INFORMATION-THEORETIC MEASURES IN SELECTED LEARNING PROBLEMS

A Dissertation

by

RUIDA ZHOU

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Chao Tian
Committee Members,	Dileep Kalathil
	P. R. Kumar
	Tie Liu
	Zhangyang Wang
Head of Department,	Costas Georghiades

August 2023

Major Subject: Electrical Engineering

ABSTRACT

We study the usage of information-theoretic measures in learning problems.

The first problem considered is the algorithm-dependent generalization error bound. Conceptually, the mutual information between the output of the learning algorithm and training samples captures the amount of information the algorithm learned from the samples, which reflects the overfitting. This motivated the studies on mutual information-based generalization error bound. We propose the individually conditional individual mutual information (ICIMI) bound based on a combination of the error decomposition technique of Bu et al. and the conditional mutual information (CMI) construction of Steinke and Zakynthinou. It combines the merits of the existing studies and provides a tighter bound, and in the process of establishing this bound, we introduce a conditional decoupling lemma, which allows us the view the existing bounds in a unified framework. We further propose a stochastic chaining method, which applies the chaining technique in conjunction with information-theoretic measures to bound the generalization error. The stochastic chaining method borrowed intuition from successive refinement and is more flexible than the previous deterministic chaining approach in conjunction with information-theoretic bounds. We finally provide a new information-theoretic generalization error bound that is exactly tight (i.e., matching even the constant) for the canonical quadratic Gaussian mean estimation problem. Despite considerable existing efforts in deriving information-theoretic generalization error bounds, applying them to this simple setting where sample average is used as the estimate of the mean value of Gaussian data has not yielded satisfying results.

Besides capturing the interplay between learning algorithms and samples, information measures can also be useful to characterize the complexity of the problems. We study the effect of reward variance heterogeneity in the approximate top-m arm identification problem. In this problem, the rewards of pulling each arm are sub-Gaussian but with different variance-proxies, and the agent needs to incorporate this knowledge to minimize the expected number of arm pulls to identify m arms with the largest means in probably approximately sense. The worst-case sample complexity of this problem is characterized by a divide-and-conquer style algorithm and a matching lower bound. The sample complexity reveals that the effect of the reward variance heterogeneity is quantified by an Entropy-like function of the variances.

In addition to bounding and characterizing certain performance metrics, information measures can also facilitate the design of algorithms. We study the policy optimization in multi-objective reinforcement learning and propose an Anchor-changing Regularized Natural Policy Gradient (ARNPG) framework, which can systematically incorporate ideas from well-performing first-order methods into the design of policy optimization algorithms for multi-objective Markov decision process (MDP) problems. The ARNPG framework introduces Kullback-Leibler divergences with changing anchors as regularization to the intermediate policy update, which enables acceleration as well as bridging the analysis between the policy gradient update and the incorporated first-order methods. Under softmax parameterization with exact gradients, the proposed algorithms inherit the advantages of the integrated first-order methods and are guaranteed to have $\tilde{O}(1/T)$ global convergence without further assumptions on the underlying MDP. Experiments are further provided to demonstrate that the proposed algorithms provide superior performance.

DEDICATION

To my wife and our parents, for their love and support.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my advisor, Professor Chao Tian. Prof. Tian's mentorship is invaluable. His guidance and unwavering belief in my abilities have been instrumental in shaping my research and academic growth. In retrospect, many memories come to my mind. He illustrated the beauty of Shannon's creative thinking to me; he taught me how to pursue an idea and formulate a problem by thinking big and concisely; he gave me the freedom to work on things that I found exciting and encouraged me to think independently. I am indebted to Prof. Tian for the countless discussions, feedback, and opportunities he has provided me throughout my Ph.D. journey. Thank you, Prof. Tian!

I am also grateful to Professor P. R. Kumar and Professor Dileep Kalathil. Your courses introduced the world of stochastic systems and reinforcement learning to me. The numerous discussions and brainstorming we had and your insightful suggestions on problem solving and presentation are wealth for my future research. Additionally, Professor Kumar's slides on “graduate studies, research, and life thereafter" helped me through the wilderness time in my graduate study. I am truly thankful for your wisdom and mentorship. I would also like to express my gratitude to Professor Tie Liu for his valuable help and suggestions on research, career, and life. Knowing that I can turn to you whenever I face troubles in research or in life is a great source of strength. I would also like to thank Professor Zhangyang Wang for his time and valuable feedback. I am grateful to have you all on my committee.

I would also like to express my gratitude to Professor Yang Shen, who kindly serve on my committee from spring 2020 to spring 2022, and Professor Xiaoning Qian, who give me a chance to explore new topics in his lab in summer 2023. I am thankful for the opportunity to work in the AT&T Lab-Research with Dr. Huahui Wang, Dr. Wenjie Zhao and Dr. Jie Chen during the summers of 2021 and 2022. I would also like to thank Professor Cong Shen and Professor Jing Yang for introducing me to the joys of research during my undergraduate studies and for inspiring me to explore new and interesting problems throughout my graduate journey. I am grateful for your mentorship and guidance. I am fortunate to have a wonderful group of collaborators and friends: Tao Liu, Prof. Hua Sun, Li Fan, Kun Yang, Chengshuai Shi, Min Cheng, with whom discussions are inspiring and enjoyable; my current and previous colleagues in lab Qiang Zhang, Dr. Tao Guo, Wenjing Chen, Chengyuan Qian, Tianli Zhou, Kai Zhang, and Lei Zheng, who make coming to work a pleasure; my previous roommates Kaixiong Zhou and Jianhao Chen, who have helped me a lot during the beginning years of my graduate study; and Siqi Fan, Yuning You, Tianlong Chen, Chenjie Luo, Pengcheng Pi, Jianfeng Song, who make my life in this small town interesting.

Finally, I want to express my gratitude to my wife Xinyu, and our parents. Your love, support, and understanding have been the driving force behind my graduate life.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Professor Chao Tian, Professor Dileep Kalathil, Professor P. R. Kumar and Professor Tie Liu of the Department of Electrical and Computer Engineering and Professor Zhangyang Wang of the Department of Computer Science.

The analyses depicted in Chapter 4 were conducted in part by Tao Liu. All other work conducted for the dissertation was completed by the student independently.

Funding Sources

Graduate study was supported by National Science Foundation under Grant CCF-1816546 and Grant CCF-2007067.

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

Information-theoretic methods have been playing important roles in many fields, such as communication, data storage, cryptography, e.c.t., among which machine learning has recently attracted a lot of attention. Lying in the heart of the information-theoretic methods, information-theoretic measures provide us with the tools for better interpretation, precise characterization, and constructive intuition. In this dissertation, we study the usage of information-theoretic measures in some learning problems. Specifically, we study (1) interpreting generalization error by mutual information-based bound, (2) characterizing complexity of top-m arm identification with heterogeneous reward variances problem via Shannon's entropy, and (3) designing policy gradient-based algorithm for multi-objective Markov decision process based on the idea of anchor changing Kullback-Leiber divergence regularization.

1.1 Generalization Error Bounded by Mutual Information

The generalization error of a supervised learning algorithm is defined to be the difference between the empirical risk and the population risk, which is used to quantify the degree to which a learning algorithm may overfit the training data. Bounding the generalization error of learning algorithms is of fundamental importance in statistical machine learning. The conventional approach is to bound it using a quantity related to the hypothesis class, such as the VC-dimension [1], and such bounds are therefore oblivious to the learning algorithm and data distribution. The obtained results are usually rather conservative, and cannot fully explain the recent success of deep learning. Recently, information-theoretic approaches that jointly take into consideration the hypothesis class, the learning algorithm, and the data distribution, have drawn considerable attention [2–11].

The effort of deriving generalization error bounds using information-theoretic approaches was perhaps first initiated in [2] and [3]. The bound was further tightened in [8], by decomposing the error, and bounding each term individually. Steinke and Zakynthinou [9] proposed a conditional mutual information (CMI) based bound, by introducing a dependence structure which resembles that in the analysis of the Rademacher complexity [1]. Combining the idea of error decomposition [8] and the CMI bound in [9], Haghifam et al. [10] subsequently provided a sharpened bound based on conditional individual mutual information (CIMI).

We propose another generalization error bound, which is also based on a combination of the error decomposition technique and the CMI construction. This bound is motivated by the observation that in a simple Gaussian setting, the CIMI bound in [10] (as well as the CMI bound in [9]) is of constant order, while the bound in [8] is of order $\Theta(\frac{1}{\sqrt{n}})$, where n is the number of training samples. The conditioning term in CIMI is the same as CMI, and it tends to reveal too much information which makes the bounds loose. The proposed bound is thus obtained by making the mutual information conditioned on an individual sample (pair), which we refer to as the individually conditional individual mutual information (ICIMI) bound.

We propose a new approach to applying the chaining technique in conjunction with information-theoretic measures to bound the generalization error of machine learning algorithms. Different from the deterministic chaining approach based on hierarchical partitions of a metric space, previously proposed by Asadi et al., we propose a stochastic chaining approach, which replaces the hierarchical partitions with an abstracted Markovian model borrowed from successive refinement source coding. This approach has three benefits over deterministic chaining: 1) the metric space is not necessarily bounded, 2) facilitation of subsequent analysis to yield a more explicit bound, and 3) further opportunity to optimize the bound by removing the geometric rigidity of the partitions. The proposed approach includes traditional chaining methods as a special case, and can therefore also utilize any deterministic chaining construction. We illustrate these benefits using the problem of estimating the Gaussian mean and that of phase retrieval. For the former, we derive a bound that provides an order-wise improvement over previous results, and for the latter, we provide a stochastic chain that allows optimization over the chaining parameter.

We provide a new information-theoretic bound that is exactly tight (i.e., matching even the constant) for the canonical quadratic Gaussian problem. In fact, most existing bounds are order-wise loose in this setting, which has raised concerns about the fundamental capability of information-theoretic bounds in reasoning the generalization behavior for machine learning. The proposed new bound adopts the individual-sample-based approach proposed by Bu et al. [8], but also has several key new ingredients. Firstly, instead of applying the change of measure inequality on the loss function, we apply it to the generalization error function itself; secondly, the bound is derived in a conditional manner; lastly, a reference distribution, which bears a certain similarity to the prior distribution in the Bayesian setting, is introduced. The combination of these components produces a general KL-divergence-based generalization error bound. We further show that although the conditional bounding and the reference distribution can make the bound exactly tight, removing them does not significantly degrade the bound, which results in a mutual-information-based bound that is also asymptotically tight in this setting.

1.2 The Sample Complexity of Top-m Arm Identification with Heterogeneous Reward Variances Measured by Entropy

Approximate top-m arm identification with fixed confidence is a formal Probably Approximately Correct (PAC)-learning formulation for the best arm identification setting, where the agent is required to identify the top-m arms, where the expected rewards of the m arms identified are not less than that of the m-th best arm by ϵ, with confidence at least $1-\delta$. In this setting, the algorithms will have a performance guarantee in terms of the confidence of success as well as the precision ϵ. We refer to this setting as (ϵ, δ) top-m arm identification.

In most previous works on multi-armed bandits, an inherent assumption is that the reward distribution of each arm is sub-Gaussian, and moreover, the variance proxies are known and the same, i.e., homogeneous among all the arms. Such an assumption may be natural when the rewards are bounded in a range, or it is reasonable to view the arms as of the same random nature (except the reward means of the arms). In other applications, this assumption is less suitable, since the reward distributions are naturally heterogeneous. We consider (ϵ, δ) best m-arm identification with sub-Gaussian distributed rewards when the variance proxies are heterogeneous and known.

Several well-known algorithms can be straightforwardly adapted to the problem under consideration. We observe that the adapted algorithms only perform well in some respective cases. More precisely, the adapted naive elimination algorithm performs well when the heterogeneity is more significant, and the adapted median elimination algorithm performs well when the heterogeneity is less significant. Given this observation, we seek for a new algorithm that can naturally account for the heterogeneity, and propose the variance-grouped median elimination algorithm. There is no need to artificially ascribe an instance as having either high or low heterogeneity in this algorithm, and its performance adapts naturally. We further establish a matching lower bound by reformulating it into an optimization problem and considering its dual. Combined with this lower bound, we show the proposed algorithm is optimal.

We show that the worst-case sample complexity of this problem is

$$\Theta\left(\sum_{i=1}^{n}\frac{\sigma_i^2}{\epsilon^2}\ln\frac{1}{\delta}+\sum_{i\in G^m}\frac{\sigma_i^2}{\epsilon^2}\ln(m)+\sum_{j\in G^l}\frac{\sigma_j^2}{\epsilon^2}\mathrm{Ent}(\sigma_{G^r}^2)\right), \tag{1.1}$$

where G^m, G^l, G^r are certain specific subsets of the overall arm set $\{1,2,\ldots,n\}$, and Ent$(\cdot)$ is an entropy-like function which measures the heterogeneity of the variance proxies. The worst-case sample complexity is in general proportional to the sum of the reward variances and has three components. The first component (with $\ln\frac{1}{\delta}$) reflects the effect of the confidence parameter, the second component reflects the impact of the more homogeneous subset of the arms, and the last term (with the $\mathrm{Ent}(\cdot)$ function) reflects the impact of the more heterogeneous subset of the arms. The result naturally degrades if the reward variances are indeed homogeneous, which essentially has only the first two components. The third component captures the impact of the heterogeneity, which is not critically related to m, but on the variances $\sigma_{1:n}^2$ through an entropy-like function. For highly heterogeneous variances, the second term will disappear, and Ent$(\sigma_{G^r}^2)$ can be of order $O(1)$, thus becoming independent of m completely.

1.3 Policy Optimization Regularized by Kullback-Leibler Divergence for Multi-Objective Markov Decision Process

In many sequential decision-making scenarios, agents usually face multiple objectives simultaneously. This motivates the study of reinforcement learning (RL) with multiple reward values

$V^{\pi}_{1:m}(\rho)$. The agent exploits certain criteria to reflect the system requirement and aims to optimize the policy such that its values $V^{\pi}_{1:m}(\rho)$ satisfies the criteria.

We study policy gradient-based approaches that optimize over parameterized policies $\Pi = \{\pi_\theta : \theta \in \Theta\}$ through policy gradient. In general, the optimization problems above may not be convex in terms of θ, not even for single-objective MDPs with direct parameterization by $\theta_{s,a} = \pi_\theta(a|s)$ [12]. Due to the non-convexity, $O(1/T)$ global convergence of policy gradient-based methods was only established very recently for single-objective MDPs with exact gradients [12, 13]. These breakthrough results have motivated the study of policy optimization for multi-objective MDPs, e.g., smooth concave scalarization [14], constrained MDPs (CMDPs) [15, 16].

However, under the exact gradients scenario, the previous approaches for multi-objective MDPs, either suffer from slow provable $O(1/\sqrt{T})$ global convergence [15], or require extra assumptions [17–19]. The compactness of Θ is assumed in [17], but this assumption forbids a very common softmax parameterization, where $\Theta = \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$. The NPG-based methods have been analyzed in [18, 19] under an ergodicity assumption, but such an assumption is not required for NPG in single-objective MDPs [12], and therefore appears artificial.

Many criteria for multi-objective MDPs could be viewed as convex optimization problems w.r.t. a value vector $v \in \mathcal{V}$, for which there is a wide array of well-performing first-order methods for convex optimization problems in general. It is desirable to take full advantage of such efficient first-order methods in a unified and flexible manner when designing policy gradient-based algorithms for multi-objective MDPs.

We propose an anchor-changing regularized natural policy gradient (ARNPG) framework that can exploit and integrate first-order methods for the design of policy gradient-based algorithms for multi-objective MDPs. We introduce Kullback-Leibler divergences with changing anchors in the ARNPG framework as regularization to the intermediate policy update. This regularization accelerates the policy update due to its local strong convexity, and meanwhile, the changing anchors reduce the bias caused by introducing regularization to the original problem. Analytically, the divergences bridge the analysis between policy gradient and the incorporated first-order methods.

We demonstrate the strength of the ARNPG framework by designing algorithms for three general criteria: smooth concave scalarization, constrained MDPs, and max-min trade-off. Under softmax parameterization with exact gradients, the proposed algorithms inherit the advantages of the integrated first-order methods and are guaranteed to have $\tilde{O}(1/T)$ global convergence without further assumptions on the underlying MDP. In addition to the theoretical advantages, we provide the results of extensive experimentation which demonstrate that the ARNPG-guided algorithms provide superior performance in exact gradient and sample-based tabular scenarios, as well as actor-critic deep RL scenarios, compared to several existing policy gradient-based approaches.

2. INFORMATION-THEORETIC BOUNDS ON GENERALIZATION ERROR*

In this chapter, we proposed the individually conditional individual mutual information (ICIMI) bound to upper bound the generalization error. To establish the proposed bound, we introduce a new conditional decoupling lemma. This lemma allows us to view the bounds in many of the previous works [3, 8–10] and the proposed bound in a unified manner, which not only yields a dichotomy of these bounds, but also makes possible a meaningful comparison among them. It also allows us to take expectation of the conditionals outside of the concave conjugate function, which may significantly tighten the bound when the Jensen gap is large. Finally, we show that in the Gaussian setting mentioned earlier, the proposed new bound is also able to provide a bound of the same order as, but with an improved leading constant than that in [8].

As an application of the proposed ICIMI bound, we apply it on a logistic regression setting where the mutual information terms need to be estimated from data, and it yields a generalization bound similar to that in [8]. The CMI bound and the CIMI bound, on the other hand, are much more difficult to estimate since they involve many more random variables. As another application, we further analyze the noisy and iterative stochastic gradient Langevin dynamics (SGLD) algorithm, which includes the Langevin dynamics algorithm as a special case when the full batch is used, and derive an upper bound on its generalization error, which is more general than previous results (e.g., no requirement for the loss function to be bounded).

2.1 Preliminaries

System model. We study the classic supervised learning setting. Denote the data domain as $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the feature domain and $\mathcal{Y}$ is the label set. The parametric hypothesis class

*Part of this chapter is reprinted with permissions from “Individually conditional individual mutual information bound on generalization error," by Ruida Zhou, Chao Tian, and Tie Liu, 2022, *IEEE Transactions on Information Theory*, vol. 68, no. 5, pp. 3304– 3316, Copyright © 2022 IEEE [20], “Stochastic chaining and strengthened information-theoretic generalization bounds" by Ruida Zhou, Chao Tian, and Tie Liu, 2023, *Journal of the Franklin Institute*, vol.360, no. 6, pp. 4114-4134, Copyright © 2023 The Franklin Institute [21], and “Exactly Tight Information-Theoretic Generalization Error Bound for the Quadratic Gaussian Problem" by Ruida Zhou, Chao Tian, and Tie Liu, 2023, *IEEE International Symposium on Information Theory (ISIT)*, Copyright © 2023 IEEE [22].

is denoted as $\mathcal{H}_{\mathcal{W}} = \{h_W : W \in \mathcal{W}\} \subseteq \mathcal{Y}^{\mathcal{X}}$, where $\mathcal{W}$ is the parameter space. During training, the learning algorithm (learner) has access to a sequence of training samples $Z_{[n]} = (Z_1, Z_2, \ldots, Z_n)$, where each Z_i is drawn independently from $\mathcal{Z}$ following some unknown probability distribution ξ. The learner can be represented by $P_{W|Z_{[n]}}$, which is a kernel (channel) that (randomly) maps $\mathcal{Z}^n$ to $\mathcal{W}$.

To complete the classification or regression task, the learner in principle would choose a hypothesis $w \in \mathcal{W}$ to minimize the following population loss, under a given loss function $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}$,

$$L_\xi(w) = \mathbb{E}_{Z \sim \xi}[\ell(w, Z)]. \tag{2.1}$$

However, since only a training data vector $Z_{[n]}$ is available, the empirical loss of w is usually computed (and minimized during training), which is given as

$$L_{Z_{[n]}}(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(w, Z_i). \tag{2.2}$$

The expected generalization error of the learner $P_{W|Z_{[n]}}$ is

$$\operatorname{gen}(\xi, P_{W|Z_{[n]}}) := \mathbb{E}\left[L_\xi(W) - L_{Z_{[n]}}(W)\right], \tag{2.3}$$

where the expectation is taken over the distribution $P_{W,Z_{[n]}}$ as the joint distribution implied by the kernel $P_{W|Z_{n]}}$ and the marginal $P_{Z_{[n]}} = \xi^n$. This quantity captures the effect of the learner's expected overfitting error due to limited training data, which we shall study in this work.

Formally we write $\mathbb{E}_{X \sim P}[f(X)] = \int_{\mathcal{X}} f(x) dP(x)$ as the expectation of $f(X)$. When the distribution of X is clear from the context, we omit P and write it as $\mathbb{E}_X[f(X)]$, where the subscript X means the expectation is taken with respect to the random variable X. When the random variable X in $f(X)$ is also clear from the context, we simply write it as $\mathbb{E}[f(X)]$.

We introduce the following quantity

$$\mathrm{gen}_{Z_{[n]}}(\xi, w) := L_\xi(w) - L_{Z_{[n]}}(w), \tag{2.4}$$

which can be viewed as stochastic process indexed by hypothesis w and the expected generalization error can by written as $\mathrm{gen}(\xi, P_{W|Z_{[n]}}) = \mathbb{E}[\mathrm{gen}_{Z_{[n]}}(\xi, W)]$. Generalization error can also be written in a different form by defining

$$\mathrm{gen}^i_{Z_i}(\xi, w) := L_\xi(w) - \ell(w, Z_i), \tag{2.5}$$

$$\mathrm{gen}^i(\xi, P_{W|Z_i}) := \mathbb{E}[L_\xi(W) - \ell(W, Z_i)]. \tag{2.6}$$

Clearly $\mathrm{gen}(\xi, P_{W|Z_{[n]}}) = \frac{1}{n}\sum_{i=1}^n \mathrm{gen}^i(\xi, P_{W|Z_i})$. It is worth noting that the distribution $P_{W|Z_i}$ is obtained by marginalizing over $P(W, Z_{[n]})$ (and dividing ξ).

We will then briefly review a few information-theoretic bounds on the generalization error relevant to this work. A more thorough discussion of their relation is deferred to Section 2.2.4 and 2.2.5, after a unified framework is given.

Mutual information based bounds. Xu and Raginsky, motivated by a previous work by Russo and Zou [2], provided a mutual information (MI) based bound on the expected generalization error [3].

Theorem 1 (MI Bound [3])**.** *Suppose $\ell(w, Z)$ is σ^2-sub-Gaussian under ξ for all $w \in \mathcal{W}$, then*

$$\mathrm{gen}(\xi, P_{W|Z_{[n]}}) \le \sqrt{\frac{2\sigma^2}{n} I\left(W; Z_{[n]}\right)}. \tag{2.7}$$

The generalization can be written in two ways

$$\begin{aligned} \mathrm{gen}(\xi, P_{W|Z_{[n]}}) &= \mathbb{E}\left[L_{\tilde{Z}_{[n]}}(\tilde{W})\right] - \mathbb{E}\left[L_{Z_{[n]}}(W)\right] && (2.8)\\ &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[(\ell(\tilde{W}, \tilde{Z}_i) - \ell(W, Z_i))\right], && (2.9)\end{aligned}$$

where $\tilde{W}$ and $\tilde{Z}_i$ are independent random variables that have the same marginal distributions as W and Z_i, respectively. Instead of bounding the difference (2.8) as in [3], Bu et al. [8] bounded each individual difference in (2.9) and derived an individual mutual information (IMI) based bound. Furthermore, the following inverse Fenchel conjugate function was utilized to obtain a tightened and more general bound. For any random variables F, its cumulant generating function (CGF) is

$$\Lambda_F(\lambda) := \ln \mathbb{E}\left[e^{\lambda F}\right]. \tag{2.10}$$

The CGF $\Lambda_F(\lambda)$ may not exist for some $\lambda \in \mathbb{R}$. Define the extended-value centered CGF of F as $\psi_F(\lambda) := \infty$ for such λ that $\Lambda_F(\lambda)$ does not exist, and $\psi_F := \Lambda_F(\lambda) - \lambda\mathbb{E}[F]$ otherwise. The inverse of its Fenchel conjugate is given as

$$\psi_F^{*-1}(\eta) := \inf_{\lambda>0} \frac{\eta + \psi_F(\lambda)}{\lambda}, \quad \eta \in [0, \infty). \tag{2.11}$$

The tightened bound is summarized in the following theorem.

