{aligned} \quad (4.60)$$

On the other hand, since $\|-\text{sgn}(\mathcal{A}(X) - \mathbf{b}) + \mathbf{y}^*\|^2 \leq 2\|\mathbf{1}\|^2 + 2\|\mathbf{y}^*\|^2$, Equation (4.57) follows directly.

Setting $\mathbf{y} = 3\mathbf{y}^*$ in eq. (4.56) yields:

$$\mathbb{E}[f_0(X) - f_0(X^*) - 3\langle \mathbf{y}^*, \mathcal{A}(X) - \mathbf{b} \rangle] \leq \varepsilon_1 + 9\varepsilon_2 \mathbb{E}\|\mathbf{y}^*\|^2. \quad (4.61)$$

By substituting eq. (4.59) into eq. (4.61), it follows:

$$-2\mathbb{E}[f_0(X) - f_0(X^*)] \leq \varepsilon_1 + 9\varepsilon_2\mathbb{E}\|\mathbf{y}^*\|^2. \quad (4.62)$$

Moreover, setting $\mathbf{y} = \mathbf{0}$ in eq. (4.56), we have

$$\mathbb{E}[f_0(X) - f_0(X^*)] \leq \varepsilon_1. \quad (4.63)$$

From eq. (4.62) and eq. (4.63), we can obtain:

$$\mathbb{E}[|f_0(X) - f_0(X^*)|] \leq \max\left(\varepsilon_1, \frac{1}{2}(\varepsilon_1 + 9\varepsilon_2\mathbb{E}\|\mathbf{y}^*\|^2)\right), \quad (4.64)$$

from which eq. (4.58) is a consequence. $\square$

To establish the convergence rate of the proposed algorithm, we recall the definition of linearly decaying step eq. (4.12):

$$\eta_1^t = \frac{\varepsilon}{t+1}, \eta_2^t = \frac{\rho}{G},$$

where $G > 1$. We can analyze the convergence performance of eq. (4.11) with the step size scheme eq. (4.12) by the following theorem.

Theorem 4.3. *Given Assumptions 4.1 to 4.4, if the step size is specified as in eq. (4.12) and $(X^t, \mathbf{y}^t)$ is the estimate pair from eq. (4.11), we have:*

$$\mathbb{E}\|X^{t+1} - X^*\|^2 \leq \frac{2\varepsilon}{t+1} \left(\Phi(X^*) + \log(t+1)\omega + \frac{G}{2\rho}\mathbb{E}\|\mathbf{y}^*\|^2 \right), \quad (4.65)$$

where

$$\Phi(X) \triangleq \frac{1}{2} \left(\frac{1}{\eta_1^0} - \xi \right) \mathbb{E}\|X^0 - X\|^2 \quad (4.66)$$

$$\omega = \varepsilon(\sigma^2 + 2v^2 + 4\rho^2(M\sqrt{N}\beta + \|\mathbf{b}\|_1)^2). \quad (4.67)$$

Moreover, we can evaluate the objective value and the feasibility of the averaged estimate $\bar{X}^t = 1/(t+1) \sum_{i=0}^t X^i$:

$$\mathbb{E}[\|\mathcal{A}(\bar{X}^t) - \mathbf{b}\|_1] \leq \frac{1}{t+1} (\Phi(X^*) + \log(t+1)\omega + \frac{GM}{\rho} + \frac{G}{\rho} \|\mathbf{y}^*\|^2), \quad (4.68)$$

and:

$$\begin{aligned} \mathbb{E}[|f_0(X^t) - f_0(X^*)|] &\leq \frac{1}{t+1} \max \left(\Phi(X^*) + \log(t+1)\omega, \right. \\ &\quad \left. \frac{1}{2} (\Phi(X^*) + \log(t+1)\omega + \frac{9G}{4\rho} \mathbb{E}\|\mathbf{y}^*\|^2) \right). \end{aligned} \quad (4.69)$$

Proof. We set $X = X^*$ in eq. (4.37) and sum it up from $t = 0$ to $\bar{t}$:

$$\begin{aligned} &\sum_{t=0}^{\bar{t}} \mathbb{E}[f_0(X^t) - \langle \mathbf{y}, \mathcal{A}(X^t) - \mathbf{b} \rangle] + \frac{1}{2\eta_1^t} \mathbb{E}\|X^{\bar{t}+1} - X\|^2 + \frac{1}{2\eta_2^t} \mathbb{E}\|\mathbf{y}^{\bar{t}+1} - \mathbf{y}\|^2 \\ &\leq \sum_{t=0}^{\bar{t}} \mathbb{E}[f_0(X^*) + f_1(X^*, \mathbf{y}^t)] + \frac{1}{2} \left(\frac{1}{\eta_1^0} - \xi \right) \mathbb{E}\|X^0 - X^*\|^2 + \frac{1}{2\eta_2^0} \mathbb{E}\|\mathbf{y}^0 - \mathbf{y}\|^2 \\ &\quad + \sum_{t=0}^{\bar{t}} \eta_1^t (\sigma^2 + 2\mathbb{E}\|\mathbf{y}^t\|_1^2) + 4\rho^2 (M\sqrt{N}\beta + \|\mathbf{b}\|_1)^2 - \sum_{t=0}^{\bar{t}} \frac{1}{2\eta_2^t} \left(\frac{\rho}{\eta_2^t} - 1 \right) \|\mathbf{y}^{t+1} - \mathbf{y}^t\|^2 \\ &\quad - \sum_{t=0}^{\bar{t}} \mathbb{E}[\langle X^t - X^*, g_0^t - \mathbb{E}[g_0^t] \rangle]. \end{aligned} \quad (4.70)$$

Since $\mathbf{y}^0 = \mathbf{0}$, $\eta_2^t \leq \rho$, $\mathbb{E}[\langle X^t - X^*, g_0^t - \mathbb{E}[g_0^t] \rangle] = 0$, and

$$\sum_{i=0}^{\bar{t}} \frac{1}{i+1} \leq \int_1^{\bar{t}+1} \frac{1}{x} dx \leq \log(\bar{t}+1), \quad (4.71)$$

we have the following upper bound for eq. (4.70):

$$\begin{aligned} &\sum_{t=0}^{\bar{t}} \mathbb{E}[f_0(X^t) - f_0(X^*) - \langle \mathbf{y}, \mathcal{A}(X^t) - \mathbf{b} \rangle] + \frac{1}{2\eta_1^t} \mathbb{E}\|X^{\bar{t}+1} - X\|^2 + \frac{1}{2\eta_2^t} \mathbb{E}\|\mathbf{y}^{\bar{t}+1} - \mathbf{y}\|^2 \\ &\leq \Phi(X^*) + \log(t+1)\omega + \frac{1}{2\eta_2^0} \mathbb{E}\|\mathbf{y}\|^2. \end{aligned} \quad (4.72)$$

We can also write:

$$\frac{1}{2\eta_1^{\bar{t}}}\mathbb{E}\|X^{\bar{t}+1} - X\|^2 \geq 0, \quad (4.73)$$

$$\frac{1}{2\eta_2^{\bar{t}}}\mathbb{E}\|\mathbf{y}^{\bar{t}+1} - \mathbf{y}^*\|^2 \geq 0, \quad (4.74)$$

$$\sum_{t=0}^{\bar{t}} \mathbb{E}[f_0(X^t) - f(X^*) - \langle \mathbf{y}^*, \mathcal{A}(X^t) - \mathbf{b} \rangle] \geq 0, \quad (4.75)$$

where the last inequality holds according to eq. (4.52). If we choose $\mathbf{y} = \mathbf{y}^*$ and drop the first and third term on the left-hand side in eq. (4.72), we obtain:

$$\mathbb{E}\|X^{\bar{t}+1} - X\|^2 \leq \frac{2\varepsilon}{\bar{t}+1}(\Phi(X^*) + \log(\bar{t}+1)\omega + \frac{G}{2\rho}\mathbb{E}\|\mathbf{y}^*\|^2), \quad (4.76)$$

from which follows eq. (4.65). Furthermore, we drop the second and the third term on the left-hand side of eq. (4.72) (without assuming $\mathbf{y} = \mathbf{y}^*$), leading to the following:

$$\begin{aligned} & \mathbb{E}\left[\sum_{t=0}^{\bar{t}} f_0(X^t) - (\bar{t}+1)f_0(X^*) - \sum_{t=0}^{\bar{t}} \langle \mathbf{y}, \mathcal{A}(X^t) - \mathbf{b} \rangle\right] \\ & \leq \Phi(X^*) + \log(\bar{t}+1)\omega + \frac{G}{2\rho}\mathbb{E}\|\mathbf{y}\|^2. \end{aligned} \quad (4.77)$$

Since $f_0(X)$ is convex and $\mathcal{A}(X)$ is linear, eq. (4.77) implies:

$$\mathbb{E}[f_0(\bar{X}^{\bar{t}}) - f_0(X^*) - \langle \mathbf{y}, \mathcal{A}(\bar{X}^{\bar{t}}) - \mathbf{b} \rangle] \leq \frac{\Phi(X^*) + \log(\bar{t}+1)\omega}{\bar{t}+1} + \frac{G}{(\bar{t}+1)2\rho}\mathbb{E}\|\mathbf{y}\|^2, \quad (4.78)$$

where $\bar{X}^{\bar{t}} \triangleq 1/(\bar{t}+1) \sum_{t=0}^{\bar{t}} X^t$. According to the definition eq. (4.51):

$$\mathbb{E}[h(\bar{X}^{\bar{t}}; X^*, \mathbf{y})] \leq \frac{\Phi(X^*) + \log(\bar{t}+1)\omega}{\bar{t}+1} + \frac{G}{(\bar{t}+1)2\rho}\mathbb{E}\|\mathbf{y}\|^2. \quad (4.79)$$

According to Lemma 4.10, we have the following two conclusions with respect to

the feasibility:

$$\mathbb{E}[\|\mathcal{A}(\bar{X}^{\bar{t}}) - \mathbf{b}\|_1] \leq \frac{1}{\bar{t} + 1} (\Phi(X^*) + \log(\bar{t} + 1)\omega + \frac{GM}{\rho} + \frac{G}{\rho} \|\mathbf{y}^*\|^2), \quad (4.80)$$

and with respect to the objective value:

$$\begin{aligned} \mathbb{E}[|f_0(X^{\bar{t}}) - f_0(X^*)|] &\leq \frac{1}{\bar{t} + 1} \max \left(\Phi(X^*) + \log(\bar{t} + 1)\omega, \right. \\ &\quad \left. \frac{1}{2} (\Phi(X^*) + \log(\bar{t} + 1)\omega + \frac{9G}{4\rho} \mathbb{E}\|\mathbf{y}^*\|^2) \right). \end{aligned} \quad (4.81)$$

This concludes the proof for Theorem 4.3. $\square$

4.3 Implementation

In Section 4.2, we have established the $\log(t)/t$ convergence rate for the specific step size eq. (4.12) under the conditions of Assumptions 4.1 to 4.4. However, as in other stochastic algorithms, the performance of the algorithm depends on the choice of step size. We introduce an adaptive step size and specify some other implementation details in this section.