Theorem 2 (IMI Bound [8]). *Suppose ψ_- is an upper bound of $\psi_{-\ell(\tilde{W},\tilde{Z}_i)}$, then*

$$\text{gen}(\xi, P_{W|Z_{[n]}}) \leq \frac{1}{n}\sum_{i=1}^{n} \psi_-^{*-1}\left(I\left(W; Z_i\right)\right), \tag{2.12}$$

where $\tilde{W}$ and $\tilde{Z}_i$ are independent random variables that have the same marginal distributions as W and Z_i, respectively.

Conditional mutual information-based bounds. Steinke and Zakynthinou [9] recently introduced a novel bounding approach. In their approach, $Z_{[n]}^{\pm} := (Z_1^{\pm 1}, Z_2^{\pm 1}, \ldots, Z_n^{\pm 1})$ is a $2 \times n$ table of samples that each Z_i^s, for $s = -1, 1$ and $i = 1, \ldots, n$ is independently drawn following ξ. The training vector $(Z_1^{R_1}, Z_2^{R_2}, \ldots, Z_n^{R_n})$ is selected from the table $Z_{[n]}^{\pm}$, where R_i's are independent Rademacher random variables, i.e., R_i takes 1 or -1 equally likely. The vector $R_{[n]} = (R_1, \ldots, R_n) \in \{-1, 1\}^n$ essentially selects one sample from each column in the table, which partitions $Z_{[n]}^{\pm}$ into a training vector and a testing vector. For simplicity, we shall write Z_i^{-1}

and Z_i^{+1} as Z_i^- and Z_i^+, when the meaning is clear from the context.

With the structure given above, the expected generalization error can be written as

$$\mathrm{gen}(\xi, P_{W|Z_{[n]}}) = \mathbb{E}_{Z^{\pm}_{[n]}}\left[\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} R_i\left(\ell(W, Z_i^-) - \ell(W, Z_i^+)\right)\Big| Z^{\pm}_{[n]}\right]\right]. \tag{2.13}$$

Steinke and Zakynthinou obtained the following conditional mutual information (CMI) based result.

Theorem 3 (CMI Bound [9])**.** *Suppose* $\sup_{w\in\mathcal{W}} |\ell(w, z_1) - \ell(w, z_2)| \leq \Delta(z_1, z_2)$ *for any* $z_1, z_2 \in \mathcal{Z}$*, then*

$$\mathrm{gen}(\xi, P_{W|Z_{[n]}}) \leq \sqrt{\frac{2}{n}\mathbb{E}[\Delta(Z_1, Z_2)^2] I\left(W; R_{[n]}|Z^{\pm}_{[n]}\right)}, \tag{2.14}$$

where Z_1, Z_2 *are independent samples distributed as* ξ*.*

Since R_i is binary, the conditional mutual information is always bounded; in contrast, mutual information-based bounds (i.e., MI and IMI bounds) can be unbounded, particularly when the random variables W, Z_i are both continuous.

Motivated by the results in [8], Haghifam et al. [10] proposed a sharpened bound by similarly bounding each term in (2.13). Moreover, they provided a conditional individual mutual information (CIMI) based bound represented by *the sample-conditioned mutual information*, which is defined as

$$I_u(X; Y) := I(X; Y|U = u). \tag{2.15}$$

Clearly $I_U(X; Y)$ is a function of the random variable U, thus also a random variable, and $\mathbb{E}[I_U(X; Y)] = I(X; Y|U)$. These sharpened bounds are summarized in the following theorem.

Theorem 4 (CIMI Bound [10]). *Suppose $\ell \in [0,1]$, then*

$$\mathrm{gen}(\xi, P_{W|Z_{[n]}}) \leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\sqrt{2I_{Z^{\pm}_{[n]}}(W;R_i)}\right] \leq \frac{1}{n}\sum_{i=1}^{n}\sqrt{2I\left(W;R_i|Z^{\pm}_{[n]}\right)}. \tag{2.16}$$

2.2 The ICIMI Bound and Its Properties

We first introduce a motivating example which shows that the CMI and CIMI bound can be order-wise worse than the IMI bound. In order to remedy this deficiency, we first introduce an instrumental (conditional decoupling) lemma, then provide the proposed bound, which we refer to as individually conditional individual mutual information (ICIMI) bound. The relation between the proposed bound and the existing bounds is discussed. As a byproduct of the unified view allowed by the aforementioned conditional decoupling lemma, several more general forms of the existing bounds are also given. Finally, we return to the motivating example and show that the proposed bound can indeed provide a bound of the same order as the IMI bound, but with a slightly better constant factor.

2.2.1 A motivating example

Let us consider the simple setting of estimating the mean from samples generated from a Gaussian distribution $N(\mu, \sigma^2)$, by averaging the $i.i.d.$ training samples under the squared loss.

Estimating the Gaussian mean The training samples $Z_{[n]}$ are drawn $i.i.d.$ following $N(\mu,\sigma^2)$ for some unknown μ. The learner deterministically estimates μ by averaging the training samples, i.e., $W = \frac{1}{n}\sum_{i=1}^{n} Z_i$, whose empirical error is

$$L_{Z_{[n]}}(W) = \frac{1}{n}\sum_{i=1}^{n}(W - Z_i)^2. \tag{2.17}$$

Bu et al. [8] showed that the mutual information term in the IMI bound is

$$I(W;Z_i) = \frac{1}{2}\log\frac{n}{n-1} = \frac{1}{2(n-1)} + o\left(\frac{1}{n}\right), \tag{2.18}$$

and obtained the following IMI-based bound

$$\sigma^2\sqrt{\frac{2(n+1)^2}{n^2}\log\frac{n}{n-1}} = \sigma^2\sqrt{\frac{2}{n-1}} + o\left(\frac{1}{\sqrt{n}}\right). \tag{2.19}$$

For this simple setting, the generalization error can in fact be calculated exactly to be $\frac{2\sigma^2}{n}$. Though the error bound above does not have the same order as the true generalization error, it is consistent with the VC dimension-based bound and is the best known for this case. Note that the MI bound will be unbounded, since $I(W; Z_{[n]})$ is unbounded.

Next, consider the CMI and CIMI bounds, and let us focus on the mutual information terms in these bounds, which give

$$I(W; R_{[n]}|Z^{\pm}_{[n]}) = n/\log_2 e, \qquad I_{Z^{\pm}_{[n]}}(W; R_i) = 1/\log_2 e, \quad a.s.. \tag{2.20}$$

It is seen that they are order-wise worse than (2.18), which suggests that the bounds obtained from the CMI and CIMI bounds would be order-wise worse than (2.19).

Theorem 3 and Theorem 4 do not apply directly in this setting, since their required conditions do not hold. In Theorem 3, the function $\Delta(z_1, z_2)$ does not exist (i.e., unbounded); even if it existed, the term $\mathbb{E}[\Delta(Z_1, Z_2)^2]$ would be a constant, thus the CMI bound would be of constant order. Similarly, if the condition $\ell \in [0, 1]$ holds, the CIMI bound would also be of constant order. As we shall show shortly, the CMI and CIMI bounds can be generalized and strengthened, yet the resultant strengthened bounds in this setting still do not diminish as $n \to \infty$, and thus would be order-wise worse than the IMI bound.

A question arises naturally: Is the looseness of the CMI and CIMI bounds here due to the introduction of the conditioning terms? As we shall show next, it is caused by too much information being revealed in the conditioning terms, and there is indeed a natural way to resolve this issue.

2.2.2 A conditional decoupling lemma

Our main result relies on a key lemma. A few more definitions are first introduced to present this lemma and the main result.

For any random variables F and U, define the sample-conditioned CGF for any realization $U = u$,

$$\Lambda_{F|U}(\lambda, u) := \ln \mathbb{E}\left[e^{\lambda F} \middle| U = u\right]. \tag{2.21}$$

Similar to the regular CGF, $\Lambda_{F|U}(\lambda, u)$ may not exist for some $\lambda \in \mathbb{R}$. Define the *extended-value centered sample-conditioned CGF* as $\psi_{F|U}(\lambda, u) := \infty$ for such λ that $\Lambda_{F|U}(\lambda, u)$ does not exist, and $\psi_{F|U}(\lambda, u) := \Lambda_{F|U}(\lambda, u) - \lambda\mathbb{E}[F|U = u]$ otherwise. It is straightforward to verify that for any realization $U = u$, $\psi_{F|U}(0, u) = \psi'_{F|U}(0, u) = 0$ and $\psi''_{F|U}(0, u) > 0$. Hence the inverse of its Fenchel conjugate

$$\psi^{*-1}_{F|U}(\eta, u) := \inf_{\lambda>0} \frac{\eta + \psi_{F|U}(\lambda, u)}{\lambda}, \quad \eta \in [0, \infty) \tag{2.22}$$

is concave and non-decreasing; see e.g., [8] and [23]. The unconditioned version of this function was utilized earlier by Asadi et al. [4] and Bu et al. [8]. When it is clear from context, we will write

$$\Psi_{F|U}(\lambda) := \psi_{F|U}(\lambda, U), \quad \Psi^{*-1}_{F|U}(\eta) := \psi^{*-1}_{F|U}(\eta, U), \tag{2.23}$$

which are functions of U, thus random. Next, define the *extended-value centered conditional CGF*

$$\bar{\psi}_{F|U} = \mathbb{E}\left[\Psi_{F|U}\right], \tag{2.24}$$

and similarly its inverse Fenchel conjugate as $\bar{\psi}^{*-1}_{F|U}$.

For a pair of random variables (X, Y), its *decoupled pair conditioned on a third random vari-*

able U is a pair of random variables $(\tilde{X}, \tilde{Y})$, such that

$$(\tilde{X}, U) \overset{D}{=} (X, U), \quad (\tilde{Y}, U) \overset{D}{=} (Y, U), \tag{2.25}$$

i.e., $(\tilde{X}, U)$ and (X, U) are identically distributed, and $(\tilde{Y}, U)$ and (Y, U) are identically distributed, and moreover

$$\tilde{X} \leftrightarrow U \leftrightarrow \tilde{Y} \tag{2.26}$$

forms a Markov string. It follows from this definition that

$$I_U(X;Y) = D(P_{X,Y|U} || P_{\tilde{X},\tilde{Y}|U}). \tag{2.27}$$

We next introduce a conditional decoupling (CD) lemma, which serves an instrumental role in our work. The unconditioned version was presented in [8].

Lemma 1 (The CD lemma)**.** *For any three random variables X, Y, U, let $\tilde{X}, \tilde{Y}$ be the decoupled pair of X, Y conditioned on U. Let $F := f(X, Y)$ and $\tilde{F} := f(\tilde{X}, \tilde{Y})$, for some real-valued measurable function f. The following inequalities hold*

$$\mathbb{E}[F] - \mathbb{E}[\tilde{F}] \leq \mathbb{E}\left[\Psi^{*-1}_{\tilde{F}|U}\left(I_U(X;Y)\right)\right] \leq \bar{\psi}^{*-1}_{\tilde{F}|U}\left(I(X;Y|U)\right). \tag{2.28}$$

2.2.3 The ICIMI bound

For each $i = 1, \ldots, n$, let $(\tilde{W}_i, \tilde{R}_i)$ be a decoupled pair of (W, R_i) conditioned on $Z_i^{\pm}$. The bound we propose is presented in Theorem 5.

Theorem 5. *(ICIMI Bound) Given an algorithm $P_{W|Z_{[n]}}$, the following bounds on the generalization hold*

$$\mathrm{gen}(\xi, P_{W|Z_{[n]}}) \leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\Psi^{*-1}_{\tilde{G}_i|Z_i^{\pm}}(I_{Z_i^{\pm}}(W;R_i))\right] \tag{2.29}$$

$$\leq \frac{1}{n}\sum_{i=1}^{n} \bar{\psi}^{*-1}_{\tilde{G}_i|Z_i^{\pm}}(I(W;R_i|Z_i^{\pm})), \tag{2.30}$$

where $\tilde{G}_i = \tilde{R}_i \left(\ell(\tilde{W}_i, Z_i^-) - \ell(\tilde{W}_i, Z_i^+)\right)$.

There are two bounds in this theorem. The stronger bound is in terms of the sample-conditioned mutual information, which is different from the conventional notion of conditional mutual information. The weaker bound is in terms of conventional mutual information.

In the proposed bounds, the mutual information is conditioned on the individual data pair $Z_i^{\pm}$, instead of the full data pair set $Z_{[n]}^{\pm}$. Intuitively, revealing only $Z_i^{\pm}$ makes it more difficult, than revealing all data pairs $Z_{[n]}^{\pm}$, to deduce information regarding R_i from W. As a consequence, the mutual information $I(W;R_i|Z_i^{\pm})$ is always smaller or equal to $I(W;R_i|Z_{[n]}^{\pm})$, which is formally shown in Lemma 2, yielding potentially tighter bound.

Proof of Theorem 5. We can rewrite the generalization error given in (2.13) as

$$\mathrm{gen}(\xi, P_{W|Z_{[n]}}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\mathbb{E}\left[R_i\left(\ell(W,Z_i^-) - \ell(W,Z_i^+)\right)|Z_i^{\pm}\right]\right]. \tag{2.31}$$

Now apply the CD lemma on each individual term in (2.31) by letting $X = W$, $Y_i = R_i$, $U_i = Z_i^{\pm}$, and $F_i = R_i\left(\ell(W,Z_i^-) - \ell(W,Z_i^+)\right)$. Since

$$\mathbb{E}[\tilde{G}_i] = \mathbb{E}[\tilde{F}_i] = \mathbb{E}\left[\tilde{R}_i\left(\ell(\tilde{W}_i,Z_i^-) - \ell(\tilde{W}_i,Z_i^+)\right)\right] = 0,$$

we have

$$\begin{aligned}\mathrm{gen}(\xi, P_{W|Z_{[n]}}) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[F_i] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[F_i] - \mathbb{E}[\tilde{F}_i] \\ &\leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\Psi^{*-1}_{\tilde{G}_i|Z_i^{\pm}}(I_{Z_i^{\pm}}(W;R_i))\right] \qquad (2.32) \\ &\leq \frac{1}{n}\sum_{i=1}^{n}\bar{\psi}^{*-1}_{\tilde{G}_i|Z_i^{\pm}}(I(W;R_i|Z_i^{\pm})), \qquad (2.33)\end{aligned}$$

which completes the proof. □

We call this bound the individually conditional individual mutual information (ICIMI) bound, since it is derived by applying the CD lemma on the individual conditional terms in (2.31). We note that Theorem 5 implies Proposition 3 in [24], which we state below as a corollary.

Corollary 1. *Suppose $\ell \in [a, b]$ with $a < b$, then*

$$\begin{aligned} \mathrm{gen}(\xi, P_{W|Z_{[n]}}) &\leq \frac{b-a}{n} \sum_{i=1}^{n} \mathbb{E}\left[\sqrt{2I_{Z_i^{\pm}}(W; R_i)}\right] && (2.34) \\ &\leq \frac{b-a}{n} \sum_{i=1}^{n} \sqrt{2I(W; R_i|Z_i^{\pm})}. && (2.35) \end{aligned}$$

Proof of Corollary 1. When $\ell \in [a, b]$ and $\tilde{G}_i \in [a-b, b-a]$, it is straightforward to verify that $\tilde{G}_i$ is $(b-a)^2$-sub-Gaussian. The definition of the sub-Gaussian distribution gives $\Psi_{\tilde{G}_i|Z_i^{\pm}}(\lambda) \leq \frac{(b-a)^2}{2}\lambda^2$, and thus $\Psi^{*-1}_{\tilde{F}_i|Z_i^{\pm}}(\eta) \leq (b-a)\sqrt{2\eta}$, from which the corollary follows. □

2.2.4 Dichotomy and generalizations of existing bounds

The CD lemma allows us to view the existing MI, IMI, CMI, and CIMI bounds in a unified framework. By applying the CD lemma in different manners, these bounds can be obtained almost directly. The technical conditions under which the bound hold can also be generalized, and the bounds themselves can be strengthened using the inverse Fenchel conjugate. These results are summarized in Table 2.1. We also provide the bounds for the bounded loss function, which eliminate the $\bar{\psi}^{*-1}$ functions and have much simpler forms.

Take the derivation of MI bound [3] as an example: the mutual information $I(W; Z_{[n]})$ measures the correlation between W and $Z_{[n]}$. We use the CD lemma to decouple such correlation by letting $X = W$ and $Y = Z_{[n]}$, but the conditioning term does not exist. Let function $F = -\frac{1}{n}\sum_{i=1}^{n} \ell(W, Z_i) + L_\xi(W)$ be the generalization error that we aim to study. Its decoupled version $\tilde{F} = -\frac{1}{n}\sum_{i=1}^{n} \ell(\tilde{W}, \tilde{Z}_i) + L_\xi(\tilde{W})$ has zero mean and the conjugate CGF $\bar{\psi}^{*-1}_{\tilde{F}}(\eta)$. The MI

Table 2.1: A dichotomy of several generalization bounds using the CD Lemma

Approach	X	Y	U	F	Generalization bound	Special case $\ell \in [0,1]$
MI [3]	W	$Z_{[n]}$		$-\frac{1}{n}\sum_{i=1}^n \ell(W,Z_i) + L_\xi(W)$	$\bar{\psi}_{\tilde{F}}^{*-1}\left(I\left(W;Z_{[n]}\right)\right)$	$\sqrt{\frac{1}{2n}I(W;Z_{[n]})}$
IMI [8]	W	Z_i		$F_i = -\ell(W,Z_i)$	$\frac{1}{n}\sum_{i=1}^n \bar{\psi}_{\tilde{F}_i}^{*-1}\left(I\left(W;Z_i\right)\right)$	$\frac{1}{n}\sum_{i=1}^n \sqrt{\frac{1}{2}I(W;Z_i)}$
CMI [9]	W	$R_{[n]}$	$Z_{[n]}^{\pm}$	$\frac{1}{n}\sum_{i=1}^n R_i\left(\ell(W,Z_i^-) - \ell(W,Z_i^+)\right)$	$\bar{\psi}_{\tilde{F}\|Z_{[n]}^{\pm}}^{*-1}\left(I\left(W;R_{[n]}\|Z_{[n]}^{\pm}\right)\right)$	$\sqrt{2I(W;R_{[n]}\|Z_{[n]}^{\pm})}$
CIMI [10]	W	R_i	$Z_{[n]}^{\pm}$	$F_i = R_i\left(\ell(W,Z_i^-) - \ell(W,Z_i^+)\right)$	$\frac{1}{n}\sum_{i=1}^n \bar{\psi}_{\tilde{F}_i\|Z_{[n]}^{\pm}}^{*-1}\left(I\left(W;R_i\|Z_{[n]}^{\pm}\right)\right)$	$\frac{1}{n}\sum_{i=1}^n \sqrt{2I(W;R_i\|Z_{[n]}^{\pm})}$
ICIMI (this and [24])	W	R_i	$Z_i^{\pm}$	$F_i = R_i\left(\ell(W,Z_i^-) - \ell(W,Z_i^+)\right)$	$\frac{1}{n}\sum_{i=1}^n \bar{\psi}_{\tilde{F}_i\|Z_i^{\pm}}^{*-1}\left(I\left(W;R_i\|Z_i^{\pm}\right)\right)$	$\frac{1}{n}\sum_{i=1}^n \sqrt{2I(W;R_i\|Z_i^{\pm})}$

bound is then obtained by applying the CD lemma on these assignments, i.e,

$$\begin{aligned}\operatorname{gen}(\xi, P_{W|Z_{[n]}}) &= \mathbb{E}\left[L_\xi(W) - L_{Z_{[n]}}(W)\right] = \mathbb{E}[F] \\ &= \mathbb{E}[F] - \mathbb{E}[\tilde{F}] \leq \bar{\psi}_{\tilde{F}|U}^{*-1}\left(I(X;Y|U)\right) = \bar{\psi}_{\tilde{F}}^{*-1}\left(I\left(W;Z_{[n]}\right)\right).\end{aligned}$$

The CD lemma separates the loss geometry captured by $\bar{\psi}^{*-1}$ from the information acquired by the algorithm which is represented as a mutual information term, and it allows us to study them individually.

The CMI and CIMI results can be further strengthened by utilizing the inverse Fenchel conjugate function together with the sample-conditioned mutual information. More precisely, let $(\tilde{R}_{[n]}, \tilde{W})$ be the decoupled pair of $(R_{[n]}, W)$ conditioned on $Z_{[n]}^{\pm}$. Further define

$$\tilde{E}_i = \tilde{R}_i\left(\ell(\tilde{W}, Z_i^-) - \ell(\tilde{W}, Z_i^+)\right), \quad \tilde{E} = \frac{1}{n}\sum_{i=1}^n \tilde{E}_i, \tag{2.36}$$

then we have the strengthened CMI and CIMI bounds:

$$\operatorname{gen}\left(\xi, P_{W|Z_{[n]}}\right) \leq \mathbb{E}\left[\Psi_{\tilde{E}|Z_{[n]}^{\pm}}^{*-1}\left(I_{Z_{[n]}^{\pm}}\left(W;R_{[n]}\right)\right)\right], \tag{2.37}$$

$$\operatorname{gen}\left(\xi, P_{W|Z_{[n]}}\right) \leq \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\Psi_{\tilde{E}_i|Z_{[n]}^{\pm}}^{*-1}\left(I_{Z_{[n]}^{\pm}}(W;R_i)\right)\right]. \tag{2.38}$$

2.2.5 Comparison of the bounds

We first consider the special case where the loss function is bounded, i.e., $\ell \in [0,1]$. For this case, it was shown in [10] that the CIMI bound (2.16) is tighter than the CMI bound (2.14). We next show that the proposed bound (2.35) is tighter than the CIMI bound (2.16) when $\ell \in [0,1]$.

Lemma 2. *For any $i = 1, \ldots, n$, we have*

$$I(W; R_i | Z_i^{\pm}) \leq I(W; R_i | Z_{[n]}^{\pm}).$$

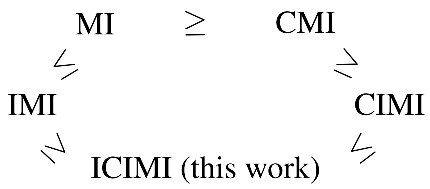


Figure 2.1: Relations among generalization bounds, when the inverse Fenchel conjugate functions are assumed to be the same.

To further understand the relation among these bounds under more general conditions when the loss function may not be bounded, let us assume the inverse Fenchel conjugate functions, which roughly capture the geometry induced by the expected loss, are the same (denoted as $\bar{\psi}^{*-1}$) for all the five approaches, i.e.,

$$\bar{\psi}^{*-1} = \bar{\psi}^{*-1}_{-\tilde{F}} = \bar{\psi}^{*-1}_{-\tilde{F}_i} = \bar{\psi}^{*-1}_{\tilde{F}|Z_{[n]}^{\pm}} = \bar{\psi}^{*-1}_{\tilde{F}_i|Z_{[n]}^{\pm}} = \bar{\psi}^{*-1}_{\tilde{F}_i|Z_i^{\pm}}.$$

Then we can focus on the information measure quantities, and compare these bounds as shown in Fig. 2.1. Here the inequalities given in black were proved previously (see [8] and [10]). Since the common function $\bar{\psi}^{*-1}$ is non-decreasing, the inequality "CIMI $\geq$ ICIMI" follows from Lemma 2. The inequality "IMI $\geq$ ICIMI" is implied by the following lemma for the same reason.

Lemma 3. *For any $i = 1, \ldots, n$, we have $I(W; R_i|Z_i^{\pm}) \leq I(W; Z_i)$.*

The inverse Fenchel conjugate functions may indeed be different for different bounds, thus although the above comparison suggests certain dominant relations, it is not clear for any specific problem, whether any given bound is tighter than the other. This is particularly true if we use the bounds based on the inverse Fenchel conjugate, however, even for the special case of $\ell \in [0, 1]$, the different multiplicative factors and the sum-square-root forms imply that the relation can be less clear.

2.2.6 Revisiting the example

We now return to the problem of estimating the Gaussian mean, and show that the proposed ICIMI bound can provide scaling behavior similar to that of IMI, thus order-wise stronger than the CMI and CIMI bounds. In fact, the bound is also strictly better than the IMI bound given in [8] asymptotically in this setting.

We first formally establish, as suspected previously, that the CMI and CIMI bounds are at least of constant order for this setting, the proof of which can be found in the appendix.

Proposition 1. *The strengthened CMI and CIMI bounds, i.e., (2.37) and (2.38), are at least $\frac{\sigma^2}{\pi\sqrt{\log e}}$ in the problem of estimating the Gaussian mean.*

The next proposition establishes a generalization error bound based on the ICIMI bound in this setting.

Proposition 2. *For the problem of estimating the mean of the Gaussian distribution, the ICIMI bound gives*

$$\text{gen}\left(\xi, P_{W|Z_{[n]}}\right) \leq \frac{2\sigma^2}{\sqrt{\pi}}\sqrt{\frac{1}{n-1}} + o\left(\frac{1}{\sqrt{n}}\right). \tag{2.39}$$

Remark: This bound scales as $\Theta(\sqrt{\frac{1}{n}})$. Compared to the IMI bound in (2.19), the ICIMI-based bound is asymptotically tighter by a factor of $\sqrt{\frac{\pi}{2}} \approx 1.25$.

Proposition 2 is proved by studying separately the sample-conditioned individual mutual information $I_{Z_i^{\pm}}(W;R_i)$ and the inverse Fenchel conjugate functions $\Psi^{*-1}_{\tilde{G}_i|Z^{\pm}}$. For the former, since the algorithm here is averaging the samples without any prior of the Gaussian distribution, without loss of generality, we can assume the mean of the Gaussian distribution to be 0, i.e., $\mu = 0$. Therefore, given $Z_i^{\pm} = z_{\pm} \in \mathbb{R}^2$, W is mixed-Gaussian distributed, which follows $N(\frac{z_+}{n}, \frac{n-1}{n^2}\sigma^2)$ when $R_i = 1$ and follows $N(\frac{z_-}{n}, \frac{n-1}{n^2}\sigma^2)$ when $R_i = -1$. The term $I_{Z_i^{\pm}}(W;R_i)$ is thus related to the scaling behavior of the differential entropy of a mixed Gaussian distribution, which the following lemma makes more precise.

Lemma 4. *Let R be a Rademacher random variable and V be a mixed-Gaussian random variable, such that $V \sim N(\nu, \sigma^2)$ when $R = 1$, and $V \sim N(-\nu, \sigma^2)$ when $R = -1$. We have*

$$I(V;R) = \frac{1}{2}\frac{\nu^2}{\sigma^2} + o\left(\frac{\nu^2}{\sigma^2}\right). \tag{2.40}$$

The next lemma gives an upper bound on the inverse Fenchel conjugate functions.

Lemma 5. *For the problem of estimating the mean of the Gaussian distribution, and any realization of $Z_i^{\pm} = z_{\pm} \in \mathbb{R}^2$ with $|z_+| \neq |z_-|$,*

$$\Psi^{*-1}_{\tilde{G}_i|Z_i^{\pm}=z_{\pm}}(\eta) \leq B_{z_{\pm},n}(\eta) = |z_+^2 - z_-^2|\sqrt{2\eta} + \Theta\left(\frac{1}{n}\right),$$

where

$$B_{z_{\pm},n}(\eta) := |z_+^2 - z_-^2|\sqrt{2\eta} + \frac{2\sigma^2(z_+ - z_-)^2}{n|z_+^2 - z_-^2|}\sqrt{2\eta} + \frac{4\max\left(z_+^2, z_-^2\right)}{n}; \tag{2.41}$$

and for $|z_+| = |z_-|$,

$$\Psi^{*-1}_{\tilde{G}_i|Z_i^{\pm}=z_{\pm}}(\eta) \leq 4\sigma\sqrt{\frac{2\eta}{n}}|z_+| + \frac{4\max\left(z_+^2, z_-^2\right)}{n}. \tag{2.42}$$

With these lemmas, Proposition 2 can be proved as follows.

2.3 Application of ICIMI Bound

2.3.1 A setting using empirical evaluations

We evaluate the proposed bound and compare it with the previous bounds in a scenario which does not have an explicit representation to facilitate the calculation of the distribution $P_{W|Z_{[n]}}$ or the corresponding mutual information term. The setup follows that in Section VI of [8], which is a logistic regression model for binary classification[1]. The loss function here is the 0-1 loss, which essentially measures the probability of prediction error. The logistic regressor itself is an empirical error minimization (ERM) algorithm that minimizes the empirical logistic loss, which is a differentiable convex surrogate of the 0-1 loss.

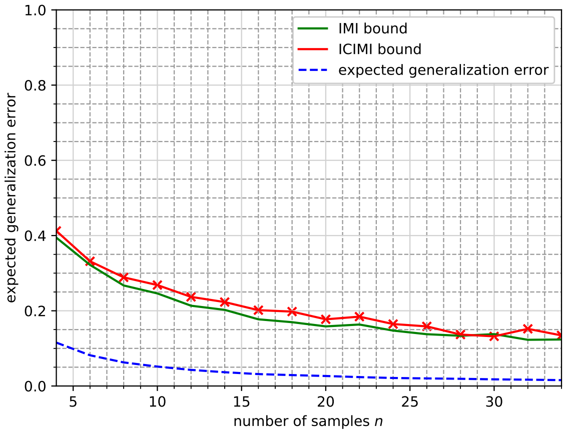


Figure 2.2: Empirical evaluation of the IMI and ICIMI bounds for the expected generalization error of logistic regression

The training data $Z_{[n]} = \{(X_i, Y_i)\}_{i\in[n]}$ are sampled in an i.i.d. fashion following some distribution ξ unknown to the algorithm, where X_i's are d-dimensional feature vectors and Y_i's are the corresponding labels with values ± 1. The data generating distribution ξ is set as follows: for any

[1] We would like to thank Dr. Bu for providing the source codes used in [8], which we adapted to perform the experiments.

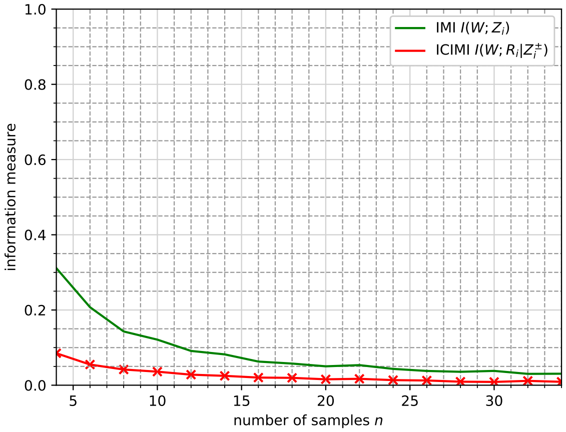


Figure 2.3: Empirical estimation of the information terms in the IMI and ICIMI bounds for logistic regression

$(X_i, Y_i) \sim \xi$, the label Y_i follows the Rademacher distribution; and the feature vector X_i, conditioned on Y_i, follows a Gaussian distribution with a mean vector $Y_i\mu$ and a covariance matrix $2I_d$, where μ is a d-dimensional non-zero vector and I_d is the $d \times d$ identity matrix.

Given the training data $Z_{[n]}$, the logistic regressor returns a hypothesis parameterized by $W \in \mathbb{R}^d$, which is the minimizer of the empirical logistic loss

$$L_{Z_{[n]}}^{logistic}(w) = \frac{1}{n}\sum_{i=1}^{n} \ln\left(1 + e^{-Y_i w^T X_i}\right).$$

The hypothesis $w \in \mathbb{R}^d$ predicts that the feature vector $x \in \mathbb{R}^d$ is associated with label $\operatorname{sign}(w^T x) \in \{\pm 1\}$. The information measure terms in the generalization error bounds, such as $I(W; Z_i)$ and $I(W; R_i|Z_i^{\pm})$, are empirically estimated by simulations. We shall refer to the process of generating n training samples and applying logistic regressor on the training data as one “simulation process". We can collect a copy of the training data $Z_{[n]}$, the output hypothesis W and auxiliary data $R_{[n]}$ and $Z_{[n]}^{\pm}$ after each simulation process.

In the experiment, we set the feature dimension $d = 2$, and $\mu = [1, 1]^T$. The IMI bound and the proposed ICIMI bound are evaluated as follows. The individual mutual information, and the individual conditional individual mutual information are estimated with the K-nearest neighbor-based mutual information estimator [25] with $K = 5$, based on the data collected by running 20,000 independent simulation processes. We do not include the CMI bound and the CIMI bound here, since they use conditioning on the whole data table $Z_n^{\pm}$, which involves many more random variables, and the nearest neighbor-based estimator cannot produce a sufficiently accurate estimate even within $1,000,000$ independent simulation processes.

The comparison between the IMI and the ICIMI generalization error bounds is shown in Figure 2.2. The ICIMI bound is comparable with the IMI bound. Recall that in the logistic regression problem, the IMI bound is $\frac{1}{n}\sum_{i=1}^{n}\sqrt{I(W;Z_i)/2}$, and the ICIMI bound is $\frac{1}{n}\sum_{i=1}^{n}\sqrt{2I(W;R_i|Z_i^{\pm})}$, and there exists a constant factor of 2 mismatch between the two bounds. In Figure 2.3, we isolate the mutual information terms in the two bounds, from which it can be seen that $I(W;R_i|Z_i^{\pm})$ is significantly less than $I(W;Z_i)$. The ICIMI bound is tighter than IMI bound when $4I(W;R_i|Z_i^{\pm}) < I(W;Z_i) = I(W;R_i,Z_i^{\pm}) = I(W;Z_i^{\pm}) + I(W;R_i|Z_i^{\pm})$, i.e., $3I(W;R_i|Z_i^{\pm}) < I(W;Z_i^{\pm})$; see [11, 24] for similar discussions.

From this example, we see that the ICIMI bound and the IMI bound have the advantage of being more amiable for estimation than the bounds that use conditioning on all samples since the latter group involves more random variables. The performance difference between the ICIMI bound and the IMI bound is however not significant in this example.

2.3.2 Application on a noisy and iterative algorithm

In this section we consider using the ICIMI bound to analyze the stochastic gradient Langevin dynamics (SGLD) algorithm discussed in [8]. The (non-stochastic) Langevin dynamic (LD) and more generally minibatch SGLD were considered in [6,10,24] using a more delicate data-dependent bounding approach, however, in this work we restrict our attention to the SGLD algorithm.

The SGLD algorithm model: We shall largely adopt the notation in [8], however with the additional data sample selection random variable R_i. Denote the parameter vector at iteration t as $W_{(t)} \in \mathbb{R}^d$. Let $W_{(0)}$ be an arbitrary initial vector for the algorithm. At each iteration $t \geq 1$, an index $V_{(t)} \in [n]$ in the data set is randomly selected, and if $V_{(t)} = i$ then the data sample $Z_i^{R_i}$ is used in the algorithm; the sample $Z_i^{R_i}$ is also denoted as $Z_{V_{(t)}}$. The gradient is computed at iteration t as $\nabla\ell(W_{(t-1)}, Z_{V_{(t)}})$. The parameter vector is updated as

$$W_{(t)} = W_{(t-1)} - \eta_{(t)}\nabla\ell(W_{(t-1)}, Z_{V_{(t)}}) + \sigma_{(t)}\epsilon_{(t)}, \tag{2.43}$$

where $\eta_{(t)}$ is the learning rate parameter, and $\epsilon_{(t)}$ is the independent zero-mean isotropic Gaussian noise with unit component variance. The parameter $\sigma_{(t)}$ controls the eventual variance of the additive Gaussian noise in (2.43).

The algorithm can take multiple iterations for training. If T iterations of training are performed, each sample is utilized T/n times in expectation. The eventual parameter obtained is denoted $W_{(T)}$. The trajectory of the parameter until iteration t is written as $W_{([t])} = (W_{(0)}, W_{(1)}, \ldots, W_{(t)})$; similarly $V_{([T])} = (V_1, V_2, \ldots, V_T)$.

Analysis of SGLD using the ICIMI bound: Let $\pi_{i,(t)}$ be a (possibly randomized) function that maps the random variables $(W_{([t-1])}, Z_i^{\pm}, V_{([T])})$ to the range $[0, 1]$, which can be viewed as an estimate of the probability of $R_i = +1$; with a slight abuse of notation, we also use $\pi_{i,(t)}$ to denote the induced random variable. Furthermore, define the following quantity

$$\Theta_{i,(t)}(W_{([t-1])}, Z_i^{\pm}, V_{([T])}) := \mathbb{E}\left[\left(\frac{R_i+1}{2} - \pi_{i,(t)}\right)^2 \middle| W_{([t-1])}, Z_i^{\pm}, V_{([T])}\right]. \tag{2.44}$$

Note that $\Theta_{i,(t)}$ is $(W_{([t-1])}, Z_i^{\pm}, V_{([T])})$-measurable, which reflects the accuracy of the estimate of $\pi_{i,(t)}$ to the true value of R_i given the condition $(W_{([t-1])}, Z_i^{\pm}, V_{([T])})$. For the purpose of bounding, we could simply use the optimal estimate $\pi^*_{i,(t)}$, however, $\pi_{i,(t)}$ can be any function, and by setting $\pi_{i,(t)} = 0.5$ (and noticing R_i only takes values ± 1), a trivial upper bound on $\Theta_{i,(t)}$ can be obtained as $\Theta_{i,(t)}(W_{([t-1])}, Z_i^{\pm}, V_{([T])}) \leq 1/4$. When there is no confusion, we shall write $\Theta_{i,(t)}(W_{([t-1])}, Z_i^{\pm}, V_{([T])})$ simply as $\Theta_{i,(t)}$.

We have the following result for the generalization error of the SGLD algorithm.

Proposition 3. *The generalization error of SGLD is upper-bounded as*

$$\mathrm{gen}(\xi, P_{W|Z_{[n]}}) \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\Psi^{*-1}_{\tilde{F}_i|Z_i^{\pm},V_{([T])}}\left(\sum_{\tau\in\mathcal{T}_i} S_{i,\tau}\right)\right], \tag{2.45}$$

where $S_{i,\tau} = \frac{\eta^2_{(\tau)}\mathbb{E}\left[\Theta_{i,(\tau)}\|\zeta_{(\tau)}(Z_i^{\pm})\|_2^2 \;\middle|Z_i^{\pm},V_{([T])}\right]}{2\sigma^2_{(\tau)}}$, *$\mathcal{T}_i$ is the set of iterations for which sample $Z_i^{R_i}$ is selected for the random sample path $V_{([T])}$, the function $\Psi^{*-1}_{\tilde{F}_i|Z_i^{\pm},V_{([T])}}$ with*

$$\tilde{F}_i = \tilde{R}_i\left(\ell(\tilde{W}_i, Z_i^-) - \ell(\tilde{W}_i, Z_i^+)\right), \tag{2.46}$$

is defined for each $V_{([T])} = v_{([T])}$ and $Z_i^{\pm} = z_i^{\pm}$, and

$$\zeta_{(\tau)}(Z_i^{\pm}) = \nabla\ell(W_{(\tau-1)}, Z_i^+) - \nabla\ell(W_{(\tau-1)}, Z_i^-) \tag{2.47}$$

is the incoherence at iteration τ for sample pair $Z_i^{\pm}$.

The bound is more general than those given in [6, 8, 10, 24] in the sense that very few assumptions are taken, such as the Lipschitz property or boundedness of the loss function in those results. This generality is accomplished through the usage of the Ψ^{*-1} function, however, it also makes the result less explicit. We shall discuss several specifications after the proof to make the bound more explicit.

We could similarly utilize the bound based on $\bar{\psi}^{*-1}$ to obtain an alternative looser bound, and we omit this derivation for brevity. Let us now specialize our result:

- If the gradients of the loss function are bounded, i.e., $\sup_{w\in\mathcal{W},z\in\mathcal{Z}}||\nabla\ell(w,z)||_2 \leq L$ for some $L > 0$, then $||\zeta_{(\tau)}(Z_i^{\pm})||_2^2 \leq 4L^2$, and thus

$$\mathrm{gen}(\xi, P_{W|Z_{[n]}}) \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\Psi^{*-1}_{\tilde{F}_i|Z_i^{\pm},V_{([T])}}\left(\sum_{\tau\in\mathcal{T}_i} S'_{i,\tau}\right)\right],$$

  where $S'_i = \frac{2\eta^2_{(\tau)}L^2}{\sigma^2_{(\tau)}}\mathbb{E}\left[\Theta_{i,(\tau)}\middle|V_{([T])}, Z_i^{\pm}\right]$.

- In addition, if the loss function is not bounded, yet conditioned on any $(Z_i^{\pm}, V_{[T]}) = (z_i^{\pm}, v_{[T]})$, $\tilde{F}_i$ is σ^2-sub-Gaussian, then we have

$$\mathrm{gen}(\xi, P_{W|Z_{[n]}}) \leq \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}\sqrt{\sum_{\tau\in\mathcal{T}_i}\frac{\sigma^2\eta_{(\tau)}^2 L^2}{\sigma_{(\tau)}^2}\mathbb{E}\left[\Theta_{i,(\tau)}\middle|V_{([T])}, Z_i^{\pm}\right]}.$$

  Note that this bound cannot be obtained using the bounds in [10,24] since their loss function must be bounded.

- As a special case, when the loss function is also bounded in $[a, b]$, then by applying Corollary 1, we have

$$\mathrm{gen}(\xi, P_{W|Z_{[n]}}) \leq \frac{2(b-a)L}{n}\sum_{i=1}^{n}\mathbb{E}\sqrt{\sum_{\tau\in\mathcal{T}_i}\frac{\eta_{(\tau)}^2}{\sigma_{(\tau)}^2}\mathbb{E}\left[\Theta_{i,(\tau)}\middle|V_{([T])}, Z_i^{\pm}\right]}. \qquad (2.48)$$

  This bound is in a similar form as those given in [10, 24] for the same setting, but can be looser due to the Jensen gap, because more expectation is taken inside the square root instead of outside. The bound in (2.48) inherently leverages certain data-dependent information: the variance term $\Theta_{i,(t)}$ will diminish, when the number of epochs is large and the estimate of R_i given the previous iterations becomes more and more accurate.

- Since the term $\mathbb{E}[\Theta_{i,(t)}|V_{([T])}, Z_i^{\pm}]$ is bounded by $\frac{1}{4}$, we can obtain the following relaxed bound for bounded and Lipschitz loss functions

$$\mathrm{gen}(\xi, P_{W|Z_{[n]}}) \leq \frac{(b-a)L}{n}\mathbb{E}_{V_{([T])}}\sum_{i=1}^{n}\sqrt{\sum_{\tau\in\mathcal{T}_i}\frac{\eta_{(\tau)}^2}{\sigma_{(\tau)}^2}},$$

  which degrades to the same form as those given in [6, 8] for the same setting (with the same or slightly worse constant factors).

From the discussion above, we see that the tightening effect of the ICIMI bound does not manifest in this SGLD algorithm setting. We suspect it is due to the specific difficulty in bounding the conditional mutual information in this context, particularly under the given assumptions. It is possible that by taking the data-dependent approach and identifying more specific assumptions as in [6, 10], the ICIMI can further tighten the result.

2.4 The Stochastic Chaining and Strengthened Bounds

2.4.1 Motivation

[4] introduced the chaining technique, which has traditionally been used in bounding random processes, into the derivation of information-theoretic generalization bounds. The technique resolves the issue that certain unbounded mutual information quantity leads to a vacuous bound, and may also yield a tighter bound in general. The main idea behind the result in [4] can be summarized as follows. The generalization error can be viewed as a random process $\{X_t\}_{t\in\mathcal{W}}$ indexed by the hypothesis parameters. If $(\mathcal{W}, d)$ is a bounded metric space under the metric d, then $\mathcal{W}$ can be divided into finer and finer partitions, with each coarse partition embedded into the next layer finer partition, and the partition cells having a decreasing radius. The generalization error can then be represented by a sum of chained quantities, each relating to two adjacent partition layers. Since the partitions are becoming finer and finer, each of these decomposed quantities can be bounded more effectively, eventually resulting in an overall tighter bound. This approach is referred to as *chaining mutual information*.

Despite the success of the chaining mutual information approach, we observe several difficulties in applying the chaining technique in this manner, which motivated the current work:

- **Restriction on the metric space to be bounded:** This chaining approach assumes a bounded metric space $(\mathcal{W}, d)$. However, even in some of the simplest settings, the parameter space may not be bounded (or impractical to assume the bound on $(\mathcal{W}, d)$ is known).

- **Difficulty in computation:** Using these deterministic and hierarchical partitions, the information measures involved in the bounds can be difficult to compute or bound analytically.

- **Restrictions in the partitions:** The hierarchical partitions place certain unnecessary geometric constraints on the covering radius sequence of the required partitions, which can

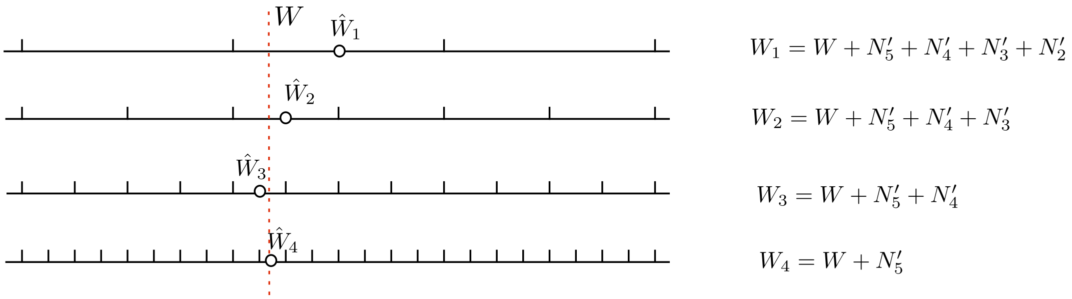


Figure 2.4: Multilevel quantization of a random value W using quantizers of different stepsize and the corresponding information-theoretic successive refinement source coding model.

impact the bound.

To make these difficulties more concrete, consider the following two simple examples.

- **Example-1**: The training samples are drawn $i.i.d.$ following a normal distribution with an unknown mean μ, and the algorithm wishes to estimate this mean. Here the parameter space is $\mathcal{W} = \mathbb{R}$, which is unbounded under any meaningful metric, particularly so for the natural Euclidean distance. Moreover, since the induced measure on $\mathcal{W}$ will not be uniform, computing the series sum of mutual information is rather difficult if not impossible.

- **Example-2**: Let $Z := (G_1, G_2) \sim \mathcal{N}(0, I_2)$ be standard normal vectors in $\mathbb{R}^2$. The learner needs to identify the phase of the vector through certain means, and the learned result is modeled as the true phase with certain additive noise. Here $\mathcal{W}$ is the bounded interval of the angle $[0, 2\pi)$. A natural sequence of partitions is to reduce the stepsize by an integer factor γ. However, this would preclude any non-integer γ values, which potentially makes the bound looser.

2.4.2 The stochastic chaining methods

The sequence of refining partitions of the metric space associated with the chaining technique is reminiscent of multilevel quantization in data compression. For example, a scalar source W distributed on the real line can be quantized with a stepsize of 2^{-k} for the k-th level quantization, resulting in its quantized representation $\hat{W}_k$. As the index k increases, the stepsize reduces and the accuracy of the quantization improves; see the left side of Fig. 2.4 for an illustration.