4.3.1 Initialization

First we explain how to initiate the Gibbs sampling in each iteration. When $t = 0$, the subgradient can be computed efficiently. When $t = 1$, we randomly generate the initial sample $x^{(0)} \sim \mathcal{N}(\mathbf{0}, I)$. If $t > 1$, we adopt the latest Gibbs sample in last iteration as the first Gibbs sample in t -th iteration. In other words, if we denote $\mathbf{x}_t^{(k)}$ as the k -th Gibbs sample in t -th iteration, we have:

$$\mathbf{x}_t^{(1)} \sim \mathcal{N}(\mathbf{0}, I), t = 0, \quad (4.82a)$$

$$\mathbf{x}_t^{(1)} = \mathbf{x}_{t-1}^{(K)}, t \geq 1. \quad (4.82b)$$

As discussed in Section 4.2.1, when the step size η_1^t decreases to 0 as t increases, the distinction between $\mathcal{N}(\mathbf{0}, \{X^t\}^{-1})$ and $\mathcal{N}(\mathbf{0}, \{X^{t+1}\}^{-1})$ will be smaller. $\mathbf{x}_t^{(K)}$ constitutes a favorable initial sample for $\mathcal{N}(\mathbf{0}, \{X^t\}^{-1})$.

Second, we have defined a closed compact convex set $\mathcal{X}^+$ which contains the primal optimal solution in Lemma 4.4. However, the values of α and β depend on X^* and $\mathbf{y}^*$ in Assumption 4.2. In practice, it is impossible to compute an exact instance of α and β . Therefore, we only restrict the smallest eigenvalue of X to be positive, i.e., X is positive definite. In practice, we found this to be sufficient to ensure convergence.

Third, we define the initialization of the primal variable X^0 and dual variable $\mathbf{y}^0$:

$$X^0 = I; \mathbf{y}^0 = \mathbf{0}. \quad (4.83)$$

4.3.2 Adaptive Step Size

The subgradient eq. (4.4) can be evaluated directly. To specify the adaptive step size, we initiate $\tilde{\mathbb{E}}[g]^0$ and $\tilde{\mathbb{E}}[g^2]^0$ as:

$$\tilde{\mathbb{E}}[g]^0 \triangleq \mathcal{L}_X(X^0, \mathbf{y}^0), \quad (4.84a)$$

$$\tilde{\mathbb{E}}[g^2]^0 \triangleq \mathcal{L}_Y(X^0, \mathbf{y}^0)^2, \quad (4.84b)$$

where the square operation is element-wise. With the above initialization, we proceed to define the step size of the primal and dual variable as:

$$\eta_1^t = \eta^t \frac{\|\tilde{\mathbb{E}}[g]^t\|^2}{\|\tilde{\mathbb{E}}[g^2]^t\|^2}, \eta_2^t = \frac{\rho}{G}, \quad (4.85a)$$

$$\tilde{\mathbb{E}}[g]^t \triangleq \theta \tilde{\mathbb{E}}[g]^{t-1} + (1 - \theta)g(X^t, \mathbf{y}^t), \quad (4.85b)$$

$$\tilde{\mathbb{E}}[g^2]^t \triangleq \theta \tilde{\mathbb{E}}[g^2]^{t-1} + (1 - \theta)g(X^t, \mathbf{y}^t)^2, \quad (4.85c)$$

where $\eta, \theta \in (0, 1)$, $t, G \geq 1$, and $g(X^t, \mathbf{y}^t)^2$ is element-wise square. On close inspection of eq. (4.85), $\tilde{\mathbb{E}}[g]^t$ and $\tilde{\mathbb{E}}[g^2]^t$ are approximately the first-order and second-order moments of the stochastic subgradient eq. (4.10). θ is the weight parameters to smoothen the estimates of $\tilde{\mathbb{E}}[g]^t$ and $\tilde{\mathbb{E}}[g^2]^t$. η is an exponentially decaying factor mitigating the influence of the variance. For the stochastic subgradient $g(X^t, \mathbf{y}^t)$ (abbr. g), we can decompose its second-order moments $\mathbb{E}[g^2]$ as below:

$$\mathbb{E}[g^2] = \mathbb{E}[g]^2 + \mathbb{V}[g]. \quad (4.86)$$

If the variance $\mathbb{V}[g]$ is large, η_1^t will be accordingly smaller. Consequently, the algorithm employs a small step size and it avoids the risk of a large step towards a wrong direction. When the variance is small and we have confidence in the stochastic gradient, the step size setting eq. (4.85a) will take a large step when updating the primal variable, thereby accelerating the convergence. Hence it accelerates the convergence. Moreover, the step size of the primal variable in eq. (4.85) decays exponentially due to the factor η^t . This exponential decrease of step size can mitigate the increasing variance for large dimensions, while also accelerating convergence. We typically set $\theta = 0.99$, $\eta = 0.9995$.

4.3.3 Termination Criterion

Finally, we shall define the convergence criteria and decide the termination condition accordingly. It is not affordable to evaluate the objective value or duality gap in our algorithm. Instead, the termination criterion is defined based on element-wise difference of the primal variable. Concretely, we average the estimates of X every $\bar{t}$ iterations as $\tilde{X}^k \triangleq 1/\bar{t} \sum_{i=0}^{\bar{t}-1} X^{k\bar{t}+i}$, where $k \in \mathbb{N}$, and halt the algorithm if the following condition is satisfied:

$$\|\tilde{X}^k - \tilde{X}^{k-1}\| < \vartheta, \quad (4.87)$$

Algorithm 3 Adaptive Primal-Dual Stochastic subGradient Descent (PDSGD-ad)
For log-determinant optimization

Input: Matrix C , $\mathcal{A}$, $\mathbf{b}$, penalty parameters μ_1, μ_2, ρ , number of Gibbs samples M , decaying factor η , updating weight θ , termination tolerance ϑ .

Output: Primal optimum X^* and dual optimum $\mathbf{y}^*$.

```

1: Initialize  $t \leftarrow 0$  and the first guess as a diagonal matrix  $X^0$  and zeros vector  $\mathbf{y}^0$  eq. (4.83);
2: while  $t < \text{Maxiter}$  do
3:   Generate  $\mathbf{x}^{(1:K)}$  via Gibbs sampling eq. (4.8);
4:   Compute the subgradients eq. (4.4b) and eq. (4.10), and the step size  $\eta_1^t, \eta_2^t$  eq. (4.85);
5:   while TRUE do
6:     Compute the updated  $X^{t+1}$  and  $\mathbf{y}^{t+1}$  as in eq. (4.11);
7:     Compute the smallest eigenvalue  $\lambda_{\min}$  of  $X^{t+1}$  via the IRL method;
8:     if  $\lambda_{\min} > 0$  then
9:        $X_{t+1} \leftarrow X_{\text{new}}$ ;
10:      break;
11:    else
12:      halve the step size  $\eta_1^t$ ;
13:    end if
14:    if Converged then
15:       $X^* \leftarrow X^{t+1}, \mathbf{y}^* \leftarrow \mathbf{y}^{t+1}$ ;
16:      break;
17:    else
18:       $t \leftarrow t + 1$ ;
19:    end if
20:  end while
21: end while

```

where ϑ is a small constant controlling the accuracy. Typically, we can set $\bar{t} = 10$ and $\vartheta = 0.001$ in our algorithm. We summarize the resulting adaptive Primal-Dual Stochastic subGradient Descent (PDSGD-ad) method in Algorithm 3.

4.4 Numerical Results

In this section, we verify the convergence of the proposed Algorithm 3. We also compare Algorithm 3 with adaptive step size against benchmark methods LogdetPPA [35, 36], ANS [7], and ASPG [7]. We follow the tolerance settings in [35] to ensure sim-

ilar accuracy of the benchmark methods. All the computations are performed on 64-bit OS Windows server 2012 with two Intel Xeon(R) CPU E5-2690 v2 @3.00GHz processors with 32.0 GB RAM.

4.4.1 Synthetic Data

We generate the random precision matrix as follows. Given the dimension N , we determine the nonzero pattern of X by randomly selecting $0.1N(N-1)/2$ entries from the lower triangular part X_L , fill the nonzero elements with random numbers uniformly distributed in $[-1, 1]$, and compute $X = X_L + X_L^T$. The value of the diagonal entries equals the sum of the absolute value of the corresponding off-diagonal entries plus 0.1. The resulting matrix is then diagonally dominant and positive definite. We further rescale X such that the variance is normalized to 1. Finally, we draw 1000 samples from $\mathcal{N}(\mathbf{0}, X^{-1})$ and compute the empirical covariance matrix from those samples, corresponding to C in eq. (1.8). For numerical stability, we add 10^{-5} to the diagonal entries of C . To generate the corresponding linear constraints, we specify the constraint according to eq. (1.16), so that ANS and ASPG are applicable. We follow a similar strategy as in [7]. We randomly select 50% of elements from the set $\{(i, j) | X_{i,j} = 0, i - j \geq 2\}$ and regard the resulting set as Ω in eq. (1.16).

We first investigate the influence of the number of Gibbs samples in the algorithm. Specifically, we fix the dimension $N = 500$ and the penalty parameter $\rho = 0.01$, and try sample sizes from 40 to 400 in steps of 40. We record how the primal objective value eq. (1.8), the violation of the constraints $\|\mathcal{A}(X) - \mathbf{b}\|/M$ (feasibility), and the mean square error $\|X - X^*\|/N^2$ (MSE) evolve as the algorithm proceeds. The results are depicted in Figure 4.1. The convergence results are also listed in Table 4.1. We observe that the algorithm will converge to the optimal even for a small sample size at the cost of more iterations. As the number of samples increases from 40 to 400, the total number of iterations till convergence decreases from nearly 4,200 to

around 1,500. In such case, the variance stemming from eq. (4.9) becomes smaller as K increases. As a consequence, the algorithm converges with fewer iterations. On the other hand, as indicated in the results, the total number of iterations is similar for $K \in \{360, 400\}$. This implies that after a certain threshold, increasing K cannot decrease the total number of iterations. In the case of large K , the total number of iterations is subject to the initialization and step size. Therefore, we prefer a moderate K in practice to achieve efficiency while avoiding redundant operations.

After investigating the impact of different values of K , we compare the performance of the proposed method with various penalty parameters ρ . Concretely, we fix $N = 500$, $K = 400$, and we increase the penalty parameter ρ from 10^{-4} to 10^4 by powers of 10. Similar to the case of Figure 4.1, we record how the primal objective value eq. (1.8), the feasibility, and the MSE change. The results are presented in Figure 4.2 and Table 4.2. We note that when $\rho \in \{10^{-4}, 10^{-3}\}$ (blue and orange line), the method fails to converge to the optimal point. As explained in Section 4.1, when the penalty parameter is too small, the augmented Lagrangian function degenerates to the Lagrangian function and cannot account for the linear constraints. In such case, the proposed method aims to minimize the objective function regardless of the constraints. Therefore, the output is not feasible. On the other hand, when $\rho \in \{10^3, 10^4\}$ (purple and light purple line), the problem becomes ill-conditioned, therefore, the method oscillates between solutions. In the case of $\rho = 10^4$, the oscillation is so significant that the method fails to converge. We terminate the algorithm in iteration 200 to illustrate the divergence. The corresponding results in the last column of Table 4.2 are not convergent. The results are convergent when $\rho \in \{10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$. There are minor differences within the results. For example, when $\rho = 10^3$, the feasibility is smaller and it takes more time to converge. When ρ is chosen larger, the weight of the term $\|\mathcal{A}(X) - \mathbf{b}\|^2$ increases, thus the violation of the linear constraints (feasibility) is smaller. Moreover, a larger ρ amplifies the variance of X and it takes more iterations to converge. This explains the increasing trend of run time from $\rho = 10^{-2}$ to $\rho = 10^3$. In conclusion, a penalty

Figure 4.1: The dimension is $N = 500$ and the number of constraints is $M = 55,917$. Under the same parameter setting, we test various number K of Gibbs samples. The objective value $f_0(X)$, the feasibility $\|\mathcal{A}(X) - \mathbf{b}\|/M$, and the mean square error $\|X - X^*\|/N^2$ converge. More details are listed in Table 4.1.

Table 4.1: Results for different sample sizes corresponding to Figure 4.1

K	40	80	120	160	200	240	280	320	360	400
Objective	461.43	460.91	460.55	460.45	460.35	460.38	460.24	460.27	460.23	460.17
Feasibility	1.35E-03	8.14E-04	7.09E-04	6.07E-04	5.12E-04	3.97E-04	4.15E-04	3.55E-04	3.92E-04	9.22E-04
MSE	1.24E-07	4.56E-07	2.56E-06	5.93E-06	8.14E-06	5.74E-06	1.68E-05	1.43E-05	4.45E-05	2.57E-05
Iteration	4200	3520	2730	2320	2150	2270	1810	1820	1340	1570
Time (secs)	160.10	129.97	137.52	126.48	122.04	105.80	137.53	120.74	112.72	136.18

parameter of a proper range (from 10^{-2} to 10^3) leads to the primal-dual optimal pair and a smaller penalty parameter within the suitable range is more efficient.

Next we validate the result of Theorem 4.3 by comparing Algorithm 3 with benchmark methods. In particular, we set $N = 500$, $\rho = 0.1$, and $K = 400$. As shown in Figure 4.3, we investigate the convergence performance in term of the primal objective value eq. (1.8), the feasibility, and the MSE, which are discussed in Theorem 4.3. The ASPG and ANS methods contain both inner iterations and outer iterations. During inner iterations, these methods optimize the objective function regardless of the matrix constraints. The sudden jumps in the red and yellow lines in Figure 4.3 correspond to the termination of a sequence of inner iterations in the ANS and ASPG methods respectively. The benchmark methods execute deterministic computations, and achieve similar convergence within fewer iterations. By contrast, PDSGD-ad spends more iterations to mitigate the influence of variance of the stochastic subgradients.

Next we proceed to explore the run time as a function of N (see Figure 4.4). We set the termination tolerance $\vartheta = 0.0005$ in Algorithm 3 in this experiment so that the convergence performance is comparable. We increase the dimension from 500 to 10,000 by powers of two with $K = 400$ and $\rho = 0.1$. We only run ANS up to $N = 2000$ and ASPG and LogdetPPA up to $N = 4000$ because they take over 10^6 seconds for higher dimensions. In order to explicitly show the increasing trend of run time as a function of the dimension, we fit a line in the log domain and list the corresponding slopes of each method. As indicated in Figure 4.4, the run time of the proposed Algorithm 3 grows approximately $\mathcal{O}(N^2)$, while the time complexity of benchmark methods is around $\mathcal{O}(N^3)$. As explained in the literature review, the time complexity of the benchmark methods is lower bounded by $\mathcal{O}(N^3)$ because of the cumbersome operations including matrix inversion, full spectral decomposition, and matrix determinant evaluation. In comparison, PDSGD-ad requires only $\mathcal{O}(N^2)$ operations. In Figure 4.4, we show the empirical scaling of the time complexity. It is clear from this graph that the numerical results are in line with the theoretical scaling

Figure 4.2: The dimension $N = 500$, the number of constraints $M = 55,917$, and sample size equals 400. The synthetic data set is the same as in Figure 4.1. Experiments are conducted for various penalty parameters $\rho \in [10^{-4}, 10^4]$. When ρ is chosen within a proper range, the estimate is convergent.

Table 4.2: Results for different penalty parameters corresponding to Figure 4.2

ρ	1.00E-04	1.00E-03	1.00E-02	1.00E-01	1.00E+00	1.00E+01	1.00E+02	1.00E+03	1.00E+04
Objective	430.4634588	436.64	460.17	460.10	460.07	460.03	460.01	460.65	545.30
Feasibility	3.28E-02	1.82E-02	9.22E-04	2.69E-04	2.38E-04	1.78E-04	9.02E-05	3.47E-05	4.11E-02
MSE	1.83E-03	2.88E-04	2.57E-05	8.71E-06	6.45E-06	5.48E-06	4.90E-06	8.06E-07	7.10E-03
Iteration	2000	2970	1570	1980	2120	2210	2350	3570	200
Time (secs)	151.6957483	218.09	136.18	153.42	153.07	156.94	151.89	238.75	133.34

Figure 4.3: Comparison of convergence performance in terms of the objective value, the feasibility, and the MSE.

behavior. Moreover, the feasibility and MSE of all methods are listed in Table 4.3. The ANS and ASPG methods force the elements in Ω to zeros. The procedure zeros the elements in Ω , therefore, these methods are only applicable to problems in the form of eq. (1.16). The feasibility of Logdet-PPA and PDSGD-ad is both on the scale of 10^{-6} , indicating that $\mathcal{A}(X) = \mathbf{b}$ is satisfied numerically. All MSE are in a comparable scale, therefore the results are in agreement with each other and all methods are convergent. In conclusion, PDSGD-ad achieves similar accuracy with only $\mathcal{O}(N^2)$ computational time, whereas it takes about $\mathcal{O}(N^3)$ time for benchmark methods.

Table 4.3: The feasibility and MSE $\|X - X^*\|/N^2$ of each method corresponding to Figure 4.4. X^* is the averaged estimate of benchmark methods.

	N	500	1000	2000	4000
ANS	Feasiblity	0.00	0.00	0.00	NA
	MSE	3.43E-06	4.95E-06	9.80E-07	NA
ASPG	Feasiblity	0.00	0.00	0.00	0.00
	MSE	2.43E-06	4.85E-06	9.70E-07	7.30E-06
Logdet-PPA	Feasiblity	2.60E-05	1.00E-05	1.01E-05	1.40E-06
	MSE	7.09E-06	2.90E-06	1.40E-06	7.30E-06
PDSGD-ad	Feasiblity	8.50E-06	2.60E-06	2.85E-06	3.60E-06
	MSE	1.10E-06	1.10E-06	1.87E-06	2.32E-06

4.4.2 Real Data

We consider four gene expression data sets: Lymph node status data [82], Arabidopsis thaliana data [71], Leukemia data [72], and hereditary breast cancer data [83]. There is no prior knowledge about the structure of the primal variable, therefore, we solve the problem eq. (1.8) without the linear constraints, i.e., $\Omega = \emptyset$. The dimension of each data set and the run time of each method is listed in Table 4.4. In conclusion, the proposed method achieves similar results at much lower computational cost when the dimension is large.

Figure 4.4: The dimension increases from 500 to 10,000 and $K = 400$. For varying dimensions, the feasibility $\|\mathcal{A}(X) - \mathbf{b}\|/M$ and the mean square error $\|X - X^*\|/N^2$ of the estimate of PDSGD-ad are presented in Table 4.3.

Table 4.4: Performance for four real gene-expression data with $K = 400$. We report the run time of each method and their relative error against X^* (the averaged estimate of benchmark methods).

Data Set	N	Method	ANS	ASPG	LogdetPPA	PDSGD-ad
Lymph	587	MSE	2.52E-06	2.65E-06	2.56E-06	2.70E-06
		Time (secs)	28.7	22.9	65.8	132.8
Arabidopsis	834	MSE	2.22E-06	2.32E-06	1.15E-06	1.26E-06
		Time (secs)	381.37	96.33	223.49	105.38
Leukemia	1255	MSE	1.43E-07	1.47E-07	1.49E-07	3.37E-07
		Time (secs)	1106	305.2	497.5	377.7
Hereditary	1869	MSE	2.77E-06	2.25E-06	2.75E-06	2.78E-06
		Time (secs)	1241.6	680.3	1822.8	428.1

4.5 Conclusion

In this chapter, we propose a novel stochastic primal-dual method tackling the general log-determinant optimization problem. We convert the original optimization problem into the problem of finding the saddle point of the augmented Lagrangian form, construct an efficient stochastic subgradient of the augmented Lagrangian function with the aid of Gibbs sampling, and adopt iterative restarted Lanczos method to ensure positive definiteness. In this manner, we achieve quadratic time complexity of solving the log-determinant optimization, whereas the time complexity of existing methods is at least cubic.