The information-theoretic model for multilevel quantization is usually referred to as successive refinement source coding [26, 27]. Particularly useful to us is a stochastic abstraction in this framework. For example, assume there are a total of K-levels, then one possible stochastic representation of the reconstruction $\hat{W}_k$ is W_k that is written as

$$W_k = \alpha_k \left(W + \sum_{i=k+1}^{K+1} N'_i \right), \tag{2.49}$$

where N'_i's are mutually independent random noises, also independent of W, and α_k's are certain fixed scalar coefficients; see the right side of Fig. 2.4. It is seen that the relation among W and $\{W_k\}_{k=1}^{K}$ is captured by the joint probability distribution among them, and we can measure the "distance" between W and W_k using $\mathbb{E}d(W, W_k)$, in contrast to the conventional chaining approach which uses the covering radius.

The main idea of this work is that these abstracted stochastic versions of $\{W_k\}_{i=1}^{K}$ can be used to replace the partition-based quantized versions in bounding the generalization error. This new approach helps to resolve the difficulties mentioned above: firstly the restriction for the metric space to be bounded is naturally removed, and secondly, it helps to simplify the computation, and lastly, the abstract model can remove the geometric constraints in designing the hierarchical partitions in some cases.

The proposed stochastic chaining approach essentially allows more flexible constructions of the chains than the more traditional deterministic chaining. One can attempt to further optimize the construction of stochastic chains based on the existing knowledge regarding the underlying metric space and the corresponding probability distribution for the given problem setting. On the other hand, when such knowledge is not available, we can safely fall back to the default construction of the original deterministic chaining partitions, which is essentially a special case of stochastic chaining.

We obtain two generalization bounds using stochastic chaining instead of the deterministic chaining in [4], built on the mutual information bound given in [3] and the individual sample mutual information bound given in [8], respectively. We further show that the proposed bound can reduce to the VC-dimension bound correctly. We then illustrate the benefits of this new approach in the context of the two examples. For the problem of estimating the Gaussian mean mentioned above, we can obtain a bound that is order-wise stronger than previously given in the literature. For the phase retrieval problem considered in [4], the bound can be naturally improved by optimizing over a continuous parameter.

We define a new notion of the stochastic chain as follows.

Definition 1 (Stochastic chain of random process and random variable pair)**.** *Let $(X_{\mathcal{W}}, W)$ be a random process and random variable pair, where W is a random variable in the set $\mathcal{W}$. A sequence of random variables $\{W_k\}_{k=k_0}^{\infty}$, each distributed in the set $\mathcal{W}$, is called a stochastic chain of the pair $(X_{\mathcal{W}}, W)$, if 1) $\lim_{k\to\infty} \mathbb{E}[X_{W_k}] = \mathbb{E}[X_W]$, 2) $\mathbb{E}[X_{W_{k_0}}] = 0$, and 3) $\{X_t\}_{t\in\mathcal{W}} \leftrightarrow W \leftrightarrow W_k \leftrightarrow W_{k-1}$ is a Markov chain for every $k > k_0$.*

We allow k_0 to take the value of $-\infty$ instead of providing another parallel definition to that effect. We are now ready to present the first main theorem of this work.

Theorem 6. *Assume $\{\mathrm{gen}_{Z_{[n]}}(\xi, w)\}_{w\in\mathcal{W}}$ is sub-Gaussian on $(\mathcal{W}, d)$, and $\{W_k\}_{k=k_0}^{\infty}$ is a stochastic chain of $(\{\mathrm{gen}_{Z_{[n]}}(\xi, w)\}_{w\in\mathcal{W}}, W)$. Then*

$$\mathrm{gen}(\xi, P_{W|Z_{[n]}}) \leq \sum_{k=k_0+1}^{\infty} \mathbb{E}\left[d(W_k, W_{k-1})\sqrt{2D(P_{Z_{[n]}|W_k}||P_{Z_{[n]}})}\right]. \tag{2.50}$$

Moreover, we have

$$\mathrm{gen}(\xi, P_{W|Z_{[n]}}) \leq \sum_{k=k_0+1}^{\infty} \sqrt{\mathbb{E}[d^2(W_k, W_{k-1})]}\sqrt{2I(Z_{[n]}; W_k)}. \tag{2.51}$$

The following theorem is based on the individual sample mutual information bound of [8].

Theorem 7. *For each $i \in [n]$, assume $\{\mathrm{gen}^i_{Z_i}(w)\}_{w\in\mathcal{W}}$ is sub-Gaussian on $(\mathcal{W}, d)$, and $\{W_{i,k}\}_{k=k_0}^{\infty}$ is a stochastic chain of $(\{\mathrm{gen}^i_{Z_i}(w)\}_{w\in\mathcal{W}}, W)$. Then*

$$\mathrm{gen}(\xi, P_{W|Z_{[n]}}) \leq \frac{1}{n}\sum_{i=1}^{n}\sum_{k=k_0+1}^{\infty}\mathbb{E}\left[d(W_{i,k}, W_{i,k-1})\sqrt{2D(P_{Z_i|W_{i,k}}||P_{Z_i})}\right]. \tag{2.52}$$

Moreover, we have

$$\mathrm{gen}(\xi, P_{W|Z_{[n]}}) \leq \frac{1}{n}\sum_{i=1}^{n}\sum_{k=k_0+1}^{\infty}\sqrt{\mathbb{E}[d^2(W_{i,k}, W_{i,k-1})]}\sqrt{2I(Z_i; W_{i,k})}. \tag{2.53}$$

These two theorems are given in the context of bounding generalization errors, which are obtained using a more general result on bounding random processes.

Theorem 8. *Assume $X_{\mathcal{W}}$ is sub-Gaussian on $(\mathcal{W}, d)$, and $\{W_k\}_{k=k_0}^{\infty}$ is a stochastic chain for $(X_{\mathcal{W}}, W)$, then*

$$\mathbb{E}\left[X_W\right] \leq \sum_{k=k_0+1}^{\infty}\mathbb{E}\left[d(W_k, W_{k-1})\sqrt{2D(P_{X_{\mathcal{W}}|W_k}||P_{X_{\mathcal{W}}})}\right]. \tag{2.54}$$

Moreover, we have

$$\mathbb{E}\left[X_W\right] \leq \sum_{k=k_0+1}^{\infty}\sqrt{\mathbb{E}[d^2(W_k, W_{k-1})]}\sqrt{2I(X_{\mathcal{W}}; W_k)}. \tag{2.55}$$

By using a deterministic sequence of partitions to form $\{W_k\}_{k_0}^{\infty}$, we recover the result in [4] which was obtained for bounded metric space $(\mathcal{W}, d)$.

Corollary 2. *Let $\{\mathcal{P}_k\}_{k=k_0}^{\infty}$ be an increasing sequence of partitions of $\mathcal{W}$, where for each $k \geq k_0$, $\mathcal{P}_k$ is a 2^{-k}-partition of the bounded metric space $(\mathcal{W}, d)$, and $2^{-k_0} \geq$ diam$(\mathcal{W}) = \max_{x,y\in\mathcal{W}} d(x, y)$. Let W_k be the center of the covering ball of the partition cell that W belongs to in the partition $\mathcal{P}_k$, then for separable process $X_{\mathcal{W}}$ on $(\mathcal{W}, d)$,*

$$\mathbb{E}\left[X_W\right] \leq \sum_{k=k_0+1}^{\infty}\mathbb{E}\left[3\cdot 2^{-k}\sqrt{2D(P_{X_{\mathcal{W}}|W_k}||P_{X_{\mathcal{W}}})}\right]$$

$$\leq \sum_{k=k_0+1}^{\infty} 3 \cdot 2^{-k} \sqrt{2I(X_{\mathcal{W}}; W_k)}. \tag{2.56}$$

Unlike existing deterministic chaining, expected distance instead of worst-case distance is used.

Proof of Theorem 3. To prove the theorem, we start by writing

$$X_W = X_{W_{k_0}} + \sum_{k=k_0+1}^{k_1} (X_{W_k} - X_{W_{k-1}}) + (X_W - X_{W_{k_1}}). \tag{2.57}$$

Because $\{W_k\}_{k=k_0}^{\infty}$ is a stochastic chain for $(\tilde{X}_{\mathcal{W}}, W)$, we have $\mathbb{E}[X_{W_{k_0}}] = 0$ and $\lim_{k_1 \to \infty} \mathbb{E}[X_{W_{k_1}}] = \mathbb{E}[X_W]$, and it follows that

$$\begin{aligned} \mathbb{E}\left[X_W\right] &= \sum_{k=k_0+1}^{\infty} \mathbb{E}\left[X_{W_k} - X_{W_{k-1}}\right] \\ &= \sum_{k=k_0+1}^{\infty} \mathbb{E}\left[\mathbb{E}[X_{W_k} - X_{W_{k-1}} | W_k, W_{k-1}]\right]. \end{aligned} \tag{2.58}$$

By the Donsker–Varadhan variational representation of the KL divergence, the expectation of a function $g(Y)$ with respect to the measure P defined on $\mathcal{Y}$ can be bounded as

$$\mathbb{E}_P[g(Y)] \leq \inf_{\lambda > 0} \frac{1}{\lambda} \left(D(P||Q) + \log \mathbb{E}_Q[e^{\lambda g(Y)}] \right), \tag{2.59}$$

where Q is another measure on $\mathcal{Y}$.

In our setting, let $Y = g(Y) = \Delta X_{w_k, w_{k-1}} = X_{w_k} - X_{w_{k-1}}$, $Q = P_{\Delta X_{w_k, w_{k-1}}}$, and $P = P_{\Delta X_{w_k, w_{k-1}} | w_k, w_{k-1}} := P_{\Delta X_{w_k, w_{k-1}} | W_k = w_k, W_{k-1} = w_{k-1}}$, then we have

$$\begin{aligned} &\mathbb{E}_{P_{\Delta X_{W_k, W_{k-1}} | w_k, w_{k-1}}}[\Delta X_{w_k, w_{k-1}}] \\ &\leq \inf_{\lambda > 0} \frac{1}{\lambda} \Big(D(P_{\Delta X_{w_k, w_{k-1}} | w_k, w_{k-1}} || P_{\Delta X_{w_k, w_{k-1}}}) + \log \mathbb{E}_{P_{\Delta X_{w_k, w_{k-1}}}} \left[e^{\lambda (X_{w_k} - X_{w_{k-1}})} \right] \Big) \\ &\leq \inf_{\lambda > 0} \frac{1}{\lambda} \Big(D(P_{\Delta X_{w_k, w_{k-1}} | w_k, w_{k-1}} || P_{\Delta X_{w_k, w_{k-1}}}) + \frac{1}{2} d^2(w_k, w_{k-1}) \lambda^2 \Big) \\ &= d(w_k, w_{k-1}) \sqrt{2D(P_{\Delta X_{w_k, w_{k-1}} | w_k, w_{k-1}} || P_{\Delta X_{w_k, w_{k-1}}})}, \end{aligned} \tag{2.60}$$

where the second inequality is because the process $X_{\mathcal{W}}$ is sub-Gaussian on $(\mathcal{W}, d)$.

Denote $\tilde{X}_{\mathcal{W}}$ as an independent copy of $X_{\mathcal{W}}$ such that $\tilde{X}_{\mathcal{W}}$ and $X_{\mathcal{W}}$ are independent and have the same distribution. Denote $\Delta_k = X_{W_k} - X_{k-1}$ and $\tilde{\Delta}_k = \tilde{X}_{W_k} - \tilde{X}_{k-1}$. It then follows that $P_{\Delta_k|W_k=w_k,W_{k-1}=w_k} = P_{\Delta X_{w_k,w_{k-1}}|w_k,w_{k-1}}$ and $P_{\tilde{\Delta}_k|W_k=w_k,W_{k-1}=w_k} = P_{\Delta X_{w_k,w_{k-1}}}$. The fact that $\{W_k\}_{k=k_0}^{\infty}$ is a stochastic chain also implies that $\lim_{k\to\infty}\mathbb{E}[X_{W_k}] = \mathbb{E}[X_W]$, and thus

$$\mathbb{E}\left[X_W\right] \leq \sum_{k=k_0+1}^{\infty} \mathbb{E}\left[d(W_k, W_{k-1})\sqrt{2D(P_{\Delta_k|W_k,W_{k-1}}||P_{\tilde{\Delta}_k|W_k,W_{k-1}})}\right]. \tag{2.61}$$

By the data processing inequality for the KL divergence, we have

$$D(P_{\Delta_k|W_k,W_{k-1}}||P_{\tilde{\Delta}_k|W_k,W_{k-1}}) \leq D(P_{X_{\mathcal{W}}|W_k,W_{k-1}}||P_{\tilde{X}_{\mathcal{W}}|W_k,W_{k-1}}). \tag{2.62}$$

Since $P_{X_{\mathcal{W}}|W_k,W_{k-1}} = P_{X_{\mathcal{W}}|W_k}$ and $P_{\tilde{X}_{\mathcal{W}}|W_k,W_{k-1}} = P_{\tilde{X}_{\mathcal{W}}} = P_{X_{\mathcal{W}}}$, from which the second inequality follows.

The mutual information-based bound can be derived by

$$\begin{aligned}
&\mathbb{E}\left[d(W_k, W_{k-1})\sqrt{2D(P_{X_{\mathcal{W}}|W_k}||P_{X_{\mathcal{W}}})}\right] \\
&\leq \sqrt{\mathbb{E}[d^2(W_k, W_{k-1})]}\sqrt{\mathbb{E}[2D(P_{X_{\mathcal{W}}|W_k}||P_{X_{\mathcal{W}}})]} \\
&= \sqrt{\mathbb{E}[d^2(W_k, W_{k-1})]}\sqrt{2I(X_{\mathcal{W}}; W_k)},
\end{aligned} \tag{2.63}$$

where the inequality is Cauchy-Schwartz inequality for random variables and the equality is by the definition of mutual information. □

Note that we can also remove the Markov chain assumption in stochastic chaining (Definition 1). The similar expected generalization upper bounds as in Theorem 3 can be derived by replacing $D(P_{X_{\mathcal{W}}|W_k}||P_{X_{\mathcal{W}}})$ and $I(X_{\mathcal{W}}; W_k)$ by $D(P_{X_{\mathcal{W}}|W_k,W_{k-1}}||P_{X_{\mathcal{W}}})$ and $I(X_{\mathcal{W}}; W_k, W_{k-1})$, respectively.

To obtain Theorem 6 from Theorem 8, we let $\mathcal{W} := \mathcal{W}$, and $X_w := \mathrm{gen}(w)$ for $w \in \mathcal{W}$. Due to the Markov chain

$$X_{\mathcal{W}} = \{\mathrm{gen}(w)\}_{w\in\mathcal{W}} \leftrightarrow Z_{[n]} \leftrightarrow W \leftrightarrow W_k, \tag{2.64}$$

for all $k \geq k_1$, we can apply the data processing inequality for KL divergence [28] and that for mutual information, respectively, to arrive at

$$\begin{aligned} D(P_{X_{\mathcal{W}}|W_k}||P_{X_{\mathcal{W}}}) &\leq D(P_{Z_{[n]}|W_k}||P_{Z_{[n]}}),\\ I(X_{\mathcal{W}};W_k) &\leq I(Z_{[n]};W_k), \end{aligned} \tag{2.65}$$

from which Theorem 6 follows immediately. Theorem 7 can be obtained similarly.

When the process is not sub-Gaussian, more general forms of these bounds can also be found in terms of the cumulant generating function. This result is given in the appendix.

2.4.3 Relations to existing results

Connection to VC theory: For binary classification problems, i.e., $|\mathcal{Y}| = 2$ with zero-one loss $\ell(w,(x,y)) = \mathbb{I}(h_w(x) \neq y)$, the generalization error of any classifier W is upper bounded as $\mathrm{gen}(\xi, P_{W|Z_{[n]}}) \leq O(\sqrt{\frac{d_{VC}(\mathcal{W})}{n}})$, where $d_{VC}(\mathcal{W})$ is the VC-dimension of the classification function class $\mathcal{H}_{\mathcal{W}}$ (c.f., [1] Ch. 6). The generalization error bound in Theorem 8, or more precisely the proposed stochastic chaining approach, can naturally recover the VC-dimension based bound, and we establish this connection in the appendix.

Discussion on the chaining construction: The conventional deterministic chaining places certain structural constraints on the hierarchical partitions. For example, consider a partition of a bounded 2-D space using congruent hexagon cells; the next partition at the higher level will be collections of such hexagons. This subsequently implies that hierarchy must follow a certain relation between consecutive levels, and the analysis of such hierarchical partitions can be complex. The stochastic chaining technique can remove the geometric constraints in the *design of hierarchical partitions* as in Corollary 2 in many cases. In the example above, we can replace the partition using either an additive Gaussian noise or additive noise with a uniform distribution on hexagons (see the second example in the next section where a similar uniform additive noise is used).

Since stochastic chains include conventional partition-based chaining as a special case, it is not more difficult to construct. The construction can be more straightforward due to its flexibility. For example, for bounded metric space, we can use the following generic construction: let $p(W_{k-1}|W_k)$ be uniformly distributed on a metric ball of radius 2^{-k} centering at W_k. If more information regarding the distribution of W is known, we can further optimize the chain, e.g., by adjusting the radius such that they are dependent on the density value of W_k; more specifically, we can let the radius be larger for W_k values of lower density, and vice versa. If the metric space is also a vector space, it can be convenient to let $p(W_{k-1}|W_k)$ be some vector Gaussian distribution with covariance scaling like 2^{-k}. This allows more opportunity for optimization for stronger bounds in a parametric form. In contrast, it is impossible to design partitions (or deterministic mappings [29]) to mimic such behaviors, let alone find an analytic bound. This issue has a natural origin in source coding: deterministic quantization design vs. probabilistic forward test channel modeling. The latter is used in source coding for mathematically precise characterization, and analytic optimization.

Comparison to the chaining technique in [29]: The alternative chaining method proposed by Hafez-Kolahi et al. (Theorem 6 in [29]) used a different chaining construction, which does not require hierarchical partitions, and to some extent, it helps resolve the difficulty in designing such hierarchical partitions. However, this simplification came with a heavy price: the learning algorithm must be *deterministic*, and the hypothesis space $\mathcal{W}$ still needs to be *bounded* (since the core steps rely on [4]), and there is a factor of 2 loss in the bound. The restrictions make it inapplicable in the two examples we study in the next section. In contrast, the proposed method applies to unbounded metric space and does not require the learning algorithm to be deterministic.

2.4.4 Illustrative examples

We analyze two simple settings, which demonstrate the effectiveness of the proposed stochastic chaining technique. The purpose of discussing the following two examples is by no means to literally characterize the generalization error, since the generalization error can be calculated directly due to the simplicity of the examples. We aim to show the effectiveness of the proposed stochastic chaining technique in these two examples by comparing it with the underlining generalization error and some previous generalization error bounds.

2.4.4.1 Estimating the Gaussian mean

Consider the case when the training samples $Z_{[n]}$ are drawn $i.i.d.$ following $N(\mu, \sigma^2)$ for some unknown μ. Here $\mathcal{W} = \mathbb{R}$, and a natural choice of the metric in this space is the (scaled) Euclidean distance. The loss function is $\ell(w, Z) = (w - Z)^2$, and by defining $\bar{Z}_n := \frac{1}{n}\sum_{i=1}^{n} Z_i$, the random process (indexed by w) of interest can be written as

$$\mathrm{gen}_{Z_{[n]}}(\xi, w) = \sigma^2 + \mu^2 - \frac{\sum_{i=1}^{n} Z_i^2}{n} + 2w(\bar{Z}_n - \mu). \tag{2.66}$$

It follows that

$$\mathrm{gen}_{Z_{[n]}}(\xi, w) - \mathrm{gen}_{Z_{[n]}}(\xi, v) = 2(w - v)\left(\bar{Z}_n - \mu\right), \tag{2.67}$$

which is $d^2(w, v)$ sub-Gaussian with $d^2(w, v) = \frac{4\sigma^2(w-v)^2}{n}$. The learner deterministically estimates μ by averaging the training samples, i.e., $W = \bar{Z}_n$. We shall use Theorem 6 to bound the generalization error in this case.

To build a stochastic chain, select a sequence of mutually independent Gaussian noise $\{N_i'\}_{i\in\mathbb{N}}$, which is independent of W, and $N_i' \sim \mathcal{N}(0, \sigma_i'^2)$, where $\sigma_i'^2 = \frac{\sigma^2}{2^i n}$. Define the cumulative noise

$$N_k := \sum_{i=k+1}^{\infty} N_i' \sim \mathcal{N}(0, \sigma_k^2), \tag{2.68}$$

where $\sigma_k^2 = \frac{\sigma^2}{2^k n}$. The stochastic chain is designed as

$$W_k - \mu = \alpha_k(W - \mu + N_k), \tag{2.69}$$

where $\alpha_k = \frac{\sigma^2/n}{\sigma^2/n+\sigma_k^2} = \frac{1}{1+2^{-k}}$. We then have

$$W_{k-1} - \mu = \frac{\alpha_{k-1}}{\alpha_k}(W_k - \mu) + \alpha_{k-1}N'_k, \tag{2.70}$$

where W_k and N'_k are independent. Under this stochastic chain, we can derive the expression for $\sqrt{\mathbb{E}[d(W_k, W_{k-1})^2]}$ and the mutual information term $I(Z_{[n]}; W_k)$. Specifically, $\mathbb{E}[d(W_k, W_{k-1})^2] \leq \frac{\sigma^4}{n^2}\frac{3}{2^{k-1}+1}$, which relies on the relations between W_k and W_{k-1} in (2.70) and the detailed calculation is given in the appendix. The mutual information can be upper bounded as

$$I(Z_{[n]}; W_k) \leq I(W; W_k) = \frac{1}{2}\ln(1 + 2^k), \tag{2.71}$$

where the inequality is due to the data processing inequality over the Markov chain $Z_{[n]} \leftrightarrow W \leftrightarrow W_k$ and the equality is by the Gaussian channel nature of the stochastic chain design. The detailed proof steps are given in the appendix. A bound of the following form can then be obtained

$$\mathbb{E}[X_W] \leq \frac{\sigma^2}{n} \sum_{k=-\infty}^{\infty} \sqrt{\frac{3\ln(1 + 2^k)}{2^{k-1} + 1}}. \tag{2.72}$$

Note that the series sum on the right-hand side of (2.72) converges, and thus the bound is of order $O(\sigma^2/n)$. Bounding the series sum using numerical methods, we can then obtain $\mathbb{E}[X_W] \leq \frac{13\sigma^2}{n}$.

Due to the simplicity of the setting, the generalization error can in fact be calculated exactly to be $\frac{2\sigma^2}{n}$. It can be seen that the generalization bound offered by Theorem 6 has the same $O(\sigma^2/n)$ order as the true generalization error. In contrast, [8] derived a generalization error bound of the order $O(\sigma^2/\sqrt{n})$ using the individual sample mutual information approach. Thus the proposed approach results in an order-wise improvement in this example case. More importantly, it can be seen that the proposed chaining approach allows us to overcome the limitation of bounded metric space (i.e., the chaining mutual information approach [4] does not even apply in this setting), and also simplify the calculation due to the introduced dependence structure in the chain. In the appendix, we further derive an improved bound (with a slightly better constant factor) using

Table 2.2: Comparison of $\mathbb{E}[X_W]$ bounds

ϵ	$1/20$	$1/30$	$1/40$	$1/50$	$1/100$	$1/200$	$1/400$
Chaining mutual information [4]	1.1013	0.7507	0.5709	0.4612	0.2364	0.1204	0.0610
stochastic chaining ($\gamma = 3.75$)	0.4951	0.3387	0.2581	0.2088	0.1074	0.0548	0.0278
$\mathbb{E}[X_W]$ true value	0.0626	0.0417	0.0313	0.0250	0.0125	0.0062	0.0031

Theorem 7.