We theoretically analyze the convergence of the proposed method, establish the upper bound of the Lagrangian multipliers and stochastic subgradient, and conclude the convergence rate $\log(t)/t$. Moreover, we empirically investigate the impact of the choice of sample sizes and penalty parameters, and suggest proper choice of the parameters. Empirical comparison of the proposed method against benchmark methods also validates the theoretical convergence results. While the proposed method requires more iterations to mitigate the influence of stochasticity, the overall computational time nevertheless scales quadratically with the dimension instead of cubically.

Chapter 5

Conclusions and Future Work

This thesis proposes the GIBSON method for structure learning problem of Gaussian graphical models, and generalizes the proposed method to tackle the general log-determinant optimization. We conclude this thesis by summarizing the main contributions and suggesting ideas for future work.

5.1 Summary of Contributions

In Chapter 3, we consider the problem of structure learning (a.k.a. inverse covariance selection) of Gaussian graphical models. The objective is to learn the graph structure from multivariate Gaussian samples. In contrast to existing deterministic methods, we apply stochastic optimization to minimize the unconstrained non-smooth convex objective function. Specifically, we construct an efficient stochastic subgradient of the objective function on the basis of Gibbs sampling. By constructing an asymptotically converging Markov chain, Gibbs sampling provides a noisy but efficient realization of the stochastic subgradient. The resulting time complexity of updating the precision matrix is thereby reduced to $\mathcal{O}(N^2)$, whereas the existing methods are subject to cumbersome operations with $\mathcal{O}(N^3)$ time complexity such as

matrix inversion and full spectral decomposition. Moreover, the optimization problem requires a positive definite estimate of the inverse covariance. Existing methods decompose the eigenvalues, compute the Cholesky decomposition, or evaluate Schur complement to check the definiteness. We employ the implicitly restarted Lanczos process to check the extreme eigenvalue of the estimate. This technique reduces the time complexity of definiteness checking from $\mathcal{O}(N^3)$ to $\mathcal{O}(N^2)$.

We also provide a detailed convergence analysis of the proposed method. We prove an analytical bound on the eigenvalues of the optimal solution and obtain the bounded closed convex set on which the optimization problem can be constrained. The objective function can be regarded as strongly convex when restricted to the bounded set. Moreover, we analyze the second-order moments of the singular Wishart distribution, which characterize the distribution of each Gibbs sample. By combining the bound on Wishart distribution and the closed convex set, we can bound the stochastic subgradient in terms of its second-order moments. The strong convexity and bound on the stochastic subgradient lead to a sublinear convergence rate of the proposed method.

Numerical analysis supports the theoretical analysis. When the precision matrix is chosen as semi-sparse and the penalty parameter is chosen small, the time complexity of benchmark methods is at least cubic. By contrast, the time complexity of the proposed method is only $\mathcal{O}(N^2)$. Moreover, we found that the efficiency of the proposed method is less sensitive to the graph density and choices of penalty parameters compared with existing methods. In conclusion, the proposed method is attractive when the underlying graph is relatively dense or a small penalty parameter is preferred, especially for high-dimensional problems.

Following the optimization problem of structure learning problem, we extend the proposed method to the general log-determinant optimization in Chapter 4. Similar to the case of GGM, the time complexity of existing methods is lower bounded by $\mathcal{O}(N^3)$, which severely impedes their scalability to high-dimensional problems.

The general matrix constraints constitute an interesting challenge. In order to incorporate the matrix constraints into the proposed method, we consider the augmented Lagrangian function of the optimization problem. By properly selecting the weight parameter, we convert the constrained optimization problem into searching for the saddle point of the augmented Lagrangian function. We modify the proposed method in the context of Uzawa iterations and construct a primal-dual stochastic subgradient method, which is also the first attempt to apply stochastic optimization to the general log-determinant optimization. The proposed method features $\mathcal{O}(N^2)$ time complexity.

Next we analyze the generalized proposed method theoretically. On the basis of reasonable assumptions, we analyze the bound of its set of Lagrangian multipliers and obtain a bound on the dual optimum, which leads to a bound on the primal optimum. We thereafter construct an equivalent strongly convex optimization problem. The equivalent formulation suggests a bound on the second-order moments of the subgradient w.r.t. the primal term and the augmented term respectively. On the basis of those bounds, we evaluate bounds on the second-order moment of the stochastic subgradient, which is essential to the convergence analysis of the proposed method. By constructing the bound, analyzing the properties of estimate from subsequent iterations, and incorporating the assumptions, we finally establish the theoretical convergence guarantee, which states the sublinear convergence rate of the primal objective value, feasibility, and primal estimate.

We further numerically analyze the performance of the proposed method on synthetic data and real data. We exploit the influence of the two important parameters, the sample size controlling the variance of the stochastic subgradient, and the penalty parameter balancing between the weight of primal objective and the constraints. The numerical results for different sample sizes indicate that the algorithm converges to the primal-dual optimal pair even when the sample size is small at the cost of more iterations. It also implies that increasing a large sample size cannot decrease the total number of iterations, which is also subject to the initialization and

the choice of step sizes. Moreover, a penalty parameter of a proper range generally leads to desired optimal solution. If the penalty parameter is too small, the weight of the constraints becomes minimal and the resulting estimate is infeasible. On the other hand, if the penalty is too large, the resulting problem is unstable and the method oscillates. The numerical results also support the theoretical convergence results of the proposed method, and validate $\mathcal{O}(N^2)$ time complexity of the proposed method. The proposed method reaches similar solutions as that of the benchmark methods in much less time in high-dimensional cases, therefore, it is attractive for problems at large scale. In this manner, we improve the scalability for the general log-determinant problems.

5.2 Recommendations for Future Work

5.2.1 Log-determinant Optimization with Nonlinear Constraints

In the log-determinant optimization problem eq. (1.8) [35,36], we typically considers linear matrix constraints only. In a more general manner, we can consider nonlinear inequality constraints, and the resulting problem is formulated as:

$$\begin{aligned} \min \quad & f_0(X) \triangleq \text{tr}(CX) - \mu_1 \log \det(X) + \mu_2 \|X\|_1 \\ \text{s.t.} \quad & f_i(X) \leq 0, i = 1, \dots, M, X \in \mathbb{S}_+^N \end{aligned} \tag{5.1}$$

where $f_i(X)$ is a differentiable convex function for each $i \in \{1, \dots, M\}$, the gradient of $f_i(X)$ can be efficiently evaluated, and $\mathcal{D} = \cap_{i=0}^M \text{dom} f_i$ is non-empty. We hereby sketch a more general version of our primal-dual stochastic subgradient method for eq. (5.1) by beginning with its augmented Lagrangian function.

In order to obtain the augmented Lagrangian function of eq. (5.1), we introduce auxiliary variables $\mathbf{z} \in \mathbb{R}_+^M$ such that the i -th constraints becomes $f_i(X) + z_i = 0$.

As a consequence, the resulting problem is formulated as:

$$\begin{aligned} \min \quad & f_0(X) \triangleq \text{tr}(CX) - \mu_1 \log \det(X) + \mu_2 \|X\|_1 \\ \text{s.t.} \quad & f_i(X) + z_i = 0, i = 1, \dots, M, \mathbf{z} \in \mathbb{R}_+^M, \end{aligned} \quad (5.2)$$

Following the augmented Lagrangian formulation of optimization problems with equality constraints, the augmented Lagrangian function of eq. (5.2) can be written as:

$$\mathcal{L}(X, \mathbf{y}, \mathbf{z}) = f_0(X) + \sum_{i=1}^M y_i (f_i(X) + z_i) + \frac{\rho}{2} \sum_{i=1}^M (f_i(X) + z_i)^2, \quad (5.3)$$

where y_i is the dual variable w.r.t. i -th constraint, $\rho > 0$ is the introduced penalty parameter. We can further minimize $\mathcal{L}(X, \mathbf{y}, \mathbf{z})$ w.r.t. $\mathbf{z}$ by deriving the corresponding gradient:

$$\mathcal{L}_{z_i}(X, \mathbf{y}, \mathbf{z}) = y_i + \rho(f_i(X) + z_i). \quad (5.4)$$

Therefore, the minimal of $\mathcal{L}(X, \mathbf{y}, \mathbf{z})$ is attained at:

$$z_i = \begin{cases} 0, & \text{if } y_i + \rho f_i(X) \geq 0 \\ -\frac{1}{\rho}(y_i + \rho f_i(X)), & \text{if } y_i + \rho f_i(X) \leq 0. \end{cases} \quad (5.5)$$

Substituting the value of $\mathbf{z}$ back into $\mathcal{L}(X, \mathbf{y}, \mathbf{z})$ leads to the residual problem:

$$\mathcal{L}(X, \mathbf{y}) = f_0(X) + \frac{1}{2\rho} \sum_{i=1}^M ([y_i + \rho f_i(X)]_+^2 - y_i^2), \quad (5.6)$$

where $[x]_+ \triangleq \max(0, x)$ and it takes precedence over the square operation. We can verify that:

$$\frac{1}{2\rho} ([y_i + \rho f_i(X)]_+^2 - y_i^2) = \begin{cases} \frac{\rho}{2} f_i(X)^2 + y_i f_i(X), & \text{if } y_i + \rho f_i(X) \geq 0 \\ -\frac{y_i^2}{2\rho}, & \text{if } y_i + \rho f_i(X) \leq 0. \end{cases} \quad (5.7)$$

On inspection we know that $\lim_{\rho \rightarrow 0} \mathcal{L}(X, \mathbf{y}) = L(X, \mathbf{y}) \triangleq f_0(X) + \sum_{i=1}^M y_i f_i(X)$, which is the Lagrangian function of eq. (5.1). Equation (5.6) is known as the aug-

mented Lagrangian function of eq. (5.1), from which we can derive a primal-dual method.

Next we can construct the stochastic subgradient w.r.t. X and $\mathbf{y}$. We can write the subgradient of eq. (5.6) as follow:

$$\mathcal{L}_X(X, \mathbf{y}) = \underbrace{C - \mu_1 X^{-1} + \Lambda}_{\in \partial f_0(X)} + \frac{1}{\rho} \sum_{i=1}^M [y_i + \rho f_i(X)]_+ \nabla f_i(X), \quad (5.8a)$$

$$\mathcal{L}_\mathbf{y}(X, \mathbf{y}) = [\mathcal{L}_{y_1}, \dots, \mathcal{L}_{y_M}]^T, \quad (5.8b)$$

where we define $\mathcal{L}_{y_i}$ according to:

$$\mathcal{L}_{y_i}(X, \mathbf{y}) = \frac{[y_i + \rho f_i(X)]_+ - y_i}{\rho} = \max \left(f_i(X), -\frac{y_i}{\rho} \right). \quad (5.9)$$

We can reduce the computational load of computing eq. (5.8) by three ideas as follows.

- Following the stochastic approximation introduced in Section 3.1, we can construct a stochastic subgradient $g_0(X)$ such that $\mathbb{E}[g_0(X)] \in \partial f_0(X)$. In this way, we can bypass the $\mathcal{O}(N^3)$ complexity stemming from X^{-1} .
- In each iteration, we can uniformly sample a batch of indexes $\mathcal{J} \triangleq \{j_1, \dots, j_m\} \subset \{1, \dots, M\}$, where m is a predetermined constant such that $m \ll M$. We compute the stochastic gradient $\sum_{i=1}^m [y_{j_i} + \rho f_{j_i}(X)]_+ \nabla f_{j_i}(X)$, whose expectation satisfies:

$$\mathbb{E} \left[\sum_{i=1}^m [y_{j_i} + \rho f_{j_i}(X)]_+ \nabla f_{j_i}(X) \right] = \sum_{i=1}^M [y_i + \rho f_i(X)]_+ \nabla f_i(X). \quad (5.10)$$

In this way, we reduce the time complexity due to the dimensionality M .

- Following the index set $\mathcal{J}$, we only update $\{y_i | i \in \mathcal{J}\}$ following the gradient eq. (5.9) in one iteration.

5.2.2 Alternative Sampling Method

In Section 4.2, we theoretically analyze the convergence performance of Algorithm 3 on the basis of Assumptions 4.1 to 4.4. We notice that Assumption 4.3 is only asymptotically true when t increases due to the burn-in bias. In practice, the burn-in bias asymptotically decreases to 0. If we can remove the burn-in bias at no significant computational cost, we can further accelerate the convergence. In the future, we could replace Gibbs sampling by alternative sampling methods which generate i.i.d. unbiased estimators. We introduce a coupling method below which can remove the burn-in bias.

In [84, 85], the author couples two Markov chains $\mathbf{x} = (\mathbf{x}_i)_{i \geq 0}$ and $\mathbf{y} = (\mathbf{y}_i)_{i \geq 0}$ with the same initial distribution and the same marginal law, whose resulting invariant distribution is $\mathcal{N}(\mathbf{0}, X_t^{-1})$. With the aid of a meticulously chosen Markov transition kernel on the coupled Markov chain, the two chains are guaranteed to meet after a finite number τ of iterations. τ is defined as meeting time of the coupled Markov chain. An unbiased random variable is constructed based on the two Markov chains after the meeting time. Under appropriate assumptions, it can be shown that:

$$\mathbb{E} \left[\frac{1}{m - k + 1} \sum_{i=k}^m \mathbf{x}_i \mathbf{x}_i^T + \sum_{i=k+1}^{\tau-1} \min \left(1, \frac{i - k}{m - k + 1} \right) (\mathbf{x}_i \mathbf{x}_i^T - \mathbf{y}_{i-1} \mathbf{y}_{i-1}^T) \right] = X^{-1}, \quad (5.11)$$

where the burn-in bias has been removed, k and m are tuning parameters satisfying $0 \leq k \leq m$. Moreover, it can also be shown that the estimator on the left-handed of eq. (5.11) has finite variance.

If we can construct a coupled Markov chain of Gibbs sampling, we can obtain unbiased estimators and maintain the time complexity as $\mathcal{O}(N^2)$. However, there are a few challenges. For example, the coupled Markov chain can remove the burn-in bias of approximating X_t^{-1} . In order to remove the burn-in bias of all iterations, we should figure out how to iteratively update the coupled Markov chain targeting

the stationary distribution $\mathcal{N}(\mathbf{0}, X_t^{-1})$ so that the updated Markov chain can yield unbiased estimators of $\{X_t - \eta_t g(X_t)\}^{-1}$. The updated coupled Markov chain more likely requires some iterations to meet again, i.e., τ in the updated Markov chain is non-zero. In this case, we need to establish a bound on τ or a bound on the bias if we discard those iterations.

5.2.3 Problem Partition

If the dimension is large and the sparsity parameter μ in eq. (1.1) is large, we can divide the problem into low-dimensional subproblems following the routine in [86]. In the following, we briefly summarize the covariance thresholding and its equivalence to the vertex-edge partition according to the precision matrix.

Specifically, given μ , we threshold the covariance matrix S of $\mathbf{x}_{1:n}$, and obtain the graph adjacency matrix G as:

$$G_{i,j} = \begin{cases} 1, & \text{if } |S_{i,j}| > \mu, i \neq j \\ 0, & \text{otherwise,} \end{cases} \quad (5.12)$$

where G defines a undirected graph $\mathcal{G}$ on nodes $\mathcal{V} = \{1, \dots, N\}$. If μ is large, it is likely that the induced covariance matrix G consists of several disconnect components. In other words, we have the decomposition $\mathcal{V} = \cup_{k=1}^K \mathcal{V}_k$, where $0 \leq k_1, k_2 \leq K$ and $k_1 \neq k_2$, K is the total number of disconnected components, $\mathcal{V}_{k_1} \cap \mathcal{V}_{k_2} = \emptyset$, and there is no path between arbitrary $i \in \mathcal{V}_{k_1}$ and $j \in \mathcal{V}_{k_2}$.

It has been shown that the partition according to the covariance thresholding is equivalent to the graph partition according to X by solving eq. (1.1). As a consequence, if μ is large, we can threshold the covariance matrix, obtain graph partition according to the covariance thresholding, divide the problem by permutation and grouping in line with the graph partition, and solve the resulting lower-dimensional subproblems. This can substantially reduce the computational load and further

increase the scalability.

Furthermore, it is possible that we can analyze eq. (1.8) following the manner in [86] and obtain similar substructure equivalence between C and X . On the basis of this equivalence, the high-dimensional problem can be decomposed into more tractable low-dimensional subproblems when μ_2 is large.

Appendix A

An Alternative Convergence Analysis of Algorithm 3

We have proven the convergence rate of Algorithm 3, where we assume the upper bound on the dual variable. In this Appendix, we remove Assumption 4.4 in Section 4.2, and derive a bound in terms of the expectation of the second-order moment of the dual variables $\mathbf{y}$. We apply a similar convergence analysis as in Theorem 4.3. The resulting method is not as efficient as Algorithm 3 but its theoretical property of the method in this appendix is attractive. Therefore, we list this step size setting in the Appendix.

Lemma A.1. *Assume that $X \in \mathcal{X}^+$ and $\mathbf{y}^0 = \mathbf{0}$. If:*

$$\frac{\eta_2^t}{\eta_1^t} \geq \eta_2^{t+1} \left(\frac{1}{\eta_1^{t+1}} - \xi \right), \rho \geq \eta_2^t, \quad (\text{A.1})$$

we have:

$$\mathbb{E}[\|\mathbf{y}^{L+1}\|^2] \leq b_0 + \sum_{t=0}^L 4\eta_2^t \eta_1^t \mathbb{E}[\|\mathbf{y}^t\|_1^2], \quad (\text{A.2})$$

where

$$b_0 \triangleq 2\eta_2^0 \left(\frac{1}{\eta_1^0} - \xi \right) \|X^0 - X^*\|^2 + 4\mathbb{E}[\|\mathbf{y}^*\|^2] + \sum_{t=0}^L 4\eta_1^t \eta_2^t (\sigma^2 + 4\rho^2 (M\sqrt{N}\beta + \|\mathbf{b}\|_1)^2). \quad (\text{A.3})$$

Proof. Let $(X, \mathbf{y})$ in eq. (4.37) be $(X^*, \mathbf{y}^*)$. Then eq. (4.37) becomes:

$$\begin{aligned} & \mathbb{E}[f_0(X^t) - \langle \mathbf{y}^*, \mathcal{A}(X^t) - \mathbf{b} \rangle] + \frac{1}{2\eta_1^t} \mathbb{E}\|X^{t+1} - X^*\|^2 + \frac{1}{2\eta_2^t} \mathbb{E}\|\mathbf{y}^{t+1} - \mathbf{y}^*\|^2 \\ & \leq \mathbb{E}[f_0(X^*) + f_1(X^*, \mathbf{y}^t)] + \frac{1}{2} \left(\frac{1}{\eta_1^t} - \xi \right) \mathbb{E}\|X^t - X^*\|^2 + \frac{1}{2\eta_2^t} \mathbb{E}\|\mathbf{y}^*\|^2 + \eta_1^t (\sigma^2 + \varrho_{\mathbf{y}}^2), \end{aligned} \quad (\text{A.4})$$

where we omitted the last two terms on the right-hand side considering the facts $\mathbb{E}[\langle X^t - X^*, g_0^t - \mathbb{E}[g_0^t] \rangle] = 0$ and $\rho/\eta_2^t \geq 1$. After permutation, eq. (A.4) can be written as:

$$\begin{aligned} & \mathbb{E}[f_0(X^t) - f_0(X^*) - \langle \mathbf{y}^*, \mathcal{A}(X^t) - \mathbf{b} \rangle] + \frac{1}{2\eta_1^t} \mathbb{E}\|X^{t+1} - X^*\|^2 + \frac{1}{2\eta_2^t} \mathbb{E}\|\mathbf{y}^{t+1} - \mathbf{y}^*\|^2 \\ & \leq \frac{1}{2} \left(\frac{1}{\eta_1^t} - \xi \right) \mathbb{E}\|X^t - X^*\|^2 + \frac{1}{2\eta_2^t} \mathbb{E}\|\mathbf{y}^t - \mathbf{y}^*\|^2 + \eta_1^t (\sigma^2 + \varrho_{\mathbf{y}}^2). \end{aligned} \quad (\text{A.5})$$