2.4.4.2 Phase retrieval

In the phase retrieval example given in [4], the data $Z := (G_1, G_2) \sim \mathcal{N}(0, I_2)$ is a standard normal vector in $\mathbb{R}^2$. The hypothesis class is $\mathcal{W} = [0, 2\pi)$, and through the transformation $t = (\cos w, \sin w)$ for $w \in \mathcal{W}$, it is the same as $\mathcal{W} = \{t \in R^2 : ||t||_2 = 1\}$; we will use them interchangeably. Define the loss function $\ell(t, Z) = -\langle t, Z\rangle$, which implies that the learner wishes to estimate an angle for the underlying data, and the generalization error process is a Gaussian process $X_t := \langle t, Z\rangle$. The metric d is the Euclidean distance, and the process $X_{\mathcal{W}}$ is sub-Gaussian. Suppose the learned parameter is

$$W := \left(\arg\max_{\phi\in[0,2\pi)} X_\phi\right) \oplus \zeta \,(\mathrm{mod}\ 2\pi), \tag{2.73}$$

where ζ is independent of $X_{\mathcal{W}}$, and has an atom with a mass ϵ on 0, and $1-\epsilon$ that is uniformly distributed in $[0, 2\pi)$. Note that $\arg\max_{\phi\in[0,2\pi)} X_\phi$ is exactly the phase of (G_1, G_2), which will be the hypothesis learned by an ERM learner, and W being retrieved here is a noisy version of the phase.

The stochastic chain can be given as

$$W_k = (W \oplus N_k)(\mathrm{mod}\ 2\pi), \tag{2.74}$$

where $N_k = \sum_{i=k+1}^{\infty} N_i'$, and N_k' is uniformly distributed on $[-\gamma^{-k}\pi, \gamma^{-k}\pi)$ for some $\gamma > 1$ to be specified later; N_k''s are mutually independent and also independent of the hypothesis parameter

W.

Since $W \oplus N_{-1}$ is independent of Z and uniformly distributed on $[0, 2\pi)$, we have $\mathbb{E}[X_{W_{-1}}] = \mathbb{E}[\langle W + N_{-1}, Z\rangle] = 0$. It is also clear that $W_k \to W$ when $k \to \infty$ a.s., and thus $\mathbb{E}[X_W] = \lim_{k\to\infty} \mathbb{E}[X_{W_k}]$ since the process is Gaussian. Since $W_{k-1} - W_k$ is exactly N'_k, the Euclidean distance between W_k and W_{k-1} (using their vector representations) is bounded by the length of the arc, i.e., $d(W_k, W_{k-1}) \leq \gamma^{-k}\pi$. We can now apply Theorem 6, where

$$\begin{aligned} I(W_k; X_{\mathcal{W}}) &= h(N_k \oplus W) - h(N_k \oplus \zeta) \\ &= \log 2\pi - h(N_k \oplus \zeta). \end{aligned} \tag{2.75}$$

The second term can be bounded as

$$h(N_k \oplus \zeta) \geq h\left(N_k \oplus \zeta \middle| \sum_{k+2}^{\infty} N'_j\right) = h(N'_{k+1} \oplus \zeta), \tag{2.76}$$

using the fact that more conditioning reduces the differential entropy. Due to the structure of the distribution of N'_{k+1} and Z, the density of $N'_{k+1} \oplus \zeta$ can be written down explicitly as

$$f(N'_{k+1} + \zeta) = \begin{cases} (2\pi)^{-1}(1-\epsilon) \\ \qquad \left[-\pi, -\gamma^{-k-1}\pi\right) \cup \left[\gamma^{-k-1}\pi, \pi\right) \\ (2\pi)^{-1}(\gamma^{k+1}\epsilon + (1-\epsilon)) \\ \qquad \left[-\gamma^{-k-1}\pi, \gamma^{-k-1}\pi\right). \end{cases} \tag{2.77}$$

Thus we can bound $h(N_k \oplus \zeta)$ and subsequently $I(W_k; X_T)$ using this density function, which eventually gives

$$\begin{aligned} \mathbb{E}[X_W] \leq \sqrt{2}\pi \sum_{k=0}^{\infty} \gamma^{-k} \Big(&(1-\epsilon)(1 - \frac{1}{\gamma^{k+1}}) \log(1-\epsilon) \\ &+ \left[\epsilon + \frac{1-\epsilon}{\gamma^{k+1}}\right] \log\left[\gamma^{k+1}\epsilon + 1 - \epsilon\right] \Big)^{1/2}. \end{aligned} \tag{2.78}$$

When choosing $\gamma = 2$, this is almost identical to the result given in [4] using the partition-based chaining, except the slightly better coefficient $\sqrt{2}\pi$ instead of $6\sqrt{2}$. This improved coefficient is mainly due to the more explicit bound on $d(W_k, W_{k-1})$ inherent in the Euclidean space, instead of the same distance derived in a generic metric space.

One advantage of the proposed approach is that we can further optimize γ over $\mathbb{R}$. Observe that the series has a faster-decaying tail if γ is large, however, the first term, i.e., $k = 0$, approaches ∞ when $\gamma \to \infty$. Thus there is an optimal γ value in between for this bound. The numerical result suggests $\gamma^* \approx 3.75$, which provides a slight improvement compared to $\gamma = 2$. As noted in [4], in this toy setting, we can calculate the exact true value $\mathbb{E}[X_W] = \epsilon\frac{\sqrt{\pi}}{2}$. A comparison of several bounds is given in Table. 2.2. To obtain (2.78), we have relaxed this bound in (2.76) for convenience using a simple property of the entropy function, and therefore loosen the bound to some extent. Moreover, we have chosen to use the geometric sequence γ^k to produce the stochastic chain, and it is possible other sequences can produce tighter bounds.

The individual sample mutual information bound in [8] requires multiple samples. In this phase retrieval example, however, there is only one sample G^2, and this bound degrades to the mutual information-based bound in [3], which in this case is vacuous since $I(W; X_W)$ is infinite.

2.5 Exactly Tight Information-Theoretic Generalization Error Bound for the Quadratic Gaussian Problem

2.5.1 Variational representation and Quadratic Gaussian Problem

2.5.1.1 Variational Representation of the KL Divergence

The Donsker-Varadhan variational representation of KL divergence for a random scalar-valued random function $F = f(X)$ on a random variable X is given by

$$D(P||Q) = \sup_f \left\{\lambda \mathbb{E}_P[F] - \ln \mathbb{E}_Q[e^{\lambda F}]\right\}, \text{where equality achieved when } \lambda F^* = \ln\frac{\mathrm{d}P}{\mathrm{d}Q} + C,$$

or in the inequality form

$$\lambda \mathbb{E}_P[F] \leq D(P||Q) + \ln \mathbb{E}_Q[e^{\lambda F}], \quad \forall \lambda \in \mathbb{R}. \tag{2.79}$$

This inequality is sometimes also referred to as the change of measure inequality. Note that P and Q can be the distributions of the underlying random variable X, or more directly, the distributions of F. In the context of bounding generalization error, examples are $F = \ell(W, Z)$ or $F = L_\xi(W) - \ell(W, Z)$.

The centered cumulant generating function of a random variable F is

$$\Lambda_{F,Q}(\lambda) = \ln \mathbb{E}_Q \left[e^{\lambda F}\right] - \lambda \mathbb{E}_Q[F]. \tag{2.80}$$

Combining it with the inequality above gives

$$D(P||Q) + \Lambda_{F,Q}(\lambda) \geq \lambda \mathbb{E}_P[F] - \lambda \mathbb{E}_Q[F], \quad \lambda \in \mathbb{R}. \tag{2.81}$$

Now if we choose $F = f(W, Z)$, then for any $Z = z$ the conditional version of the above inequality is

$$D(P_{W|Z=z}||Q_{W|Z=z}) + \Lambda_{F|Z=z,Q_{W|Z=z}}(\lambda) \geq \lambda \mathbb{E}_P[F|Z = z] - \lambda \mathbb{E}_Q[F|Z = z], \quad \lambda \in \mathbb{R}, \tag{2.82}$$

where

$$\Lambda_{F|Z=z,Q_{W|Z=z}}(\lambda) = \ln \mathbb{E}_{Q_{W|Z=z}} \left[e^{\lambda F}|Z = z\right] - \lambda \mathbb{E}_{Q_{W|Z=z}}[F|Z = z]. \tag{2.83}$$

We will simply replace $Z = z$ in the condition by Z when the exact conditional value realization is not specified.

With a negative λ we therefore obtain

$$\mathbb{E}_Q[F] - \mathbb{E}_P[F] \leq \inf_{\lambda<0} \left\{ \frac{D(P||Q) + \Lambda_{F,Q}(\lambda)}{-\lambda} \right\}$$

$$= \inf_{\lambda > 0} \left\{ \frac{D(P||Q) + \Lambda_{-F,Q}(\lambda)}{\lambda} \right\}, \tag{2.84}$$

where equality is achieved if and only if

$$\ln \frac{\mathrm{d}P}{\mathrm{d}Q} \in \left\{ \lambda^{-1} F + b : \lambda \in \mathbb{R}_-,\ b \in \mathbb{R} \right\}. \tag{2.85}$$

When P is the joint distribution of underlying random variables, and Q is the product distribution of their marginals, then $D(P||Q)$ reduces to a mutual information term. Similarly, with a positive λ, we obtain

$$\mathbb{E}_P[F] - \mathbb{E}_Q[F] \leq \inf_{\lambda > 0} \left\{ \frac{D(P||Q) + \Lambda_{F,Q}(\lambda)}{\lambda} \right\}. \tag{2.86}$$

To be consistent with past results in the literature, we will sometimes use the following definition. The Legendre dual function on the interval $[0, b)$ for some $0 < b \leq \infty$ is

$$\Lambda^*(x) := \sup_{\lambda \in [0,b)} (\lambda x - \Lambda(\lambda)). \tag{2.87}$$

$\Lambda(\lambda)$ is convex and $\Lambda(0) = \Lambda'(0) = 0$. It can be shown that the inverse dual function is

$$\Lambda^{*-1}(y) = \inf_{\lambda \in [0,b)} \left(\frac{y + \Lambda(\lambda)}{\lambda} \right). \tag{2.88}$$

2.5.1.2 The Quadratic Gaussian Problem

In the canonical Gaussian-mean-estimation problem, data samples are $Z_1, Z_2, \ldots, Z_n \overset{i.i.d.}{\sim} \xi = N(\mu, \sigma^2)$ and the sample-average algorithm chooses the following hypothesis $W = \frac{1}{n} \sum_{i=1}^n Z_i$. Then the expected generalization error is

$$\mathrm{gen}(\xi, P_{W|Z_{[n]}}) = \mathbb{E} \left[(\tilde{Z} - W)^2 - \frac{1}{n} \sum_{i=1}^n (Z_i - W)^2 \right]$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left(\sigma^2 + \mu^2 - Z_i^2 + 2(Z_i - \mu)W\right), \tag{2.89}$$

where $\tilde{Z}_{[n]}$ are n i.i.d. testing samples, independent of everything else, and the expectation is with respect to distribution $P_{\tilde{Z}}P_{Z^n,W}$, where the joint distribution $P_{Z^n,W}$ is induced by the algorithm $W = \frac{1}{n}\sum_{i=1}^{n} Z_i$. It is straightforward to show that the true generalization error is $2\sigma^2/n$.

In this work, we shall consider a slightly more general version of the sample-average algorithm that $W = \sum_{i=1}^{n}\alpha_i Z_i + N$, where N is a Gaussian noise $\sim N(0, \sigma_N^2)$, independent of $Z_{[n]}$, and α_i's are nonnegative weights such that $\sum_{i=1}^{n}\alpha_i = 1$. It can be shown that the true generalization error is also $2\sigma^2/n$ (see the Appendix).

2.5.2 A New Information-Theoretic Generalization Error Bound

The new information-theoretic generalization error bound is summarized in the following theorem.

Theorem 9. *Let $F_i = L_\xi(W) - \ell(W, Z_i)$, then we have*

$$\begin{aligned}
\text{gen}(\xi, P_{W|Z_{[n]}}) &\le \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{P_{Z_i}}\left[\inf_{\lambda>0}\frac{D(P_{W|Z_i}\|Q_W^i) + \Lambda_{F_i|Z_i,Q_W^i}(\lambda)}{\lambda}\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{P_{Z_i}}\left[\Lambda^{*-1}_{F_i|Z_i,Q_W^i}\left(D(P_{W|Z_i}\|Q_W^i)\right)\right], \qquad (2.90)
\end{aligned}$$

for any $Q^i_{W,Z_i} = Q^i_W P_{Z_i}$, $i = 1, 2, \ldots, n$, i.e., a distribution Q^i where W is independent of Z_i.

The reference distribution Q can be optimized, which would provide the tightest bound for a fixed learning algorithm. This bears a certain resemblance to those used in [30] which considers the computation of tight generalization bound using the PAC-Bayesian approach.

Proof. We start from

$$\begin{aligned}
\text{gen}(\xi, P_{W|Z_{[n]}}) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[(\ell(W, \tilde{Z}_i) - \ell(W, Z_i))\right] && (2.91)\\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[L_\xi(W) - \ell(W, Z_i))\right], && (2.92)
\end{aligned}$$

and consider each summand on the right-hand side

$$
\begin{aligned}
&\mathbb{E}_{P_{W,Z_i}}\left[L_\xi(W)-\ell(W,Z_i)\right]=\mathbb{E}_{P_{Z_i}}\Big[\mathbb{E}_{P_{W|Z_i}}\big((L_\xi(W)-\ell(W,Z_i)\big|Z_i)\Big]\\
&\le \mathbb{E}_{P_{Z_i}}\left[\inf_{\lambda>0}\frac{D(P_{W|Z_i}\|Q_W^i)+\Lambda_{F_i|Z_i,Q_W^i}(\lambda)}{\lambda}+\mathbb{E}_{Q_W^i}\left((L_\xi(W)-\ell(W,Z_i)\middle|Z_i\right)\right]\\
&=\mathbb{E}_{P_{Z_i}}\left[\inf_{\lambda>0}\frac{D(P_{W|Z_i}\|Q_W^i)+\Lambda_{F_i|Z_i,Q_W^i}(\lambda)}{\lambda}\right],
\end{aligned}
\tag{2.93}
$$

where the first equality is by the tower rule, the inequality is by (2.82), and the second equality is due to that for any algorithm $Q_{W|Z_{[n]}}$ that is independent of $Z_{[n]}$,

$$
\mathrm{gen}(\xi,Q_{W|Z_{[n]}})=\mathbb{E}_Q\left[L_\xi(W)-L_{Z_{[n]}}(W)\right]=0. \tag{2.94}
$$

Summing over i gives the bound stated in the theorem. □

As will be shown in the next section, this bound is exactly tight for the quadratic Gaussian setting, and therefore, it can be viewed as a tight bound in the sense that it cannot be strictly improved uniformly, either in terms of the constant or in the scaling. This bound can be loosened in several ways, which are stated in the following corollaries.

Corollary 3. *Let $F_i=L_\xi(W)-\ell(W,Z_i)$, then we have*

$$
\begin{aligned}
\mathrm{gen}(\xi,P_{W|Z_{[n]}})&\le\frac{1}{n}\sum_{i=1}^n\inf_{\lambda>0}\mathbb{E}\left[\frac{D(P_{W|Z_i}\|Q_W^i)+\Lambda_{F_i,Q_W^i}(\lambda)}{\lambda}\right]\\
&\le\inf_{\lambda>0}\left[\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[\frac{D(P_{W|Z_i}\|Q_W^i)+\Lambda_{F_i,Q_W^i}(\lambda)}{\lambda}\right]\right],
\end{aligned}
\tag{2.95}
$$

for any $Q_{W,Z_i}^i=Q_W^iP_{Z_i}$, $i=1,2,\ldots,n$.

The first inequality is obtained by exchanging expectation and the infimum operation, and the second is obtained by exchanging the summation and the infimum.

Corollary 4. *Let $F_i = L_\xi(W) - \ell(W, Z_i)$, then we have*

$$\begin{aligned}
\operatorname{gen}(\xi, P_{W|Z_{[n]}}) &\le \mathbb{E} \inf_{\lambda>0} \left[\frac{1}{n} \sum_{i=1}^{n} \left[\frac{D(P_{W|Z_i} \| Q_W^i) + \Lambda_{F_i, Q_W^i}(\lambda)}{\lambda} \right] \right] \\
&\le \inf_{\lambda>0} \left[\frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[\frac{D(P_{W|Z_i} \| Q_W^i) + \Lambda_{F_i, Q_W^i}(\lambda)}{\lambda} \right] \right] \qquad (2.96)
\end{aligned}$$

for any $Q_{W,Z_i}^i = Q_W^i P_{Z_i}$, $i = 1, 2, \ldots, n$.

The first inequality is obtained by exchanging the expectation and the summation, and the second by exchanging the infimum and the expectation.

Remark. The second bounds in Corollaries 3 and 4 are the same, while the first bounds are not directly comparable.

Notice that when $Q_{W,Z_i}^i = P_W \otimes P_{Z_i}$, i.e., the product of the marginals of P_{W,Z_i}, we have $\mathbb{E}[D(P_{W|Z_i} \| Q_W^i)] = I(W; Z_i)$. This leads to the following corollary.

Corollary 5. *Let $F_i = L_\xi(W) - \ell(W, Z_i)$, then we have*

$$\begin{aligned}
\operatorname{gen}(\xi, P_{W|Z_{[n]}}) &\le \frac{1}{n} \sum_{i=1}^{n} \inf_{\lambda>0} \left[\frac{I(W; Z_i) + \mathbb{E}\Lambda_{F_i, P_W}(\lambda)}{\lambda} \right] \\
&\le \frac{1}{n} \sum_{i=1}^{n} \inf_{\lambda>0} \left[\frac{I(W; Z_i) + \Lambda_{F_i, P_W P_{Z_i}}(\lambda)}{\lambda} \right], \\
&= \frac{1}{n} \sum_{i=1}^{n} \Lambda^{*-1}_{F_i, P_W P_{Z_i}} (I(W; Z_i)) \qquad (2.97)
\end{aligned}$$

where the second inequality is due to the concavity of the $\ln(\cdot)$ *function.*

By exchanging the infimum and the summation, we straightforwardly obtain further that

$$\begin{aligned}
\operatorname{gen}(\xi, P_{W|Z_{[n]}}) &\le \inf_{\lambda>0} \left[\frac{1}{n} \sum_{i=1}^{n} \left[\frac{I(W; Z_i) + \mathbb{E}\Lambda_{F_i, P_W}(\lambda)}{\lambda} \right] \right] \\
&\le \inf_{\lambda>0} \left[\frac{1}{n} \sum_{i=1}^{n} \left[\frac{I(W; Z_i) + \Lambda_{F_i, P_W P_{Z_i}}(\lambda)}{\lambda} \right] \right]. \qquad (2.98)
\end{aligned}$$

The second bound in (2.97) is quite similar to the main theorem in [8]. However, there is a major difference even when we assume the reference distribution Q is the same as the product of the marginals in P: the function F we choose to bound is different.

2.5.3 Bounding the Quadratic Gaussian Problem Generalization Error

2.5.3.1 Exactly Tight Bounds for the Quadratic Gaussian Setting

The expected generalization error of interest in the quadratic Gaussian setting is

$$\mathrm{gen}(\xi, P_{W|Z_{[n]}}) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\sigma^2+\mu^2-Z_i^2+2(Z_i-\mu)W|Z_i\right]\right]. \tag{2.99}$$

For any fixed i, define

$$F_i = f_{Z_i}(W) := \sigma^2+\mu^2-Z_i^2+2(Z_i-\mu)W.$$

Note the conditional distribution

$$W|Z_i \overset{P}{\sim} N\left(\mu+\alpha_i(Z_i-\mu), \sum_{j\neq i}\alpha_j^2\sigma^2+\sigma_N^2\right). \tag{2.100}$$

We will choose the reference distribution Q_W^i as

$$W \overset{Q_W^i}{\sim} N\left(\mu, \sum_{j\neq i}\alpha_j^2\sigma^2+\sigma_N^2\right), \tag{2.101}$$

which is indeed independent of Z_i.

Remark. In the reference distribution Q_{W,Z_i}^i, W and Z_i are independent, and the marginal distribution Q_W^i is not the same as that marginalized from $P_{W,Z_{[n]}}$. More specifically, the latter is

$$P_W \sim N\left(\mu, \sum_{i=1}^{n}\alpha_i^2\sigma^2+\sigma_N^2\right),$$

which can be compared with (2.101).

With these conditional distributions, we can derive (see appendix) that

$$
\begin{aligned}
D(P_{W|Z_i}||Q_W^i) &= \alpha_i^2(Z_i-\mu)^2\frac{1}{2\sum_{j\neq i}\alpha_j^2\sigma^2+2\sigma_N^2};\\
\Lambda_{F_i,Q_W^i}(\lambda) &= 2(Z_i-\mu)^2\left(\sum_{j\neq i}\alpha_j^2\sigma^2+\sigma_N^2\right)\lambda^2. \qquad (2.102)
\end{aligned}
$$

Therefore

$$
\begin{aligned}
\mathbb{E}[D(P_{W|Z_i}||Q_W^i)] &= \alpha_i^2\sigma^2\frac{1}{2\sum_{j\neq i}\alpha_j^2\sigma^2+2\sigma_N^2};\\
\mathbb{E}[\Lambda_{F_i,Q_W^i}(\lambda)] &= 2\sigma^2\left(\sum_{j\neq i}\alpha_j^2\sigma^2+\sigma_N^2\right)\lambda^2. \qquad (2.103)
\end{aligned}
$$

Applying the first bound in Corollary 3, we obtain

$$
\begin{aligned}
\mathrm{gen}(\xi, P_{W|Z_{[n]}}) &\leq \frac{1}{n}\sum_{i=1}^{n}\inf_{\lambda>0}\mathbb{E}\left[\frac{D(P_{W|Z_i}\|Q_W^i)+\Lambda_{F_i,Q_W^i}(\lambda)}{\lambda}\right]\\
&= \frac{1}{n}\sum_{i=1}^{n}\inf_{\lambda>0}\left[\frac{\mathbb{E}[D(P_{W|Z_i}\|Q_W^i)]+\mathbb{E}[\Lambda_{F_i,Q_W^i}(\lambda)]}{\lambda}\right]\\
&= \frac{2\sigma^2}{n}, \qquad (2.104)
\end{aligned}
$$

where the last equality is by choosing the minimizer λ_i^* as

$$
\lambda_i^* = \frac{\alpha_i}{2\sum_{j\neq i}\alpha_j^2\sigma^2+2\sigma_N^2}. \qquad (2.105)
$$

In contrast, the second bound in Corollary 3 and the first bound in Corollary 4 are not tight for general assignments of α_i's, due to the fact that the optimal λ_i^* is index-dependent. In the extreme case, consider setting $\alpha_1 = 1$ and $\alpha_i = 0$ for $i = 2, 3, \ldots, n$. Then the second bound in Corollary

3 gives

$$\begin{aligned}
\operatorname{gen}(\xi, P_{W|Z_{[n]}}) &= \frac{1}{n}\inf_{\lambda>0}\left[\frac{\frac{\sigma^2}{2\sigma_N^2} + 2(n-1)\sigma^2\left(\sigma^2+\sigma_N^2\right)\lambda^2 + 2\sigma^2\sigma_N^2\lambda^2}{\lambda}\right] \\
&= \frac{2\sigma^2}{n}\sqrt{\frac{2(n-1)\left(\sigma^2+\sigma_N^2\right)+\sigma_N^2}{2\sigma_N^2}}, \qquad (2.106)
\end{aligned}$$

which is of order $\mathcal{O}(1/\sqrt{n})$. However, when $\alpha_i = 1/n$, this dependence disappears and the loosened bounds also become tight. Indeed, consider the second bound in Corollary 3 for this case, we have

$$\operatorname{gen}(\xi, P_{W|Z_{[n]}}) = \frac{1}{n}\inf_{\lambda>0}\left[\frac{\sum_{i=1}^{n}(\mathbb{E}[D(P_{W|Z_i}\|Q_W^i)] + \mathbb{E}[\Lambda_{F_i,Q_W^i}(\lambda)])}{\lambda}\right] = \frac{2\sigma^2}{n}, \qquad (2.107)$$

where the last step is obtained by choosing

$$\lambda^* = \frac{\alpha_i}{2\sum_{j\neq i}\alpha_j^2\sigma^2 + 2\sigma_N^2} = \frac{n}{2(n-1)\sigma^2 + 2n\sigma_N^2}. \qquad (2.108)$$

2.5.3.2 *Looseness of Mutual Information Based Bounds*

One remaining question is whether we can obtain a tight or asymptotically tight generalization error bound in the quadratic Gaussian setting using a mutual-information-based bound. To understand this issue, we consider the bounds in Corollary 5 assuming the coefficients $\alpha_i = 1/n$ for $i = 1, 2, \ldots, n$. Note that in this case, the choice of the reference distribution Q_W^i is fixed as the marginal of P_W.

The various term we need when applying Corollary 5 in this setting can be shown to be (see the appendix)

$$\begin{aligned}
I(W; Z_i) &= \frac{1}{2}\log\frac{n}{n-1} \\
\mathbb{E}\Lambda_{F_i,Q_W^i}(\lambda) &= \frac{2\sigma^4(n-1)}{n^2}\lambda^2
\end{aligned}$$

$$\Lambda_{F_i,Q^i_{W,Z_i}}(\lambda) = \lambda\sigma^2 - \frac{1}{2}\log\left[1 - 2\left(\frac{2\lambda^2\sigma^4}{n} - \lambda\sigma^2\right)\right].$$

With these quantities, it follows that the first bound in Corollary 5 is

$$\mathrm{gen}(\xi, P_{W|Z_{[n]}}) \le \frac{2\sigma^2}{n}\sqrt{\left(\log\frac{n}{n-1}\right)(n-1)}. \tag{2.109}$$

The bound is of order $\mathcal{O}(1/n)$; in fact, it is asymptotically optimal in the sense that it approaches $\frac{2\sigma^2}{n}$. Therefore, the first mutual-information-based bound in Corollary 5 does not lose the tightness in a significant manner compared to the KL-based bound of those in Corollaries 3 and 4.

The second bound in Corollary 5 has the form

$$\mathrm{gen}(\xi, P_{W|Z_{[n]}}) \le \sigma^2 + \inf_{\lambda>0}\left[\frac{1}{2\lambda}\log\frac{n}{n-1} - \frac{1}{2\lambda}\log\left[1 - 2\left(\frac{2\lambda^2\sigma^4}{n} - \lambda\sigma^2\right)\right]\right], \tag{2.110}$$

for any $\delta \in (0, 1/2]$, and any $\epsilon > 0$, by choosing $\lambda = 1/(2n^\delta\sigma^2)$, it can be seen that for sufficiently large n, we have $\mathrm{gen}(\xi, P_{W|Z_{[n]}}) \le (1+\epsilon)\frac{2\sigma^2}{n^{1-\delta}}$. Therefore, the bound can be also viewed as asymptotically optimal.

Similarly, we can apply the bounds in (2.98). Since in this case, the optimal choice of λ does not depend on the index-i, they are also asymptotically optimal. It should be noted that when the weight coefficients α_i are not chosen to be uniform, then the optimal λ becomes dependent on the index i, and the bounds in (2.98) will be looser, in a similar manner as that for the KL-based bounds.

2.6 Conclusion

We proposed an information-theoretic generalization error bound, referred to as the ICIMI bound, based on a combination of the error decomposition technique and the conditional mutual information structure. Due to the reduced information content in the conditioning term, the proposed bound can be tighter, and in some cases significantly tighter, than several existing bounds. Particularly, when the loss function is bounded, it can be shown that the proposed bound is always tighter than the CMI and the CIMI bounds. A conditional decoupling lemma is provided which leads to a unified framework to study and compare these bounds, and it may be of independent interest. As applications, we studied a logistic regression setting where the mutual information value needs to be estimated from the data and also analyzed the SGLD algorithm and derived an upper bound on its generalization error with minimum restrictions on the loss function. We also proposed a new chaining-based approach to bound the generalization error by replacing the hierarchical partitions with a stochastic chain. The proposed approach can firstly remove naturally the restriction for the metric space to be bounded, and secondly, it helps to simplify the computation, and lastly, it can remove the geometric constraints in designing the hierarchical partitions in some cases. Two examples are used to illustrate that the proposed approach can overcome some difficulties in applying the chaining mutual information approach. The roles that chaining can play in bounding generalization error in conjunction with other information-theoretic approaches, such as the conditional mutual information [9], information density [31], and Wasserstein distance [32], as well as the possible application in noisy and stochastic learning algorithms, call for further research.

3. APPROXIMATE TOP-M ARM IDENTIFICATION WITH HETEROGENEOUS VARIANCES*

Besides capturing the interplay between learning algorithms and samples as illustrated in Chapter 2, information measures can also be useful to characterize the complexity of the problems. In this chapter, we consider the (ϵ, δ) top-m arm identification problem, where the reward variances are heterogeneous. We propose an optimal divide-and-conquer style algorithm with a matching lower bound. The characterized worst-case sample complexity, where the variance heterogeneity is measured by an Entropy-like function.

3.1 Preliminary

System model: We largely follow the canonical sub-Gaussian bandit model, except for the additional component related to the reward variances. A bandit instance I is represented by a set of arm indices $[n] := \{1, 2, \ldots, n\}$ and the tuple of reward distributions $(\nu_1, \nu_2, \ldots, \nu_n)$. For any $i \in [n]$, pulling the i-th arm returns a reward observation, which is independently sampled from distribution ν_i, where ν_i is a sub-Gaussian distribution with mean μ_i and variance proxy σ_i^2. A random variable X follows some σ^2-sub-Gaussian distribution, if $\ln \mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}] \leq \frac{\sigma^2\lambda^2}{2}$, $\forall \lambda \in \mathbb{R}$, and σ^2 is called the variance proxy. An arm is ϵ-approximate top-m if the mean reward of that arm is at least $\max^m_{i\in[n]} \mu_i - \epsilon$, where $\max^m_{i\in[n]}$ indicates the m-th largest (mean reward) value among the arms in $[n]$. With the knowledge of variance proxy values $\sigma^2_{1:n}$, but without the knowledge of mean values $\mu_{1:n}$, the agent actively learns the parameters of the sub-Gaussian bandit instance I by observing independent reward samples. When there is no ambiguity from the context, we omit “proxy" and simply refer to $\sigma^2_{1:n}$ as the reward variances.

(ϵ, δ) **top-m arm identification:** In the (ϵ, δ) top-m arm identification problem, the agent is required to identify some subset $R \subset [n]$ with $|R| = m$, such that, with probability at least $1 - \delta$,

[*] Part of this chapter is reprinted with permission from “Approximate top-m arm identification with heterogeneous reward variances,” by Rudia Zhou and Chao Tian, 2022, *International Conference on Artificial Intelligence and Statistics*, pp. 7483–7504, PMLR, Copyright © Ruida Zhou and Chao Tian [33].

any arm in R is ϵ-approximate top-m.

Algorithm class: Taking the parameters $(\epsilon, \delta, m, [n], \sigma^2_{1:n})$ as input, an algorithm A deployed by the agent is represented by a tuple $(\pi_t, \rho_t)_{t \geq 1}$. During the learning process, the function π_t selects an arm in $[n]$ based on the inputs of the algorithm as well as the previous observations before time step t (i.e., the arms that were pulled). The function ρ_t decides whether to stop based on the inputs of the algorithm as well as the available observations (the current observation and the previous observations before time step t). If ρ_t decides to stop, it returns a set of arms $R^{\mathsf{A}} \subset [n]$; otherwise, the process continues. Let T^{A} be the time that the process stops, which is the number of samples observed by algorithm A. We only study the *valid* algorithms that solve the (ϵ, δ) top-m arm identification when dealing with any bandit instance.

Worst-case sample complexity: The number of samples observed by the algorithm T^{A} is a stopping time, whose expectation the agent aims to minimize. We study the *worst-case sample complexity* for (ϵ, δ) top-m arm identification, which is an intrinsic quantity that measures the difficulty of the problem, and thus independent of the algorithm and $\mu_{1:n}$. Formally, the worst-case sample complexity of the (ϵ, δ) top-m arm identification problem under algorithm inputs $(\epsilon, \delta, m, [n], \sigma^2_{1:n})$ is

$$\mathrm{SC}(\epsilon, \delta, m, [n], \sigma^2_{1:n}) := \inf_{\mathsf{A}} \sup_{I \in \mathcal{I}(\sigma^2_{1:n})} \mathbb{E}_I[T^{\mathsf{A}}], \tag{3.1}$$

where the infimum is taken over all valid algorithms, the supremum is taken over the instance class $\mathcal{I}(\sigma^2_{1:n})$ containing all the distribution tuples $\nu_{1:n}$ with variances $\sigma^2_{1:n}$, and the subscript I in the expectation $\mathbb{E}_I[\cdot]$ indicates that it is with respect to the bandit model I.

Measure of heterogeneity: For any positive vector $a_{1:n}$, define the entropy function as $\mathrm{Ent}(a_{1:n}) := -\sum_{j=1}^n \hat{a}_i \ln \hat{a}_i$ with $\hat{a}_i = \frac{a_j}{\sum_{i=1}^n a_i}$. It measures the heterogeneity of the vector $a_{1:n}$, and takes value within $(0, \ln(n)]$. Note that the entropy function is usually defined on the probability simplex, and we slightly abused the notation by defining it for a positive vector. We study the worst-case sample complexity, which is gap-independent.

Related works: Multi-armed bandit problems have been extensively studied in the machine learning community in the past decades. A canonical setting is to maximize the cumulative reward, whose asymptotically optimal behavior was first characterized in the seminal work by [34]. Good tutorials and books [35–37] are readily available.

An alternative setting is to instead identify the best arm. There are in general two lines of research: minimizing the misidentification probability within a fixed budget of samples [38–40], and fast identification with a fixed confidence guarantee [41]. The (ϵ, δ) best arm identification problem belongs to the latter and was introduced in [42, 43], where several elimination-based algorithms, such as naive elimination, successive elimination, and median elimination algorithms, were proposed. The median elimination algorithm was shown to be worst-case optimal for which a matching lower bound was derived by [44]. The asymptotic (large number of arms) optimal elimination algorithm was recently discovered [45], which was inspired by the idea of identifying the "good arms" [46]. The case of exact best arm identification, i.e., $\epsilon = 0$, motivates algorithms that adapt to the underlining model which usually perform well in an instance-dependent manner [47–51].

There are multiple variants of the problem [52–56]. One of the most natural generalizations of the best arm identification problem is to identify multiple best arms. The (ϵ, δ) top-m arm identification was studied in [57], in which an algorithm named "halving" was proposed, and it bears similarity to the median elimination algorithm. It was later shown that the halving algorithm is indeed worst-case optimal [58]. Though more adaptive algorithms were proposed later, such as LUCB [57] and UGapE [58, 59], they are not worst-case optimal. For the case of exact top-m arm identification, efforts toward understanding the instance-dependent sample complexity were also made [60–62].

Gaussian rewards with heterogeneous variances was considered in the earliest work on best arm identification [63] in the fixed confidence setting, though without a theoretical analysis of the stopping time. The possible variance heterogeneity among arms gained attention recently in the fixed budget setting [64], where the confidence bounds are designed based on the central limit theorem. Identifying the best arms in multiple bandits with possible heterogeneous variances was studied in the fixed budget setting [65], where an elimination-based algorithm was proposed to take variances into designing confidence bound. In addition to the fixed budget setting, most recently [66] studied the best arm identification with unknown heterogeneous variances in the fixed confidence setting. They assumed the support of reward distribution is bounded, and proposed an elimination-based algorithm by first estimating the variances (with known upper bound on the variances) and then utilizing the estimated variances in identifying the unique best arm based on Bernstein-style confidence bounds. The algorithm achieves near-optimal instance-dependent performance. In comparison, we aim to study the *worst-case sample complexity* with known variance proxies as inputs (the support of reward distribution may be unbounded), in the *top-m identification* problem. We propose an optimal algorithm with an *exact matching* lower bound and studied the impact of variances transition from the homogeneous setting to the heterogeneous setting in terms of the parameter m.

3.2 Worst-case Sample Complexity

The main result of this work is the characterization of the worst-case sample complexity SC$(\epsilon, \delta, m, [n], \sigma_{1:n}^2)$. To present this result, we first introduce some additional notation. Let $\underline{\sigma} := \min_{i\in[n]} \sigma_i$, and partition $[n]$ into k disjoint subsets $G_1, \ldots, G_k$, such that for any $j \in [k]$,

$$G_j := \{i \in [n] : 2^{j-1} \leq \sigma_i^2/\underline{\sigma}^2 < 2^j\}. \tag{3.2}$$

Define two disjoint sets

$$G^m := \cup_{j:|G_j|>2m} G_j, \quad G^l := \cup_{j:|G_j|\leq 2m} G_j, \tag{3.3}$$

where $|\cdot|$ denotes the cardinality of the set. For each j with $G_j \subset G^l$, let $G_j' = G_j$; for each j with $G_j \subset G^m$, select $G_j' \subset G_j$ with $|G_j'| = 2m$, and denote $G^r := \cup_{j\geq 1} G_j'$ as a subset of the arms, such that $\mathrm{Ent}\,(\sigma_{G^r}^2)$ is maximized. (The superscripts of G^m, G^l, G^r indicate “more", “less" and “reduced", respectively).

The worst-case sample complexity $\mathrm{SC}(\epsilon, \delta, m, [n], \sigma^2_{1:n})$ is summarized in the following theorem.

Theorem 10. *Suppose $n > 2m$, $\epsilon > 0$ and $0 < \delta < 0.1$, then the worst-case sample complexity is*

$$\mathrm{SC}(\epsilon, \delta, m, [n], \sigma^2_{1:n}) = \Theta\left(\sum_{i\in[n]} \frac{\sigma_i^2}{\epsilon^2}\ln\frac{1}{\delta} + \sum_{i\in G^m} \frac{\sigma_i^2}{\epsilon^2}\ln(m) + \sum_{j\in G^l} \frac{\sigma_j^2}{\epsilon^2}\mathrm{Ent}(\sigma^2_{G^r})\right). \tag{3.4}$$

The following lemma upper bounds the entropy $\mathrm{Ent}(\sigma^2_{G^r})$ in the third component.

Lemma 6. *For any $m \geq 2$, $\mathrm{Ent}(\sigma^2_{G^r}) \leq 8\ln(m)$.*

This lemma indicates that the worst-case sample complexity in the heterogeneous variance setting is upper bounded by $O\left(\sum_{i\in[n]} \frac{\sigma_i^2}{\epsilon^2}\ln\frac{m}{\delta}\right)$ in general. In a certain sense, the heterogeneity makes the problem “easier" to solve. To further illustrate this point, let us consider two special cases:

- When the variances are more homogeneous, e.g., in the extreme case $\sigma_i^2 = \sigma^2,\ \forall i \in [n]$, we have $G^m = [n]$ and $G^l = \emptyset$. Theorem 10 naturally degrades to the worst-case sample complexity in the homogeneous setting characterized in [58], which is $\Theta\left(\frac{n\sigma^2}{\epsilon^2}\ln\frac{m}{\delta}\right)$.
- When the variances are highly heterogeneous, e.g., in the extreme case $|G_j| = 1, \forall j = 1, 2, \ldots, k$, we have $G^m = \emptyset$ and $G^l = [n]$. Theorem 10 shows that the worst-case sample complexity is $\Theta\left(\sum_{i\in[n]} \frac{\sigma_i^2}{\epsilon^2}\ln\frac{1}{\delta}\right)$, which is independent of m.

Comparing the two cases and assuming the sum of the variances remains the same, the latter clearly has a more desirable sample complexity. The sets G^m and G^l describe the transition between the homogeneous and the heterogeneous. In the rest of this article, we present the optimal algorithm and the matching lower bound to establish Theorem 10.

3.3 Algorithms

We first revisit several existing algorithms designed mostly under the assumption of homogeneous variances. By adapting them to the heterogeneous variance case, we analyze their advantages and disadvantages. As will become clear shortly, these adapted algorithms still perform well in certain respective cases. Based on this observation, we will propose an optimal divide-and-conquer style algorithm.

3.3.1 Adapting Existing Algorithms

3.3.1.1 Weighted naive elimination:

In this adapted algorithm, the agent simply pulls each arm-i a total of $\frac{2\sigma_i^2}{(\epsilon/2)^2}\ln\frac{1}{\omega_i}$ times, calculates the sample mean $\hat{\mu}_i$, and returns the m arms with the largest sample means. We call it "weighted" because the numbers of pulls for the arms are determined by the reward variances $\sigma_{1:n}^2$ and the confidence parameters $\omega_{1:n}$. The parameters $\omega_{1:n}$ need to be optimized in order to provide the performance guarantee, and the following lemma provides one such assignment of the optimized $\omega_{1:n}$.

Lemma 7. *Let $\omega_i = \delta\frac{\sigma_i^2}{\sum_{j=1}^n \sigma_j^2}$, the weighted naive elimination algorithm takes*

$$8\sum_{i\in[n]}\frac{\sigma_i^2}{\epsilon^2}\left(\ln\frac{1}{\delta}+\mathrm{Ent}(\sigma_{1:n}^2)\right) \tag{3.5}$$

samples, and solves the (ϵ,δ) top-m arm identification problem for any $\epsilon > 0$ and $0 < \delta < 1$.

We will use $\mathsf{WNElim}(\epsilon,\delta,m,[n],\sigma_{1:n}^2)$ to denote the weighted naive elimination algorithm with the choices of $\omega_{1:n}$ in Lemma 7. The entropy function $\mathrm{Ent}(\sigma_{1:n}^2)$ appears naturally as a multiplicative factor in the second item of Equation (3.5), which measures the heterogeneity of the variances. If the variance heterogeneity is high, the entropy term $\mathrm{Ent}(\sigma_{1:n}^2)$ can be significantly less than $\log n$. As mentioned earlier, when $\sigma_i^2 = 2^i$, the entropy term is $O(1)$, i.e., no longer a function of n and m. On the other hand, by the principle of maximum entropy [67], it has the maximum value $\ln(n)$ when the variances are homogeneous. Thus the weighted naive elimination algorithm will provide good performance when the arm variances are highly heterogeneous but will lose efficiency when they are more homogeneous.

3.3.1.2 Adapted median elimination:

Median Elimination (“Halving" algorithm in [57]) is known to achieve the worst-case optimal performance in the homogeneous variance setting. One simple method to adapt it to the heterogeneous setting is to ignore the knowledge of the heterogeneity, and simply assume that all the arms have the largest variance $\max_{i\in[n]} \sigma_i^2$. The original median elimination algorithm can be applied without any change, and the expected number of samples taken is thus $O\left(\frac{n \max_{i\in[n]} \sigma_i^2}{\epsilon^2}\left(\ln\frac{1}{\delta}+\ln m\right)\right)$, as shown in [57].

If the variances are more homogeneous, e.g., $\sigma_i^2/\sigma_j^2 \leq 2, \forall i,j \in [n]$, then $\sum_{i\in[n]}\sigma_i^2 \leq n\max_{i\in[n]}\sigma_i^2 \leq 2\sum_{i\in[n]}\sigma_i^2$ and the expected number of samples is $O\left(\frac{\sum_{i\in[n]}\sigma_i^2}{\epsilon^2}\left(\ln\frac{1}{\delta}+\ln m\right)\right)$. For the same example, the weighted naive elimination uses $O\left(\frac{\sum_{i\in[n]}\sigma_i^2}{\epsilon^2}\left(\ln\frac{1}{\delta}+\ln n\right)\right)$ samples. Thus this simple adaptation of the median elimination algorithm is able to perform well for the highly homogeneous case but will induce a loss of performance for the more heterogeneous cases.

3.3.1.3 Adapting other algorithms:

The adaptation of several instance-dependent algorithms, such as LUCB and UGapE, is straightforward. For the problem in consideration, both algorithms require $O\left(\frac{\sum_{i\in[n]}\sigma_i^2}{\epsilon^2}\left(\ln\frac{1}{\delta}+\ln\frac{\sum_{i\in[n]}\sigma_i^2}{\epsilon^2}\right)\right)$ number of samples in expectation in the worst case. They are not worst-case optimal in the homogeneous variance setting, and certainly not in the heterogeneous variance setting since the latter is a more general setting.

3.3.2 The Optimal Variance-Grouped MedElim Algorithm

It was shown in the previous subsection that the weighted naive elimination algorithm and the median elimination algorithm have advantages in the respective cases. In order to retain the advantages of both algorithms, we take a “divide and conquer" approach. Recall the minimum variance is $\underline{\sigma} = \min_{i\in[n]}\sigma_i$, and the disjoint subsets $G_1,\ldots,G_k$ form a partition of $[n]$, and for any

$j \in [k]$,

$$G_j = \left\{ i \in [n] : 2^{j-1} \le \sigma_i^2 / \underline{\sigma}^2 < 2^j \right\}. \tag{3.6}$$

The largest variance ratio within each subset is at most 2, while the variances among subsets are well separated. We wish to apply median elimination to each subset and select “good" arms within that subset, and then apply weighted naive elimination over all the selected “good" arms. However, the “good" arms within a subset can be “bad" in terms of the overall arm set $[n]$. To see this, consider the following example instance: m arms have a mean reward ϵ, and the rest of $n-m$ arms have a mean reward $-\epsilon$. Then any ϵ-approximate top-m arms need to have mean ϵ. Suppose the subset G_1 contains $m' < m$ arms with mean ϵ and some other arms with mean $-\epsilon$. Ideally, we would like to apply median elimination to find those top-m' arms with mean ϵ within G_1. However, parameter m' is not known, and we will apply median elimination on G_1 by selecting some l arms. If $l < m'$, then the returned l arms will not include all the top-m' arms in G_1, and therefore fail to identify the final top-m arms. On the other hand, if $l > m'$, then $\max^l_{i \in G_1} \mu_i = -\epsilon$. Any arm in G_1 is ranked in the top-l within G_1, and the problem is trivial to solve. The returned l arms, even though are top-l within G_1, are not guaranteed to contain those top-m' arms with mean reward ϵ.

To successfully apply the divide-and-conquer approach, we need a “blind” algorithm that returns a subset containing the approximate top-m' arms, ideally with a graceful transition of the confidence values.

Definition 2. *The algorithm $\mathsf{A}(\epsilon, \delta, m, [n], \sigma^2_{1:n})$ is said to satisfy the (ϵ, δ') top-m' condition, where $m' \le m$, if with probability at least $1-\delta'$, $\max^{m'}_{j \in R^{\mathsf{A}}} \mu_j \ge \max^{m'}_{j \in [n]} \mu_i - \epsilon$.*

The condition is equivalent to the standard (ϵ, δ) top-m arm identification requirement, if $m' = m$ and $\delta' = \delta$. We first restate the median elimination algorithm presented in Algorithm 1 (the halving algorithm [57]), with the necessary changes on the constants and the variance values taken into account (note the input $2m$).

The following lemma summarizes the sample complexity of the MedElim algorithm with the

```
Algorithm 1: MedElim($\epsilon, \delta, 2m, [n], \sigma^2_{1:n}$)
  Initialize $S_1 = [n]$, $\ell = 1$ and $\epsilon_\ell = (\epsilon/3)\frac{3^\ell}{4^\ell},\ \delta_\ell = \frac{\delta/4}{2^\ell}$
  while $|S_\ell| > 2m$ do
      Pull arm-$i$ $t_{i,\ell} = \frac{2\sigma_i^2}{(\epsilon_\ell/2)^2}\ln\frac{m}{\delta_\ell}$ times and calculate their sample mean $\hat{\mu}_{i,\ell}$ for each $i \in S_\ell$
      Update the candidate set as $S_{\ell+1} = \arg\max_{i\in S_\ell}^{1:\max(\lfloor|S_\ell|/2\rfloor, 2m)} \hat{\mu}_{i,\ell}$
      Let $\ell = \ell + 1$
  Return $S_\ell$
```

aforementioned transition in the confidence values for $m' = 1, 2, \ldots, m$ for the $2m$ return arms. This algorithm will be used as a building block for the variance-grouped median elimination algorithm given next.