According to the definition eq. (4.51), we have the following equation by multiplying $2\eta_2^t$ on both sides:

$$\begin{aligned} & 2\eta_2^t \mathbb{E}[h(X^t; X^*, \mathbf{y}^*)] + \frac{\eta_2^t}{\eta_1^t} \mathbb{E}\|X^{t+1} - X^*\|^2 + \mathbb{E}\|\mathbf{y}^{t+1} - \mathbf{y}^*\|^2 \\ & \leq \eta_2^t \left(\frac{1}{\eta_1^t} - \xi \right) \mathbb{E}\|X^t - X^*\|^2 + \mathbb{E}\|\mathbf{y}^t - \mathbf{y}^*\|^2 + 2\eta_2^t \eta_1^t (\sigma^2 + \varrho^2) \\ & \leq \frac{\eta_2^{t-1}}{\eta_1^{t-1}} \mathbb{E}\|X^t - X^*\|^2 + \mathbb{E}\|\mathbf{y}^t - \mathbf{y}^*\|^2 + 2\eta_2^t \eta_1^t (\sigma^2 + \varrho_{\mathbf{y}}^2), \end{aligned} \quad (\text{A.6})$$

where the last line holds only when $t \geq 1$. Summing up eq. (A.6) from $t = 0$ to L ,

and noticing that $\mathbf{y}^0 = \mathbf{0}$ and $h(X^k; X^*, \mathbf{y}^*) \geq 0$, we have the following result:

$$\begin{aligned} & \mathbb{E}[\|\mathbf{y}^{L+1} - \mathbf{y}^*\|^2] \\ & \leq 2\eta_2^0\left(\frac{1}{\eta_1^0} - \xi\right)\|X^0 - X^*\|^2 + \|\mathbf{y}^*\|^2 + \sum_{t=0}^L 2\eta_2^t\eta_1^t(\sigma^2 + 2\mathbb{E}[\|\mathbf{y}^t\|_1^2] + 4\rho^2(M\sqrt{N}\beta + \|\mathbf{b}\|_1)^2). \end{aligned} \quad (\text{A.7})$$

Moreover,

$$\begin{aligned} \mathbb{E}[\|\mathbf{y}^{L+1}\|^2] &= \mathbb{E}[\|\mathbf{y}^{L+1} - \mathbf{y}^* + \mathbf{y}^*\|^2] \\ &\leq 2\mathbb{E}[\|\mathbf{y}^{L+1} - \mathbf{y}^*\|^2] + 2\mathbb{E}[\|\mathbf{y}^*\|^2], \end{aligned} \quad (\text{A.8})$$

from which follows eq. (A.2). $\square$

Remark. Lemma A.1 requires a condition eq. (A.1) of the step size $\{\eta_1^t, \eta_2^t | t \in \mathcal{R}^*\}$.

If we set:

- constant step size: $\eta_1 = \varepsilon/\sqrt{L}, \eta_2 = \rho/\sqrt{L}$,
- linearly decaying step size: $\eta_1^t = \varepsilon/(t+1), \eta_2^t = \rho/\log(L+1)$, where $\varepsilon \geq 1/\xi$,

where L is a predetermined maximum number of iterations, for example, 10^5 , we can verify that eq. (A.1) is satisfied. We proceed to analyze the convergence performance of the linearly decaying step size.

Lemma A.2. Assume

$$\eta_1^t = \frac{\varepsilon}{t+1}, \eta_2^t = \frac{\rho}{\log(L+1)}, \text{ with } t \in \mathcal{R}^*, 0 \leq t \leq L. \quad (\text{A.9})$$

We have the bound on $\mathbb{E}[\|\mathbf{y}^t\|^2] \leq v^2$, where

$$v \triangleq \sqrt{\frac{d_1}{1 - 8\varepsilon\rho}}, \quad (\text{A.10})$$

$$d_1 \triangleq \frac{2\rho(\varepsilon - \xi)}{\log(L+1)}\|X^0 - X^*\|^2 + 4\mathbb{E}[\|\mathbf{y}^*\|^2] + 4\varepsilon\rho(\sigma^2 + 4\rho^2(M\sqrt{N}\beta + \|\mathbf{b}\|_1)^2). \quad (\text{A.11})$$

Proof. By the definition of η_1^t and η_2^t in eq. (A.9), we know eq. (A.1) holds, therefore, we have eq. (A.2):

$$\begin{aligned} \mathbb{E}[\|\mathbf{y}^{t+1}\|^2] &\leq \frac{2\rho(\varepsilon - \xi)}{\log(L+1)} \|X^0 - X^*\|^2 + 4\mathbb{E}[\|\mathbf{y}^*\|^2] \\ &\quad + \sum_{i=0}^t \frac{4\varepsilon\rho}{(i+1)\log(L+1)} (\sigma^2 + 2\mathbb{E}[\|\mathbf{y}^i\|_1^2] + 4\rho^2(M\sqrt{N}\beta + \|\mathbf{b}\|_1)^2). \end{aligned} \quad (\text{A.12})$$

We have the following inequality:

$$\begin{aligned} \sum_{i=0}^t \frac{1}{(i+1)\log(L+1)} &= \frac{1}{\log(L+1)} \sum_{i=0}^t \frac{1}{i+1} \\ &\leq \frac{1}{\log(L+1)} \int_1^{t+1} \frac{1}{x} dx \leq \frac{1}{\log(L+1)} \log(t+1) \leq 1. \end{aligned} \quad (\text{A.13})$$

In view of eq. (A.13), we can construct an upper bound for eq. (A.12):

$$\mathbb{E}[\|\mathbf{y}^{t+1}\|^2] \leq d_1 + \sum_{i=0}^t \frac{8\varepsilon\rho}{(i+1)\log(L+1)} \mathbb{E}[\|\mathbf{y}^i\|_1^2]. \quad (\text{A.14})$$

We know that $\mathbb{E}[\|\mathbf{y}^t\|^2] \leq v^2$ holds for $t = 0$ because $\mathbf{y}^0 = \mathbf{0}$. Assuming it holds for $i \leq t \leq L$, we have the following equation by combining eq. (A.12) and eq. (A.13):

$$\mathbb{E}[\|\mathbf{y}^{t+1}\|] \leq d_1 + 8\varepsilon\rho \frac{d_1}{1 - 8\varepsilon\rho} = \frac{d_1}{1 - 8\varepsilon\rho}. \quad (\text{A.15})$$

□

We now formulate an alternative theorem to Theorem 4.3 without the assumption of bound on dual variable.

Theorem A.1. *Given Assumptions 4.1 to 4.3, if the step size is specified as in eq. (A.9) and $(X^t, \mathbf{y}^t)$ is the pair from eq. (4.11), we have:*

$$\mathbb{E}\|X^{t+1} - X^*\|^2 \leq \frac{2\varepsilon}{t+1} (\Phi(X^*) + \frac{\log(t+1)}{2\rho} \mathbb{E}\|\mathbf{y}^*\|^2), \quad (\text{A.16})$$

where

$$\Phi(X) \triangleq \frac{1}{2} \left(\frac{1}{\eta_1^0} - \xi \right) \mathbb{E} \|X^0 - X^*\|^2 + \varepsilon \log(\bar{t} + 1) (\sigma^2 + 4\rho^2 (M\sqrt{N}\beta + \|\mathbf{b}\|_1)^2 + v^2). \quad (\text{A.17})$$

Moreover, we can evaluate the objective value and the feasibility of the averaged estimate $\bar{X}^t = 1/(t+1) \sum_{i=0}^t X^i$:

$$\mathbb{E}[\|\mathcal{A}(\bar{X}^t) - \mathbf{b}\|_1] \leq \frac{1}{t+1} (\mathbb{E}[\Phi(X^*)] + \frac{M \log(L+1)}{\rho} + \frac{\log(L+1)}{\rho} \|\mathbf{y}^*\|^2), \quad (\text{A.18})$$

$$\mathbb{E}[|f_0(\bar{X}^t) - f_0(X^*)|] \leq \frac{1}{t+1} \max(\mathbb{E}[\Phi(X^*)], \frac{1}{2} (\mathbb{E}[\Phi(X^*)] + \frac{9 \log(L+1)}{2\rho} \mathbb{E} \|\mathbf{y}^*\|^2)). \quad (\text{A.19})$$

Proof. Given the conditions in Lemma A.2, we set $X = X^*$ in eq. (4.37), and sum it up from $t = 0$ to $\bar{t} \leq L$:

$$\begin{aligned} & \sum_{t=0}^{\bar{t}} \mathbb{E}[f_0(X^t) - \langle \mathbf{y}, \mathcal{A}(X^t) - \mathbf{b} \rangle] + \frac{1}{2\eta_1^t} \mathbb{E} \|X^{\bar{t}+1} - X\|^2 + \frac{1}{2\eta_2^t} \mathbb{E} \|\mathbf{y}^{\bar{t}+1} - \mathbf{y}\|^2 \\ & \leq \sum_{t=0}^{\bar{t}} \mathbb{E}[f_0(X^*) + f_1(X^*, \mathbf{y}^t)] + \frac{1}{2} \left(\frac{1}{\eta_1^0} - \xi \right) \mathbb{E} \|X^0 - X^*\|^2 + \frac{1}{2\eta_2^0} \mathbb{E} \|\mathbf{y}^0 - \mathbf{y}\|^2 \\ & \quad + \sum_{t=0}^{\bar{t}} \eta_1^t (\sigma^2 + 2\mathbb{E}[\|\mathbf{y}^t\|_1^2] + 4\rho^2 (M\sqrt{N}\beta + \|\mathbf{b}\|_1)^2) - \sum_{t=0}^{\bar{t}} \frac{1}{2\eta_2^t} \left(\frac{\rho}{\eta_2^t} - 1 \right) \|\mathbf{y}^{t+1} - \mathbf{y}^t\|^2 \\ & \quad - \sum_{t=0}^{\bar{t}} \mathbb{E}[\langle X^t - X^*, g_0^t - \mathbb{E}[g_0^t] \rangle]. \end{aligned} \quad (\text{A.20})$$

Since $\mathbf{y}^0 = \mathbf{0}$, $\eta_2^t \leq \rho$, $\mathbb{E}[\langle X^t - X^*, g_0^t - \mathbb{E}[g_0^t] \rangle] = 0$, we have the following upper

bound for eq. (A.20):

$$\begin{aligned} & \sum_{t=0}^{\bar{t}} \mathbb{E}[f_0(X^t) - f_0(X^*) - \langle \mathbf{y}, \mathcal{A}(X^t) - \mathbf{b} \rangle] + \frac{1}{2\eta_1^{\bar{t}}} \mathbb{E}\|X^{\bar{t}+1} - X\|^2 + \frac{1}{2\eta_2^{\bar{t}}} \mathbb{E}\|\mathbf{y}^{\bar{t}+1} - \mathbf{y}\|^2 \\ & \leq \Phi(X^*) + \frac{1}{2\eta_2^0} \mathbb{E}\|\mathbf{y}\|^2. \end{aligned} \quad (\text{A.21})$$