Lemma 8. *For any $\sigma^2_{1:n}$, if $\max_{i\in[n]}\sigma_i^2/\min_{j\in[n]}\sigma_j^2 \le 2$, the* MedElim *algorithm has an expected stopping time*

$$O\left(\frac{\sum_{i\in[n]}\sigma_i^2}{\epsilon^2}\left(\ln\frac{1}{\delta} + \ln(m)\right)\right). \tag{3.7}$$

Moreover, for any $m' \le m$, the MedElim algorithm satisfies the $(\epsilon, \frac{m'}{m}\delta)$ top-m' condition.

Now we are in a position to provide the proposed algorithm below, which we refer to as the variance-grouped median elimination algorithm.

```
Algorithm 2: V-MedElim($\epsilon, \delta, m, [n], \sigma^2_{1:n}$)
  Partition $[n]$ into groups $G_1, \ldots, G_k$ by (3.6)
  for $j \in 1 : k$ do
      $R_j =$ MedElim($\epsilon/2, \delta/2, 2m, G_j, \sigma^2_{G_j}$)
  Let $G = \cup_{j=1}^k R_j$
  $R =$ WNElim($\epsilon/2, \delta/2, m, G, \sigma^2_G$)
  Return $R$
```

The performance of the proposed algorithm is summarized in the following theorem.

Theorem 11. *The variance-grouped median elimination algorithm solves the (ϵ, δ) top-m arm identification problem for any $\epsilon > 0$ and $0 < \delta < 1$, and the expected number of samples is*

$$O\left(\sum_{i\in[n]} \frac{\sigma_i^2}{\epsilon^2} \ln \frac{1}{\delta} + \sum_{i \in G^m} \frac{\sigma_i^2}{\epsilon^2} \ln(m) + \sum_{j\in G^l} \frac{\sigma_j^2}{\epsilon^2} \mathrm{Ent}(\sigma_{G^r}^2)\right). \tag{3.8}$$

Proof of Theorem 11. Without loss of generality, assume $[m]$ is the set of top-m arms. For any j with $G_j \cap [m] \neq \emptyset$ and $i \in G_j \cap [m]$, arm-i must be one of top-$|G_j \cap [m]|$ arms in G_j. Let $m'_j = |G_j \cap [m]|$ be the number of top-m arms contained in G_j. By Lemma 8, with probability at least $1 - \frac{m'_j}{m}\frac{\delta}{2}$,

$$\max{}^{m'_j}_{l\in R_j} \mu_l \geq \max{}^{m'_j}_{l \in G_j \cap [m]} \mu_l - \epsilon/2 \geq \max{}^{m}_{l\in[n]} \mu_l - \epsilon/2. \tag{3.9}$$

It implies that with probability at least $1 - \sum_{j=1}^{k} \frac{m'_j}{m}\frac{\delta}{2} = 1 - \frac{\delta}{2}$, there are at least $\sum_{j=1}^{k} m'_j = m$ arms in $G = \cup_{j=1}^{k} R_j$ that are $\epsilon/2$-approximate top-m. In other words, event $\max^m_{l\in G} \mu_l \geq \max^m_{l\in[n]} \mu_l - \epsilon/2$ occurs with probability at least $1 - \frac{\delta}{2}$.

Conditioned on this event occurring, Lemma 7 implies that with probability at least $1 - \frac{\delta}{2}$, the returned set R of the weighted naive elimination over $G = \cup_{j=1}^{k} R_j$ satisfies

$$\min_{l\in R} \mu_l \geq \max{}^{m}_{l\in G} \mu_l - \epsilon/2 \geq \max{}^{m}_{l\in[n]} \mu_l - \epsilon. \tag{3.10}$$

Thus with probability at least $1 - \delta$, all arms in R are ϵ-approximate top-m.

Recall the definition of G^l, G^m, G^r in Section 3.2. The total number of samples used in the median elimination subroutine is $O\left(\sum_{i\in G^m} \frac{\sigma_i^2}{\epsilon^2} \left(\ln \frac{1}{\delta} + \ln(m)\right)\right)$. The number of samples used in the weighted naive elimination subroutine is $O\left(\sum_{i\in[n]} \frac{\sigma_i^2}{\epsilon^2} \left(\ln \frac{1}{\delta} + \mathrm{Ent}(\sigma_{G^r}^2)\right)\right)$. By Lemma 6, the expected total number of samples is $O\left(\sum_{i\in[n]} \frac{\sigma_i^2}{\epsilon^2} \ln \frac{1}{\delta} + \sum_{i\in G^m} \frac{\sigma_i^2}{\epsilon^2} \ln(m) + \sum_{j\in G^l} \frac{\sigma_j^2}{\epsilon^2} \mathrm{Ent}(\sigma_{G^r}^2)\right)$.

□

An illustrative example In the following example, we show the number of required samples by the variance-grouped median elimination algorithm given in Theorem 11 achieves an order-wise improvement over $\frac{\sum_{i\in[n]}\sigma_i^2}{\epsilon^2}(\ln(1/\delta)+\text{Ent}(\sigma_{1:n}^2))$ and $\frac{\sum_{i\in[n]}\sigma_i^2}{\epsilon^2}(\ln(1/\delta)+\ln(m))$. Take some integer $k \geq 2$ as an auxiliary parameter in this problem setting, and denote $\ell = \lceil \log(k) \rceil$. Let $\log(m) = k$ and $\log(n) = k^2$. We aim to approximately identify the top-m arms out of n arms. Among these n arms, there are 2^i arms with the same variance 2^{-i} for each $i = 0, 1, \ldots, \ell - 1$, and the rest $n - \sum_{i=0}^{\ell-1} 2^i = 2^{k^2} - 2^\ell + 1$ arms have the same variance $2^{-k^2}\ell/k$. Then G^m is the set of arms with variances $2^{-k^2}\ell/k$, and G^l is the set of arms with variances 2^{-i} for $i = 0, 1, \ldots, \ell - 1$. It is seen that

$$\sum_{j\in G^m} \sigma_j^2 = (2^{k^2} - 2^\ell + 1)2^{-k^2}\ell/k = \Theta(\ell/k), \tag{3.11}$$

$$\sum_{j\in G^l} \sigma_j^2 = \sum_{i=0}^{\ell-1} 2^i 2^{-i} = \ell = \Theta(\log(k)), \tag{3.12}$$

which implies $\sum_{j\in[n]} \sigma_j^2 = \Theta(\log(k))$. Furthermore, we can calculate that

$$\text{Ent}(\sigma_{G^r}^2) = \Theta(\text{Ent}(\sigma_{G^l}^2)) = \Theta(\log(k)). \tag{3.13}$$

Thus the number of required samples by the variance-grouped median elimination algorithm is of order

$$\Theta(\ln(k)\ln(1/\delta) + \ln(k)^2/\epsilon^2). \tag{3.14}$$

Since $\text{Ent}(\sigma_{1:n}^2) = \Theta(k)$ and $\ln(m) = \Theta(k)$, it is seen that $\frac{\sum_{i\in[n]}\sigma_i^2}{\epsilon^2}(\ln(1/\delta) + \text{Ent}(\sigma_{1:n}^2))$ and $\frac{\sum_{i\in[n]}\sigma_i^2}{\epsilon^2}(\ln(1/\delta) + \ln(m))$ are of the same order

$$\Theta(\ln(k)\ln(1/\delta) + k\ln(k)/\epsilon^2). \tag{3.15}$$

The detailed calculation of the entropy values used above is given in the supplementary material. Fix $\delta > 0$ as constant, comparing the numbers of required samples in (3.14) and (3.15), which are of order $\Theta(\ln(k)^2/\epsilon^2)$ and $\Theta(k\ln(k)/\epsilon^2)$, respectively, it is seen that the variance-grouped median elimination algorithm provides an order-wise improvement in this example setting by reducing a factor k to $\ln(k)$.

Remark. Our result establishes the theoretical optimality of the proposed algorithm through a matching lower bound provided in the following section. However, the empirical performance of the proposed algorithm suffers from large multiplicative factors introduced by the Median Elimination subroutine. More aggressive elimination-based algorithms, such as the algorithms proposed in [45], can be used as a subroutine to improve the multiplicative factor while maintaining the same order.

3.4 The Lower Bound

In the homogeneous variance setting, the previous lower bound [58] on worst-case (ϵ, δ)-PAC top-m identification leveraged the change-of-measure technique and was proved by contradiction. The approach leads to a large multiplicative factor and is also difficult to utilize in the heterogeneous variance case. The lower bound was later tightened and generalized to the instance-dependent case in [61] and [62]. Their approach assumed that the algorithms have a uniform preference over the arms at the beginning, which is reasonable in the homogeneous setting but not in the heterogeneous setting.

We derive a flexible simple inequality to better take into account the heterogeneous variances, given in Lemma 9. Applying this lemma, we formulate the lower bound as an optimization problem, whose dual formulation (Lemma 10) is then studied. The eventual lower bound is given in the following theorem, obtained by considering several feasible solutions to the dual problem.

Theorem 12. *There exists some universal constant $c > 0$, that for any $0 < \epsilon$, $0 < \delta < 0.1$, $m < n/2$, $\sigma_{1:n}^2$ and any valid algorithm, there exists an instance with the given variances such that the expected number of samples of the algorithm is at least*

$$c\left(\sum_{i\in[n]}\frac{\sigma_i^2}{\epsilon^2}\ln\frac{1}{\delta}+\sum_{i\in G^m}\frac{\sigma_i^2}{\epsilon^2}\ln(m)+\sum_{j\in G^l}\frac{\sigma_j^2}{\epsilon^2}\mathrm{Ent}(\sigma_{G^r}^2)\right). \tag{3.16}$$

3.4.1 Dual Formulation of the Lower Bound

We first introduce an inequality in the lemma below, which helps us connect the sample complexity with a multi-hypothesis testing problem.

Lemma 9. *For any two probability measure P, Q on the same measurable space $(\Omega, \mathcal{F})$, if $\mathcal{E} \in \mathcal{F}$ with $P(\mathcal{E}) \geq 1 - \delta > Q(\mathcal{E})$, we have*

$$Q(\mathcal{E}) \geq B(\delta) e^{-\frac{D(P||Q)}{1-\delta}}, \tag{3.17}$$

where $D(\cdot||\cdot)$ is the Kullback-Leibler divergence and $B(\delta) = e^{-\frac{\mathrm{Ent}(\delta, 1-\delta)}{1-\delta}}$ is a strictly decreasing function with $B(0.1) > 0.69$.

Fix any algorithm A with inputs $(\epsilon, \delta, m, [n], \sigma_{1:n}^2)$ that solves the (ϵ, δ) top-m arm identification problem. Consider the Gaussian instances where the i-th arm has a Gaussian distribution with variance σ_i^2. Denote P_I as the probability measure induced by the learning process of applying algorithm A on Gaussian bandit instance $I \in \mathcal{I}(\sigma_{1:n}^2)$.

Let $\epsilon' > \epsilon$ be some parameter that can be arbitrarily close to ϵ. For any subset $M \subset [n]$ with $|M| = m$ and any index $l \in [n] \setminus M$, we first construct an instance $I_{l,M} \in \mathcal{I}(\sigma_{1:n}^2)$ by specifying the reward means of each arm as follows: the l-th arm has mean 0, the arms in M have mean ϵ', and the rest have mean $-\epsilon'$. The only ϵ-approximate top-m arms of instance $I_{l,M}$ are clearly M. Similarly, for each subset $F \subset [n]$ with $|F| = m - 1$ and any index $l \in [n] \setminus F$, we then construct an instance $I_{l,F} \in \mathcal{I}(\sigma_{1:n}^2)$. In instance $I_{l,F}$, the l-th arm has mean 0, the arms in F have mean ϵ', and the rest arms have mean $-\epsilon'$. The only ϵ-approximate top-m arm set of instance $I_{l,F}$ is clearly $F \cup \{l\}$. These are the possible hypotheses we will consider.

Given an instance $I_{l,M}$, if $F = M \setminus \{i\}$ for some $i \in M$, it is clear that instances $I_{l,M}$ and $I_{l,F}$ differ only at the i-th arm. Denote $t_{l,F,i}$ as the expected number of pulls of the i-th arm by algorithm A on instance $I_{l,F}$. The KL-divergence can be calculated as $D(P_{I_{l,F}}||P_{I_{l,M}}) = \frac{2\epsilon'^2}{\sigma_i^2} t_{l,F,i}$; see Lemma 5.1 in [37] for more details. Since A solves the (ϵ, δ) top-m arm identification problem, we have $P_{I_{l,F}}(R^{\mathsf{A}} = F \cup \{l\}) \geq 1 - \delta > \delta \geq P_{I_{l,M}}(R^{\mathsf{A}} = F \cup \{l\})$. Applying Lemma 9 on $P_{I_{I,F}}$,

$P_{I_{l,M}}$ and event $\{R^{\mathsf{A}} = F \cup \{l\}\}$ gives

$$P_{I_{l,M}}(R^{\mathsf{A}} = F \cup \{l\}) \geq B(\delta) e^{-\frac{D(P_{I_{l,F}} || P_{I_{l,M}})}{1-\delta}} = B(\delta) e^{-\frac{2\epsilon'^2}{\sigma_i^2}\frac{t_{l,F,i}}{1-\delta}}. \tag{3.18}$$

This inequality holds for any $F = M \setminus \{i\}$ with $i \in M$. In addition, events $\{R^{\mathsf{A}} = M \cup \{l\} \setminus \{i\}\}$'s are disjoint for any $i \in M \cup \{l\}$, and they are also disjoint with the event $\{R^{\mathsf{A}} = M\}$. It follows that $\sum_{i \in M} P_{I_{l,M}} \left(R^{\mathsf{A}} = M \cup \{l\} \setminus \{i\}\right) \leq 1 - P_{I_{l,M}}(R^{\mathsf{A}} = M) \leq \delta$. Summing inequality (3.18) for all $i \in M$ gives

$$\delta \geq \sum_{i \in M} P_{I_{l,M}}(R^{\mathsf{A}} = M \cup \{l\} \setminus \{i\}) \geq \sum_{i \in M} B(\delta) \exp\left(-\frac{2\epsilon'^2}{\sigma_i^2}\frac{t_{l,M\setminus\{i\},i}}{1-\delta}\right). \tag{3.19}$$

In the worst-case, algorithm A takes at least $\max_{F, l \notin F} \sum_{j \notin F \cup \{l\}} t_{l,F,j}$ samples in expectation. Any valid algorithm has to satisfy (3.19), and thus the sample complexity $\mathrm{SC}(\epsilon, \delta, m, [n], \sigma_{1:n}^2)$ is lower bounded by the optimal value of the following optimization problem:

$$\begin{aligned} \text{minimize:} \quad & \max_{F \subset [n]:|F|=m-1,\ l \notin F} \sum_{j \notin F \cup \{l\}} t_{l,F,j} && (3.20) \\ \text{subject to:} \quad & \sum_{i \in M} \exp\left(-t_{l,M\setminus\{i\},i}/\theta_i\right) \leq \delta', \\ & \forall M \subset [n], |M| = m,\ \forall l \notin M, && (3.21) \end{aligned}$$

where $\theta_i = \frac{(1-\delta)\sigma_i^2}{2\epsilon^2}, \forall i \in [n]$ and $\delta' = \frac{\delta}{B(\delta)}$. Though this problem is convex, it is difficult to solve explicitly. Therefore, we consider its (restricted) dual formulation in the following lemma.

Lemma 10. *For $\epsilon > 0$, $\delta < 0.25$, $m < n/2$, $(\sigma_i^2)_{i \in [n]}$, $\mathrm{SC}(\epsilon, \delta, m, [n], \sigma_{1:n}^2) \geq \frac{1-\delta}{2\epsilon^2} v^*$, where v^* is the optimal value of the following optimization problem:*

$$\begin{aligned} \textit{maximize:} \quad & \sum_{M \subset [n]:|M|=m} \left(\sum_{l \in M} \eta_{M\setminus\{l\}}\sigma_l^2\right) \times \left(\ln\frac{B(\delta)}{\delta} + \mathrm{Ent}(\{\eta_{M\setminus\{l\}}\sigma_l^2\}_{l \in M})\right) && (3.22) \\ \textit{subject to:} \quad & \sum_{F \subset [n]:|F|=m-1} \eta_F = 1, \end{aligned}$$

$$\eta_F \geq 0,\ \forall F \subset [n], |F| = m - 1. \tag{3.23}$$

Though the dual formulation is still difficult to solve, by the weak duality, we can derive lower bounds for the primal problem by assigning specific feasible values to the dual variables η_F's. In addition, each η_F is a probability mass function and has a clear operational meaning, which is the worst-case prior distribution of the underlining instance being one of $\{I_{l,F}\}_{l \notin F}$.

3.4.2 Dichotomy of the lower bound

As shown in Theorem 12, the lower bound of the sample complexity consists of three terms

$$\underbrace{\sum_{i\in[n]} \frac{\sigma_i^2}{\epsilon^2} \ln \frac{1}{\delta}}_{\text{I}} + \underbrace{\sum_{i\in G^m} \frac{\sigma_i^2}{\epsilon^2} \ln(m)}_{\text{II}} + \underbrace{\sum_{j\in G^l} \frac{\sigma_j^2}{\epsilon^2} \mathrm{Ent}(\sigma_{G^r}^2)}_{\text{III}}. \tag{3.24}$$

We will discuss each term from the viewpoint of the dual formulation in Lemma 10. The optimal value v^* of the optimization in Lemma 10 can be lower bounded by the average of the objective function values v_1, v_2, v_3 when assigning the variables certain feasible values in the dual optimization problem, i.e., $v^* = \Omega(v_1 + v_2 + v_3)$. We construct three sets of feasible dual variables η_F's, and the resultant values $v_{1:3}$ will induce Term I-III, respectively.

It is straightforward to see that Term I can be obtained by assigning η_F's uniformly, and thus we can focus on Term II and Term III. More precisely, we aim to lower bound the optimal value of the following optimization problem:

$$\begin{aligned} \text{maximize:} \quad & \sum_{M\subset[n]:|M|=m} \left(\sum_{l\in M} \eta_{M\setminus\{l\}} \sigma_l^2 \right) \times \mathrm{Ent}(\{\eta_{M\setminus\{l\}}\sigma_l^2\}_{l\in M}) && (3.25) \\ \text{subject to:} \quad & \sum_{F\subset[n]:|F|=m-1} \eta_F = 1, \\ & \eta_F \geq 0,\ \forall F \subset [n], |F| = m - 1. && (3.26) \end{aligned}$$

Firstly, to study the sample complexity induced by $\sigma_{G^m}^2$, we specify a feasible assignment of dual variables η_F's as follows. For any $F \subset G^m$ with $|F| = m-1$, let $\eta_F = \frac{\prod_{i\in F}\sigma_i^2}{\sum_{F'\subset G^m:|F'|=m-1}\prod_{j\in F}\sigma_i^2}$; and for any $F \not\subset G^m$ with $|F| = m-1$, set $\eta_F = 0$. Then $\mathrm{Ent}(\{\eta_{M\setminus\{l\}}\sigma_l^2\}_{l\in M}) = \ln(m)$ for any $M \subset G^m$ with $|M| = m$. Formally, Term II is introduced by the following lemma.

Lemma 11. *The optimal value of the optimization (3.25) is lower-bounded by* $\frac{1}{3}\sum_{j\in G^m}\sigma_j^2\ln(m)$.

Secondly, to study the complexity induced by $\sigma_{G^l}^2$, we consider the reduced arm set $G^r \supset G^l$. Define $L \subset G^r$ with $|L| = 2m$ as the arms with $2m$ largest variances in G^r. We can verify that $\sum_{i\in L}\sigma_i^2$ dominates $\sum_{j\in G^r}\sigma_j^2$. Moreover, $\mathrm{Ent}(\sigma_{G^r}^2)$ and $\mathrm{Ent}(\sigma_L^2)$ behave similarly, and thus we can focus on the arms in L. Rigorously, the following lemma justifies this choice.

Lemma 12. *Let* $\eta_F = \binom{2m}{m-1}^{-1}$ *for any* $F \subset L$ *with* $|F| = m-1$ *and* $\eta_F = 0$ *otherwise. The objective function of the optimization problem (3.25) is at least* $c'\sum_{i\in G^l}\sigma_i^2\mathrm{Ent}(\sigma_{G^r}^2) - \ln(2)\sum_{i\in L}\sigma_i^2$, *for some constant* $c' > 0$.

The first item in Lemma 12 is exactly Term III, and the second item $-\ln(2)\sum_{i\in G^l}\sigma_i^2$ can be absorbed into Term I.

3.5 Conclusion

We studied the worst-case sample complexity of (ϵ, δ) top-m arm identification problem with known heterogeneous reward variances. The heterogeneity of reward variances is measured by a certain entropy-like function. We propose the variance-grouped median elimination algorithm, which combines the advantages of the median elimination algorithm and the weighted naive elimination algorithm in a divide-and-conquer manner. Matching the lower bound of the worst-case sample complexity was devised using a dual formulation and finding suitable feasible solutions.

A natural direction is to study the problem without the knowledge of variance proxies. However, estimating a good (i.e., a valid sub-Gaussian coefficient as small as possible) variance proxy for general sub-Gaussian distributions is still an open problem. It is due to that any coefficient in the feasible set $\{\sigma^2 : \ln\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}] \le \frac{\sigma^2\lambda^2}{2},\ \forall\lambda \in \mathbb{R}\}$ is a valid variance-proxy for the sub-Gaussian random variable X. Ideally, we would like to estimate the best variance proxy, i.e., the minimum value in the feasible set. However, the variance $\mathbb{E}[(X - \mathbb{E}[X])^2]$ of the random variable is not a valid variance proxy, and canonical variance estimators would not suffice. The study of a good estimator for the variance proxy is beyond the scope of the draft.

4. KL-REGULARIZED POLICY-GRADIENT FOR MULTI-OBJECTIVE REINFORCEMENT LEARNING

In addition to bounding performance metrics of interest as shown in Chapter 2 and characterizing the complexity of the problems as shown in Chapter 3, information measures can also facilitate the design of algorithms. In this chapter, we consider policy optimization in *single-policy* multi-objective MDPs, where the agent aims to find a policy satisfying certain criteria by policy gradient-based algorithms. We propose an Anchor-changing Regularized Natural Policy Gradient (ARNPG) framework, which introduces Kullback-Leibler divergences with changing anchors as regularization.

4.1 Preliminaries

System model A Markov decision process (MDP) is represented by a tuple $(\mathcal{S}, \mathcal{A}, P, \rho, \gamma, r)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ the action space, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ the transition kernel, $\rho \in \Delta(\mathcal{S})$ the initial state distribution, $\gamma \in (0, 1)$ the discount factor, and $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ the reward function. Given any policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ and any reward function $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$, we define the state value function $V_r^\pi : \mathcal{S} \to [0, \frac{1}{1-\gamma}]$, and the state-action value function $Q_r^\pi : \mathcal{S} \times \mathcal{A} \to [0, \frac{1}{1-\gamma}]$, as

$$V_r^\pi(s) := \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s, \pi], \quad Q_r^\pi(s, a) := \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s, a_0 = a, \pi],$$

where expectation $\mathbb{E}$ is taken over the random trajectory of the Markov chain induced by the policy π and the transition kernel P. With a slight abuse of notation, we denote $V_r^\pi(\rho) := \mathbb{E}_{s\sim\rho}[V_r^\pi(s)]$. Define the discounted state-action visitation distribution (state-action visitation for short) of policy π with initial state distribution ρ by $d_\rho^\pi(s, a) := (1 - \gamma)\mathbb{E}_{s_0\sim\rho}[\sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = s, a_t = a|s_0, \pi)]$. It

[*] Part of this chapter is reprinted with permission from "Anchor-changing regularized natural policy gradient for multi-objective reinforcement learning," by Ruida Zhou, Tao Liu, Dileep Kalathil, P. R. Kumar, and Chao Tian, 2022, *Advances in Neural Information Processing Systems*, vol. 35, pp. 13584–13596, Copyright © Ruida Zhou, Tao Liu, Dileep Kalathil, P. R. Kumar, and Chao Tian [68].

then follows that $V_r^\pi(\rho) = \frac{1}{1-\gamma}\langle d_\rho^\pi, r\rangle$ by viewing d_ρ^π and r as $|\mathcal{S}||\mathcal{A}|$-dimensional vectors indexed by $(s,a) \in \mathcal{S} \times \mathcal{A}$. When it is clear from the context, we denote the state visitation distribution by $d_\rho^\pi(s) := \mathbb{E}_{s_0\sim\rho}\left[(1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}(s_t = s|s_0)\right]$, which is the marginal distribution of the state-action visitation $d_\rho^\pi(s,a)$, i.e., $d_\rho^\pi(s) = \sum_{a\in\mathcal{A}} d_\rho^\pi(s,a)$.