We also notice that:

$$\frac{1}{2\eta_1^{\bar{t}}} \mathbb{E}\|X^{\bar{t}+1} - X\|^2 \geq 0, \quad (\text{A.22})$$

$$\frac{1}{2\eta_2^{\bar{t}}} \mathbb{E}\|\mathbf{y}^{\bar{t}+1} - \mathbf{y}^*\|^2 \geq 0, \quad (\text{A.23})$$

$$\sum_{t=0}^{\bar{t}} \mathbb{E}[f_0(X^t) - f(X^*) - \langle \mathbf{y}^*, \mathcal{A}(X^t) - \mathbf{b} \rangle] \geq 0. \quad (\text{A.24})$$

If we choose $\mathbf{y} = \mathbf{y}^*$ and drop the first and third term on the left-hand side in eq. (A.21), it follows:

$$\mathbb{E}\|X^{\bar{t}+1} - X\|^2 \leq \frac{2\varepsilon}{\bar{t}+1} (\Phi(X^*) + \frac{\log(L+1)}{2\rho} \mathbb{E}\|\mathbf{y}^*\|^2), \quad (\text{A.25})$$

from which follows eq. (A.16). Furthermore, we drop the second and third term on the left-hand side of eq. (A.21) (without assuming $\mathbf{y} = \mathbf{y}^*$, different from eq. (A.25)), leading to:

$$\mathbb{E}\left[\sum_{t=0}^{\bar{t}} f_0(X^t) - (\bar{t}+1)f_0(X^*) - \sum_{t=0}^{\bar{t}} \langle \mathbf{y}, \mathcal{A}(X^t) - \mathbf{b} \rangle\right] \leq \mathbb{E}[\Phi(X^*)] + \frac{\log(L+1)}{2\rho} \mathbb{E}\|\mathbf{y}\|^2. \quad (\text{A.26})$$

Since $f_0(X)$ is convex and $\mathcal{A}(X)$ is linear, eq. (A.26) implies:

$$\mathbb{E}[f_0(\bar{X}^{\bar{t}}) - f_0(X^*) - \langle \mathbf{y}, \mathcal{A}(\bar{X}^{\bar{t}}) - \mathbf{b} \rangle] \leq \frac{1}{\bar{t}+1} \mathbb{E}[\Phi(X^*)] + \frac{\log(L+1)}{2(\bar{t}+1)\rho} \mathbb{E}\|\mathbf{y}\|^2, \quad (\text{A.27})$$

where $\bar{X}^{\bar{t}} \triangleq 1/(\bar{t} + 1) \sum_{t=0}^{\bar{t}} X^t$. According to the definition eq. (4.51):

$$\mathbb{E}[h(\bar{X}^{\bar{t}}; X^*, \mathbf{y})] \leq \frac{1}{\bar{t} + 1} \mathbb{E}[\Phi(X^*)] + \frac{\log(L + 1)}{2(\bar{t} + 1)\rho} \mathbb{E}\|\mathbf{y}^*\|^2. \quad (\text{A.28})$$

According to Lemma 4.10, we have the following two conclusions with respect to the feasibility:

$$\mathbb{E}[\|\mathcal{A}(\bar{X}^{\bar{t}}) - \mathbf{b}\|_1] \leq \frac{1}{\bar{t} + 1} (\mathbb{E}[\Phi(X^*)] + \frac{M \log(L + 1)}{\rho} + \frac{\log(L + 1)}{\rho} \mathbb{E}\|\mathbf{y}^*\|^2), \quad (\text{A.29})$$

and with respect to the objective value:

$$\mathbb{E}[|f_0(\bar{X}^{\bar{t}}) - f_0(X^*)|] \leq \frac{1}{\bar{t} + 1} \max(\mathbb{E}[\Phi(X^*)], \frac{1}{2}(\mathbb{E}[\Phi(X^*)] + \frac{9 \log(L + 1)}{2\rho} \mathbb{E}\|\mathbf{y}^*\|^2)). \quad (\text{A.30})$$

This concludes the proof for Theorem A.1. $\square$

Publications

Journal Papers:

Songwei Wu, Hang Yu, Justin Dauwels, GIBSON: Gibbs Sampling-based Stochastic Subgradient Descent Algorithms for Inverse Covariance Estimation, *IEEE Transactions on Signal Processing* (under peer review)

Songwei Wu, Hang Yu, Justin Dauwels, "Primal-Dual Stochastic Subgradient Method for Log-determinant Optimization", *IEEE Transactions on Signal Processing* (drafted)

Hang Yu, Songwei Wu, Justin Dauwels, BISN: Low-complexity Bayesian Inference of Sparse Networks, *Journal of Machine Learning Research* (submitted)

Conference Papers:

Songwei Wu, Hang Yu, and Justin Dauwels. "Efficient Stochastic Subgradient Descent Algorithms for High-dimensional Semi-sparse Graphical Model Selection." ICASSP 2019-2019 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.

Songwei Wu, Hang Yu, and Justin Dauwels. "Primal-Dual Stochastic Subgradient Method for Log-determinant Optimization" ICASSP 2020-2020 *IEEE International*

Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020. (accepted)

Bibliography

- [1] O. Banerjee, L. El Ghaoui, and A. d’Aspremont, “Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data,” *The Journal of Machine Learning Research*, vol. 9, pp. 485–516, 2008.
- [2] J. Friedman, T. Hastie, and R. Tibshirani, “Sparse inverse covariance estimation with the graphical lasso,” *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.
- [3] S. Mandt and D. Blei, “Smoothed gradients for stochastic variational inference,” in *Advances in Neural Information Processing Systems*, 2014, pp. 2438–2446.
- [4] K. Scheinberg and I. Rish, “SINCO—a greedy coordinate ascent method for sparse inverse covariance selection problem,” *preprint*, 2009.
- [5] A. d’Aspremont, O. Banerjee, and L. El Ghaoui, “First-order methods for sparse covariance selection,” *SIAM Journal on Matrix Analysis and Applications*, vol. 30, no. 1, pp. 56–66, 2008.
- [6] Z. Lu, “Smooth optimization approach for sparse covariance selection,” *SIAM Journal on Optimization*, vol. 19, no. 4, pp. 1807–1827, 2009.
- [7] —, “Adaptive first-order methods for general sparse inverse covariance selection,” *SIAM Journal on Matrix Analysis and Applications*, vol. 31, no. 4, pp. 2000–2016, 2010.

- [8] L. Li and K.-C. Toh, “An inexact interior point method for l_1 -regularized sparse covariance selection,” *Mathematical Programming Computation*, vol. 2, no. 3-4, pp. 291–315, 2010.
- [9] X. Yuan, “Alternating direction methods for sparse covariance selection,” *preprint*, vol. 2, no. 1, 2009.
- [10] K. Scheinberg, S. Ma, and D. Goldfarb, “Sparse inverse covariance selection via alternating linearization methods,” in *Advances in neural information processing systems*, 2010, pp. 2101–2109.
- [11] J. Duchi, S. Gould, and D. Koller, “Projected subgradient methods for learning sparse Gaussians,” *arXiv preprint arXiv:1206.3249*, 2012.
- [12] B. Rolfs, B. Rajaratnam, D. Guillot, I. Wong, and A. Maleki, “Iterative thresholding algorithm for sparse inverse covariance estimation,” in *Advances in Neural Information Processing Systems*, 2012, pp. 1574–1582.
- [13] C.-J. Hsieh, M. A. Sustik, I. S. Dhillon, and P. Ravikumar, “QUIC: quadratic approximation for sparse inverse covariance estimation.” *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2911–2947, 2014.
- [14] C.-J. Hsieh, A. Banerjee, I. S. Dhillon, and P. K. Ravikumar, “A divide-and-conquer method for sparse inverse covariance estimation,” in *Advances in Neural Information Processing Systems*, 2012, pp. 2330–2338.
- [15] C.-J. Hsieh, M. A. Sustik, I. S. Dhillon, P. K. Ravikumar, and R. Poldrack, “BIG & QUIC: Sparse inverse covariance estimation for a million variables,” in *Advances in neural information processing systems*, 2013, pp. 3165–3173.
- [16] E. Treister and J. S. Turek, “A block-coordinate descent approach for large-scale sparse inverse covariance estimation,” in *Advances in neural information processing systems*, 2014, pp. 927–935.
- [17] S. L. Lauritzen, *Graphical Model*. Oxford University Press, Oxford, 1996.

- [18] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.
- [19] L. R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.
- [20] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael, “Learning low-level vision,” *International journal of computer vision*, vol. 40, no. 1, pp. 25–47, 2000.
- [21] M. Fishelson and D. Geiger, “Exact genetic linkage computations for general pedigrees,” *Bioinformatics*, vol. 18, no. suppl_1, pp. S189–S198, 2002.
- [22] A. Raval, Z. Ghahramani, and D. L. Wild, “A Bayesian network model for protein fold and remote homologue recognition,” *Bioinformatics*, vol. 18, no. 6, pp. 788–801, 2002.
- [23] T. Zhang, “Solving large scale linear prediction problems using stochastic gradient descent algorithms,” in *Proceedings of the twenty-first international conference on Machine learning*. ACM, 2004, p. 116.
- [24] C. Lanczos, *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*. United States Governm. Press Office Los Angeles, CA, 1950.
- [25] G. W. Stewart, “A Krylov–Schur algorithm for large eigenproblems,” *SIAM Journal on Matrix Analysis and Applications*, vol. 23, no. 3, pp. 601–614, 2002.
- [26] L. Vandenberghe, S. Boyd, and S.-P. Wu, “Determinant maximization with linear matrix inequality constraints,” *SIAM journal on matrix analysis and applications*, vol. 19, no. 2, pp. 499–533, 1998.
- [27] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

- [28] M. R. Hestenes, “Multiplier and gradient methods,” *Journal of optimization theory and applications*, vol. 4, no. 5, pp. 303–320, 1969.
- [29] M. J. Powell, “A method for nonlinear constraints in minimization problems,” *Optimization*, pp. 283–298, 1969.
- [30] R. T. Rockafellar, “The multiplier method of Hestenes and Powell applied to convex programming,” *Journal of Optimization Theory and applications*, vol. 12, no. 6, pp. 555–562, 1973.
- [31] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.
- [32] M. Wright, “The interior-point revolution in optimization: history, recent developments, and lasting consequences,” *Bulletin of the American mathematical society*, vol. 42, no. 1, pp. 39–56, 2005.
- [33] Y. Nesterov, “Smooth minimization of non-smooth functions,” *Mathematical programming*, vol. 103, no. 1, pp. 127–152, 2005.
- [34] I. Daubechies, M. Defrise, and C. De Mol, “An iterative thresholding algorithm for linear inverse problems with a sparsity constraint,” *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, vol. 57, no. 11, pp. 1413–1457, 2004.
- [35] C. Wang, D. Sun, and K.-C. Toh, “Solving log-determinant optimization problems by a Newton-CG primal proximal point algorithm,” *SIAM Journal on Optimization*, vol. 20, no. 6, pp. 2994–3013, 2010.
- [36] J. Yang, D. Sun, and K.-C. Toh, “A proximal point algorithm for log-determinant optimization with group lasso regularization,” *SIAM Journal on Optimization*, vol. 23, no. 2, pp. 857–893, 2013.

- [37] C. Wang, “On how to solve large-scale log-determinant optimization problems,” *Computational Optimization and Applications*, vol. 64, no. 2, pp. 489–511, 2016.
- [38] T. Nakagaki, M. Fukuda, S. Kim, and M. Yamashita, “A dual spectral projected gradient method for log-determinant semidefinite problems,” *Computational Optimization and Applications*, pp. 1–36, 2020.
- [39] E. G. Birgin, J. M. Martínez, and M. Raydan, “Nonmonotone spectral projected gradient methods on convex sets,” *SIAM Journal on Optimization*, vol. 10, no. 4, pp. 1196–1211, 2000.
- [40] Y. Nesterov, “A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$,” in *Doklady AN USSR*, vol. 269, 1983, pp. 543–547.
- [41] J. M. Hammersley and P. Clifford, “Markov fields on finite graphs and lattices,” *Unpublished manuscript*, vol. 46, 1971.
- [42] N. Meinshausen, P. Bühlmann *et al.*, “High-dimensional graphs and variable selection with the lasso,” *The annals of statistics*, vol. 34, no. 3, pp. 1436–1462, 2006.
- [43] J. Dieudonné, *Foundations of modern analysis*. Read Books Ltd, 2011.
- [44] G. H. Golub and C. F. Van Loan, *Matrix computations*. JHU press, 2012, vol. 3.
- [45] J. M. Steele, *The Cauchy-Schwarz master class: an introduction to the art of mathematical inequalities*. Cambridge University Press, 2004.
- [46] W. H. Young, “On classes of summable functions and their fourier series,” *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, vol. 87, no. 594, pp. 225–229, 1912.
- [47] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

- [48] S. Boyd, L. Xiao, and A. Mutapcic, “Subgradient methods,” *lecture notes of EE392o, Stanford University, Autumn Quarter*, vol. 2004, pp. 2004–2005, 2003.
- [49] K. Healy and L. W. Schruben, “Retrospective simulation response optimization,” in *1991 Winter Simulation Conference Proceedings*. IEEE, 1991, pp. 901–906.
- [50] S. M. Robinson, “Analysis of sample-path optimization,” *Mathematics of Operations Research*, vol. 21, no. 3, pp. 513–528, 1996.
- [51] H. Robbins and S. Monro, “A stochastic approximation method,” *The annals of mathematical statistics*, pp. 400–407, 1951.
- [52] J. Kiefer, J. Wolfowitz *et al.*, “Stochastic estimation of the maximum of a regression function,” *The Annals of Mathematical Statistics*, vol. 23, no. 3, pp. 462–466, 1952.
- [53] H. Robbins and S. Monro, “A stochastic approximation method,” *The annals of mathematical statistics*, pp. 400–407, 1951.
- [54] D. A. Harville, “Use of the Gibbs sampler to invert large, possibly sparse, positive definite matrices,” *Linear algebra and its applications*, vol. 289, no. 1, pp. 203–224, 1999.
- [55] S. Geman and D. Geman, “Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images,” *IEEE Transactions on pattern analysis and machine intelligence*, no. 6, pp. 721–741, 1984.
- [56] E. S. Coakley and V. Rokhlin, “A fast divide-and-conquer algorithm for computing the spectra of real symmetric tridiagonal matrices,” *Applied and Computational Harmonic Analysis*, vol. 34, no. 3, pp. 379–414, 2013.
- [57] D. Chu and M. Chu, “Reachable matrices by the Qr iteration with shift.”

- [58] D. C. Sorensen, “Implicit application of polynomial filters in ak-step Arnoldi method,” *Siam journal on matrix analysis and applications*, vol. 13, no. 1, pp. 357–385, 1992.
- [59] D. Calvetti, L. Reichel, and D. C. Sorensen, “An implicitly restarted Lanczos method for large symmetric eigenvalue problems,” *Electronic Transactions on Numerical Analysis*, vol. 2, no. 1, p. 21, 1994.
- [60] R. B. Lehoucq, D. C. Sorensen, and C. Yang, *ARPACK users’ guide: solution of large-scale eigenvalue problems with implicitly restarted Arnoldi methods*. Siam, 1998, vol. 6.
- [61] H. Uhlig, “On singular Wishart and singular multivariate beta distributions,” *The Annals of Statistics*, pp. 395–405, 1994.
- [62] M. Johnson, J. Saunderson, and A. Willsky, “Analyzing hogwild parallel Gaussian Gibbs sampling,” in *Advances in Neural Information Processing Systems*, 2013, pp. 2715–2723.
- [63] A. Galli and H. Gao, “Rate of convergence of the gibbs sampler in the Gaussian case,” *Mathematical Geology*, vol. 33, no. 6, pp. 653–677, 2001.
- [64] A. S. Householder, *The theory of matrices in numerical analysis*. Courier Corporation, 2013.
- [65] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge university press, 2012.
- [66] A. Rakhlin, O. Shamir, and K. Sridharan, “Making gradient descent optimal for strongly convex stochastic optimization,” *arXiv preprint arXiv:1109.5647*, 2011.
- [67] O. Shamir and T. Zhang, “Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes.” in *ICML (1)*, 2013, pp. 71–79.

-
- [68] M. D. Zeiler, “ADADELTA: an adaptive learning rate method,” *arXiv preprint arXiv:1212.5701*, 2012.
- [69] H. Yu and J. Dauwels, “Modeling spatio-temporal extreme events using graphical models,” *IEEE Transactions on Signal Processing*, vol. 64, no. 5, pp. 1101–1116, 2016.
- [70] A. Dobra, “Variable selection and dependency networks for genomewide data,” *Biostatistics*, vol. 10, no. 4, pp. 621–639, 2009.
- [71] A. Wille, P. Zimmermann, E. Vranová, A. Fürholz, O. Laule, S. Bleuler, L. Hennig, A. Prelić, P. von Rohr, L. Thiele *et al.*, “Sparse graphical Gaussian modeling of the isoprenoid gene network in *Arabidopsis thaliana*,” *Genome biology*, vol. 5, no. 11, p. R92, 2004.
- [72] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri *et al.*, “Molecular classification of cancer: class discovery and class prediction by gene expression monitoring,” *science*, vol. 286, no. 5439, pp. 531–537, 1999.
- [73] K. Y. Yeung, R. E. Bumgarner, and A. E. Raftery, “Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data,” *Bioinformatics*, vol. 21, no. 10, pp. 2394–2402, 2005.
- [74] H. Liu, K. Roeder, and L. Wasserman, “Stability approach to regularization selection (stars) for high dimensional graphical models,” in *Advances in neural information processing systems*, 2010, pp. 1432–1440.
- [75] D. P. Bertsekas, “Nonlinear programming,” *Journal of the Operational Research Society*, vol. 48, no. 3, pp. 334–334, 1997.
- [76] J. H. Bramble, J. E. Pasciak, and A. T. Vassilev, “Analysis of the inexact Uzawa algorithm for saddle point problems,” *SIAM Journal on Numerical Analysis*, vol. 34, no. 3, pp. 1072–1092, 1997.

- [77] A. Nedić and A. Ozdaglar, “Subgradient methods for saddle-point problems,” *Journal of optimization theory and applications*, vol. 142, no. 1, pp. 205–228, 2009.
- [78] M. Baes, M. Buringisser, and A. Nemirovski, “A randomized mirror-prox method for solving structured large-scale matrix saddle-point problems,” *SIAM Journal on Optimization*, vol. 23, no. 2, pp. 934–962, 2013.
- [79] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex analysis and minimization algorithms I: Fundamentals*. Springer science & business media, 2013, vol. 305.
- [80] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, “Robust stochastic approximation approach to stochastic programming,” *SIAM Journal on optimization*, vol. 19, no. 4, pp. 1574–1609, 2009.
- [81] Y. Xu, “Primal-dual stochastic gradient method for convex programs with many functional constraints,” *arXiv preprint arXiv:1802.02724*, 2018.
- [82] J. Pittman, E. Huang, H. Dressman, C.-F. Horng, S. H. Cheng, M.-H. Tsou, C.-M. Chen, A. Bild, E. S. Iversen, A. T. Huang *et al.*, “Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes,” *Proceedings of the National Academy of Sciences*, vol. 101, no. 22, pp. 8431–8436, 2004.
- [83] I. Hedenfalk, D. Duggan, Y. Chen, M. Radmacher, M. Bittner, R. Simon, P. Meltzer, B. Gusterson, M. Esteller, M. Raffeld *et al.*, “Gene-expression profiles in hereditary breast cancer,” *New England Journal of Medicine*, vol. 344, no. 8, pp. 539–548, 2001.
- [84] P. E. Jacob, J. O’Leary, and Y. F. Atchadé, “Unbiased Markov chain Monte Carlo with couplings,” *arXiv preprint arXiv:1708.03625*, 2017.
- [85] J. Heng and P. E. Jacob, “Unbiased Hamiltonian Monte Carlo with couplings,” *Biometrika*, vol. 106, no. 2, pp. 287–302, 2019.

- [86] R. Mazumder and T. Hastie, “Exact covariance thresholding into connected components for large-scale graphical lasso,” *Journal of Machine Learning Research*, vol. 13, no. Mar, pp. 781–794, 2012.