We study an MDP with m objectives represented by $(\mathcal{S}, \mathcal{A}, P, \rho, \gamma, r_{1:m})$, where $r_i : \mathcal{S} \times \mathcal{A} \to [0,1]$ is the i-th reward function for each $i \in [m]$. For simplicity, denote $V_i^\pi(\cdot) := V_{r_i}^\pi(\cdot)$ and $V_{1:m}^\pi(\cdot) := (V_1^\pi(\cdot), \ldots, V_m^\pi(\cdot))$. We consider parameterized policies in $\Pi = \{\pi_\theta : \theta \in \Theta\}$, where $\Theta \subset \mathbb{R}^n$ is the parameter space. For example, the softmax policy is $\pi_\theta(a|s) = \frac{\exp(\theta_{s,a})}{\sum_{a'}\exp(\theta_{s,a'})}$ with $\Theta = \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$; and neural softmax policy is $\pi_\theta(a|s) = \frac{\exp(\mathrm{NN}_\theta(s,a))}{\sum_{a'}\exp(\mathrm{NN}_\theta(s,a'))}$, where NN_θ is some neural network parameterized θ. Define $\mathcal{V} := \{V_{1:m}^{\pi_\theta}(\rho) : \theta \in \Theta\}$ as the achievable region of value vectors. The agent wishes to optimize the policy in Π for a given specific multi-objective criterion on value vectors in $\mathcal{V}$. For example,

1. Proportional fairness [69]: Given $a_{1:m} > 0$, find $v \in \mathcal{V}$ that $\sum_{i=1}^m a_i \frac{v_i' - v_i}{v_i} \le 0,\ \forall v' \in \mathcal{V}$.

2. Hard constraints [70]: Given $b_{2:m}$, maximize$_{v\in\mathcal{V}}$ v_1, subject to $v_i \ge b_i, \forall i = 2, \ldots, m$.

3. Max-min trade-off [71]: Given $c_{1:m} > 0$, maximize$_{v\in\mathcal{V}}$ $\min_{i\in[m]}(v_i/c_i)$.

Mirror ascent As one of the most well-known iterative optimization methods, mirror descent (actually ascent in the context of our formulation as a maximization problem) [72, 73] is a general class that encompasses many first-order methods in convex optimization. Given a variable x in a compact convex set $\mathcal{X} \subset \mathbb{R}^n$ and an ascent direction $g \in \mathbb{R}^n$, the variational representation of the mirror ascent update is

$$x' \in \arg\max_{y\in\mathcal{X}}\{\langle g, y\rangle - \alpha B_h(y||x)\}, \tag{4.1}$$

where $B_h(x||y) := h(x) - h(y) - \langle \nabla h(y), x - y\rangle$ is some Bregman divergence generated by a differentiable convex function $h : \mathcal{X} \to \mathbb{R}$. When analyzing the convergence of first-order methods, certain fundamental inequalities are usually established to facilitate the proof. One such inequality is

$$\langle g, x' \rangle - \alpha B_h(x'||x) \geq \langle g, y \rangle - \alpha B_h(y||x) + \alpha B_h(y||x'), \quad \forall y \in \mathcal{X}, \tag{4.2}$$

which is a critical step in many previous works, e.g., [74–76].

It is desirable to construct a similar fundamental inequality for multi-objective MDPs that can facilitate the analysis of convergence. As we will show in the next section, such an inequality can indeed be established in a new framework, which we refer to as the Anchor-Changing Regularized Natural Policy Gradient (ARNPG).

Denote KL-divergence between two n-dimensional probability vectors x, y by $D(x||y) := \sum_{i=1}^{n} x_i \log(x_i/y_i)$, which is a widely-used Bregman divergence. For any policies π, π' and state visitation distribution d, define $D_d(\pi||\pi') := \sum_{s\in\mathcal{S}} d(s) D(\pi(\cdot|s)||\pi'(\cdot|s))$. A *uniform policy* is one which chooses actions uniformly at random.

4.2 Anchor-changing Regularized Natural Policy Gradient

Let us consider a hypothetical mirror ascent update on decision value vector $v_k \in \mathcal{V}$ according to (4.1). Given an ascent direction $\tilde{G}_k$ along which to improve v_k, the updated value vector is

$$v' \in \arg\max_{v\in\mathcal{V}} \{ \langle \tilde{G}_k, v \rangle - \alpha B_h(v||v_k) \}. \tag{4.3}$$

Suppose the value vector v_k is achieved by a policy π_{θ_k}, i.e., $v_k = V_{1:m}^{\pi_{\theta_k}}(\rho)$. Denote the reward function in the ascent direction as $\tilde{r}_k(s,a) = \langle \tilde{G}_k, r_{1:m}(s,a) \rangle$. It follows that $\langle \tilde{G}_k, v_k \rangle = V_{\tilde{r}_k}^{\pi_{\theta_k}}(\rho)$. Note that $B_h(v||v_k)$ in (4.3) serves the role of a soft constraint on v by keeping v within a vicinity of v_k. Replacing $B(v||v_k)$ by $\frac{D_{d_\rho^{\pi_\theta}}(\pi_\theta||\pi_{\theta_k})}{1-\gamma}$ will induce a similar soft constraint that prefers the vicinity of the "anchor" policy π_{θ_k}. Therefore we consider replacing the variational update in (4.3) by

$$\theta' \in \arg\max_{\theta\in\Theta} \left\{ \tilde{V}_{k,\alpha}^{\pi_\theta}(\rho) \right\}, \quad \text{where} \quad \tilde{V}_{k,\alpha}^{\pi_\theta}(\rho) := V_{\tilde{r}_k}^{\pi_\theta}(\rho) - \alpha \frac{D_{d_\rho^{\pi_\theta}}(\pi_\theta||\pi_{\theta_k})}{1-\gamma}. \tag{4.4}$$

ARNPG Motivated by the intuition above, we propose the Anchor-Changing Regularized Natural Policy Gradient (ARNPG) framework. At (macro) step k, the ARNPG framework determines the reward function in the ascent direction $\tilde{r}_k$ and the anchor policy π_{θ_k}, which can exploit well-performed first-order methods in convex optimization literature utilizing the features of the specific criteria in use. With $\tilde{r}_k$ and π_{θ_k}, we wish to solve for (4.4) to improve the value vector. However the optimal solution θ' of (4.4) is generally not determinable explicitly. ARNPG therefore approaches the optimal solution via a subroutine that executes a natural policy gradient (NPG) algorithm w.r.t. the KL-regularized value function $\tilde{V}_{k,\alpha}^{\pi_\theta}(\rho)$. We refer to this subroutine, given in Algorithm 3, as InnerLoop($\tilde{r}_k, \pi_{\theta_k}, \alpha, \eta, t_k$). It iteratively updates the parameter $\theta_k^{(t)}$ for t_k (micro) steps according to the NPG update rule as in (4.5), where $\mathcal{F}_\rho(\theta)^\dagger$ is the Moore-Penrose inverse of the Fisher information matrix $\mathcal{F}_\rho(\theta) := \mathbb{E}_{(s,a)\sim d_\rho^{\pi_\theta}} \left[\nabla_\theta \log \pi_\theta(a|s) \left(\nabla_\theta \log \pi_\theta(a|s)\right)^\top\right]$.

```
Algorithm 3: InnerLoop($\tilde{r}_k, \pi_{\theta_k}, \alpha, \eta, t_k$)
Initialize $\theta_k^{(0)} = \theta_k$
for $t = 0, 1, \dots t_k - 1$ do
    $\theta_k^{(t+1)} \leftarrow \theta_k^{(t)} + \eta \mathcal{F}_\rho(\theta_k^{(t)})^\dagger \nabla_\theta \tilde{V}_{k,\alpha}^{\pi_k^{(t)}}(\rho)$        (4.5)
Return $\theta_k^{(t_k)}$
```

The choice of the number of iterations in InnerLoop (i.e., t_k) involves a trade-off between the variational update precision and the overall efficiency. On the one hand, a larger t_k leads to a more accurate approximation of the optimal solution θ' to (4.4), but it may cause the algorithm to spend unnecessary computational resources on the regularized objective $\tilde{V}_{k,\alpha}^{\pi_\theta}(\rho)$, instead of on the true optimization problem. On the other hand, a smaller t_k saves inner loop iterations but the update follows less closely to the underlying mirror-ascent update in improving the value vector. In our experiments, we choose t_k within 10 to strike a balance and empirically observe $t_k > 1$ has better performance.

We note that when $t_k = 1$, the gradient $\nabla_\theta \tilde{V}_{k,\alpha}^{\pi_{\theta_k}}(\rho) = \nabla_\theta V_{\tilde{r}_k}^{\pi_{\theta_k}}(\rho)$, since $D_{d_\rho^{\pi_\theta}}(\pi_\theta||\pi_{\theta_k})$ has zero gradient at $\theta = \theta_k$. The update in (4.5) reduces to an NPG update on the unregularized value function $\tilde{V}_{\tilde{r}_k}^{\pi_\theta}(\rho)$. For single-objective MDPs, it reduces to the canonical NPG method.

4.2.1 Theoretical guarantee of ARNPG

We now present the main theoretical tool for the analysis of the ARNPG framework. Recall the discussion of the fundamental inequality after (4.2). Proposition 4 establishes such a fundamental inequality with controllable approximation error under the softmax policy parameterization, i.e., $\pi_\theta(a|s) = \frac{\exp(\theta_{s,a})}{\sum_{a'} \exp(\theta_{s,a'})}$. We will omit θ in π_θ when it is clear from the context, but it should be noted that all updates of policies are performed on the parameters.

Proposition 4. *Under the softmax parameterization, given $\epsilon_k > 0$, for any $\tilde{r}_k$, $t_k \geq \frac{1}{1-\gamma} \log(\frac{5\|\tilde{r}_k\|_\infty}{(1-\gamma)^2 \epsilon_k}) +$ 1, $\alpha > 0$ and $\eta = \frac{1-\gamma}{\alpha}$, the update $\pi_{k+1} \leftarrow$ InnerLoop$(\pi_k, \tilde{r}_k, \alpha, \eta, t_k)$ satisfies*

$$V_{\tilde{r}_k}^{\pi_{k+1}}(\rho) - \alpha \frac{D_{d_\rho^{\pi_{k+1}}}(\pi_{k+1}||\pi_k)}{1-\gamma} \geq V_{\tilde{r}_k}^{\pi}(\rho) - \alpha \frac{D_{d_\rho^\pi}(\pi||\pi_k) - D_{d_\rho^\pi}(\pi||\pi_{k+1})}{1-\gamma} - \epsilon_k, \quad \forall \pi. \tag{4.6}$$

The inequality (4.6) is critical to the convergence proof. Its right hand side allows telescoping, which by summing over k can iteratively cancel the terms $D_{d_\rho^\pi}(\pi||\pi_k)$. Since $t_k = \Theta(\log(1/\epsilon_k))$ it suffices to use very few iterations in InnerLoop for maintaining precision.

Remark. It has been shown that for the entropy-regularized MDP, i.e., KL-regularized with the uniform policy as the anchor policy, NPG converges linearly (i.e., geometrically fast) to the regularized optimal policy [77]. It is natural to anticipate that for the KL-regularized MDP $\tilde{V}_{k,\alpha}^{\pi}(\rho)$ with anchor π_k, NPG would similarly converge linearly (i.e., $\tilde{V}_{k,\alpha}^{\pi_k} \geq \tilde{V}_{k,\alpha}^{\pi_k^*} - \epsilon$ for $t_k = \Theta(\log(1/\epsilon))$) to a corresponding optimal policy, denoted as π_k^*. In contrast, the right hand side of inequality (4.6) has a *positive drift* $\alpha \frac{D_{d_\rho^\pi}(\pi||\pi_{k+1})}{1-\gamma}$ *for any policy* π, which is considerably stronger.

Proof sketch of Proposition 4. We can show that InnerLoop approximately solves the variational update in (4.4) with linear convergence as anticipated. However to establish (4.6), the difficulty lies in the introduction of positive drift, since $V_{\tilde{r}_k}^{\pi_\theta}(\rho)$ is not concave w.r.t. θ and $D_{d_\rho^{\pi_\theta}}(\pi_\theta||\pi_{\theta_k})$ may not be a Bregman divergence. We tackle this difficulty by showing that optimizing π_θ in InnerLoop implicitly performs a mirror ascent update for state action visitation $d_\rho^{\pi_\theta}$. □

As demonstrated in the next section, Proposition 4 ensures that the convergence rate of the algorithms derived from the ARNPG framework is of the same rate as the underlying first-order methods with only extra logarithmic factors.

4.3 Theoretical Applications

In this section, we apply the ARNPG framework to several important multi-objective MDP scenarios and obtain new policy optimization algorithms by integrating first-order methods in convex optimization. All the theoretical results presented in this section are under the softmax parameterization with exact gradients. However, the obtained algorithms can be implemented in more general settings such as neural softmax and sample-based scenarios, as in the next section. We theoretically establish $\tilde{O}(1/T)$ convergence of these algorithms by leveraging the fundamental inequality in Proposition 4.

4.3.1 Smooth concave scalarization function

We start by considering the following optimization problem

$$\max_{\theta} F(V_{1:m}^{\pi_\theta}(\rho)), \tag{4.7}$$

where F is a concave function, and β-smooth w.r.t. $\|\cdot\|_\infty$ norm, i.e., $\|\nabla F(v) - \nabla F(v')\|_1 \leq \beta\|v - v'\|_\infty$. Since the set of achievable values $\mathcal{V} \subseteq \left[0, \frac{1}{1-\gamma}\right]^m$, it can be verified that $\|\nabla F(v)\|_1 \leq L$ for some factor $L > 0$.

The proportional fair criterion can be approximated by $F(v) := \sum_{i=1}^m a_i \log(\delta + v_i)$, where $\delta > 0$ is some constant introduced to circumvent the pathological case $v_i = 0$ for some $i \in [m]$. Under this criterion, $\beta = \sum_{i=1}^m a_i/\delta^2$ and $L = \sum_{i=1}^m a_i/\delta$.

When v is viewed as the decision variable, at macro step k with value vector $V_{1:m}^{\pi_k}(\rho)$, the ascent direction in a typical gradient ascent step is the gradient $\tilde{G}_k = \nabla_v F(V_{1:m}^{\pi_k}(\rho))$. This naturally determines the reward in the ascent direction as $\tilde{r}_k(s,a) = \langle \tilde{G}_k, r_{1:m}(s,a) \rangle$. Adapting the ARNPG framework to this specific context, we present the algorithm for solving the program (4.7) in Algorithm 4. We refer to it as “implicit mirror descent" because the algorithm implicitly employs mirror descent.

Algorithm 4: ARNPG Implicit Mirror Descent (ARNPG-IMD)

```
Input $\pi_0, \alpha, \eta, t_{0:K-1}, K$
for $k = 0, 1, \ldots, K-1$ do
    Update $\pi_{k+1} \leftarrow$InnerLoop($\pi_k, \tilde{r}_k, \alpha, \eta, t_k$)
Return the policy in $\{\pi_k\}_{k=1}^{K}$ with the largest $F(V_{1:m}^{\pi_k}(\rho))$
```

Let π^* be the optimal policy for (4.7). Based on Proposition 4, we present the following theorem which guarantees the convergence of ARNPG-IMD with appropriately selected parameters π_0, α, η, t_k.

Theorem 13. *For any $K \geq 1$, take uniform policy π_0, $\alpha \geq \frac{\beta}{(1-\gamma)^3}$, $\eta = \frac{1-\gamma}{\alpha}$, and $t_k = \lceil \frac{1}{1-\gamma} \log(\frac{5LK}{\beta \log(|\mathcal{A}|)}) + 1 \rceil$. The optimality gap of ARNPG-IMD (Algorithm 4) satisfies*

$$F(V_{1:m}^{\pi^*}(\rho)) - \max_{k \in [1:K]} F(V_{1:m}^{\pi_k}(\rho)) \leq F(V_{1:m}^{\pi^*}(\rho)) - \frac{1}{K} \sum_{k=1}^{K} F(V_{1:m}^{\pi_k}(\rho)) \leq \frac{2\alpha \log(|\mathcal{A}|)}{(1-\gamma)K}. \qquad (4.8)$$

There are a total of K macro steps, and the total number of iterations is $T = \sum_{k=0}^{K-1} t_k = \Theta(\frac{K}{1-\gamma} \log(K))$. The following corollary provides the convergence rate in terms of T.

Corollary 6. *Under the same conditions as in Theorem 13, the ARNPG-IMD algorithm satisfies $F(V_{1:m}^{\pi^*}(\rho)) - \frac{1}{K} \sum_{k=1}^{K} F(V_{1:m}^{\pi_k}(\rho)) = O\left(\frac{\beta \log(T)}{(1-\gamma)^5 T}\right)$.*

Remark. In the absence of knowledge of K, we can select time-varying numbers of InnerLoop iterations, such as $t_k = \Theta(\log(k))$, and ARNPG-IMD will still have the same $\tilde{O}(1/T)$ convergence.

4.3.2 Constrained Markov decision process

Another way of trading off the objectives is to optimize one while setting hard constraints on the others. This can be formulated as the following constrained MDP (CMDP) problem:

$$\max_{\theta} V_1^{\pi_\theta}(\rho), \quad \text{s.t. } V_i^{\pi_\theta}(\rho) \geq b_i, \ \forall i \in [2:m], \qquad (4.9)$$

where $b_{2:m} \in [0, \frac{1}{1-\gamma}]^{m-1}$. Let $\pi^* = \pi_{\theta^*}$ be the optimal policy of the CMDP problem in (4.9).

Define the Lagrangian of the CMDP problem as $\mathcal{L}(\pi_\theta, \lambda) = V_1^{\pi_\theta}(\rho) + \sum_{i=2}^m \lambda_i (V_i^{\pi_\theta}(\rho) - b_i)$, where λ_i is the Lagrange multiplier (dual variable) corresponding to the constraint $V_i^{\pi_\theta} \geq b_i$, for each $i \in [2:m]$. The Lagrange dual function $\max_\pi \mathcal{L}(\pi, \cdot)$ is a convex function of dual variables $\lambda \geq 0$. Denote by λ^* the optimal dual variables that minimize the Lagrange dual function. We assume λ^* is finite, which is guaranteed by Slater's condition, i.e., there is some π_θ and $\xi > 0$ with $V_i^{\pi_\theta}(\rho) - b_i \geq \xi$ for any $i \in [2:m]$. Note (π^*, λ^*) is a saddle point of the Lagrangian $\mathcal{L}(\pi, \lambda)$. This motivates the primal-dual approach, which iteratively performs gradient ascent for π_θ and gradient descent for λ. This is suitable for the CMDP setting, since for any fixed λ, the Lagrangian $\mathcal{L}(\pi, \lambda)$ corresponds to an MDP for which policy gradient can be employed.

The canonical primal-dual gradient ascent-descent method for constrained convex optimization can only guarantee $O(1/\sqrt{T})$ convergence, and consequently the primal-dual policy gradient-based approach for CMDPs [15] has the same convergence. Recently, Yu et al. [78] have proposed a primal-dual-based method with $O(1/T)$ convergence under the Euclidean setting, i.e., $B_h(x||y) = \frac{1}{2}||x - y||_2^2$. Adopting ideas from [78], we next propose the ARNPG with Extra Primal-Dual (ARNPG-EPD) algorithm (Algorithm 5). To the best of our knowledge, this new primal-dual update appears in the CMDP-related literature for the first time.

Note that $b_i - V_i^\pi(\rho)$ is the amount of constraint violation. There are two key ideas we adopt from [78]. The first is the design of the reward in the ascent direction

$$\tilde{r}_k(s,a) := r_1(s,a) + \textstyle\sum_{i=2}^m (\lambda_{k,i} + \eta'(b_i - V_i^{\pi_k}(\rho))) r_i(s,a),$$

where an extra constraint violation term is added to the dual variables. The second idea is that the update of dual variables should not fall below the negative constraint violation (the first term in (4.10)), and it can alleviate the overshooting of dual variables. The extra constraint violation terms in $\tilde{r}_k$ and the dual update work jointly to ensure the $\tilde{O}(1/T)$ convergence.

Theorem 14. *For any $K \geq 1$ and $\eta' \in (0,1]$, take uniform policy π_0, $\alpha \geq \frac{2\eta' m}{(1-\gamma)^3}$, $\eta = \frac{1-\gamma}{\alpha}$, and choose $t_k = \lceil \frac{1}{1-\gamma} \log(\frac{5 L_k K}{2\eta' m \log(|\mathcal{A}|)}) + 1 \rceil$ with $L_k = 1 + \frac{\eta'(m-1)}{1-\gamma} + \sum_{i=2}^m \lambda_{k,i}$. The average optimality*

Algorithm 5: ARNPG with Extra Primal Dual (ARNPG-EPD)

```
Input $\pi_0, \eta', \alpha, \eta, t_{0:K-1}, K$
Initialize $\lambda_{0,i} = \max\{\eta'(V_i^{\pi_0}(\rho) - b_i), 0\},\ \forall i \in [2:m]$
for $k = 0, 1, \dots, K-1$ do
    Update $\pi_{k+1} \leftarrow$InnerLoop$(\pi_k, \tilde{r}_k, \alpha, \eta, t_k)$
    Update $\lambda_{k+1,i} = \max\left\{\eta'(V_i^{\pi_{k+1}}(\rho) - b_i), \lambda_{k,i} + \eta'(b_i - V_i^{\pi_{k+1}}(\rho))\right\},\ \forall i \in [2:m]$   (4.10)
Return: a policy randomly chosen from $\{\pi_k\}_{k=1}^K$
```

gap and the average constraint violation of ARNPG-EPD (Algorithm 5) satisfy

$$V_1^{\pi^*}(\rho) - \frac{1}{K}\sum_{k=1}^{K} V_1^{\pi_k}(\rho) \le \frac{3\alpha\log(|\mathcal{A}|)}{(1-\gamma)K}, \tag{4.11}$$

$$b_i - \frac{1}{K}\sum_{k=1}^{K} V_i^{\pi_k}(\rho) \le \frac{1}{K}\left(\frac{2\|\lambda^*\|_2}{\eta'} + 3\sqrt{\frac{\alpha\log(|\mathcal{A}|)}{(1-\gamma)\eta'}}\right) \quad \forall i \in [2:m]. \tag{4.12}$$

Note that the number of micro steps t_k is chosen according to the dual variables λ_k in the previous theorem. Denote by $T := \sum_{k=0}^{K-1} t_k$ the total number of iterations.

Corollary 7. *Under the same conditions as in Theorem 14, the ARNPG-EPD algorithm satisfies* $V_1^{\pi^*}(\rho) - \frac{1}{K}\sum_{k=1}^{K} V_1^{\pi_k}(\rho) = O(\frac{m\log(T)}{(1-\gamma)^5 T})$*, and* $b_i - \frac{1}{K}\sum_{k=1}^{K} V_i^{\pi_k}(\rho) = O(\frac{\sqrt{m}\log(T)}{(1-\gamma)^{2.5} T})$.

The theorem and corollary establish convergence of the average optimality gap and the average constraint violation, in the same manner as many previous works [15, 16, 79, 80] on CMDPs. However, a guarantee on the last iterate is more preferable. This drawback is inherited from the primal-dual algorithm for convex optimization, where the primal-dual algorithm with sublinear convergence can only be guaranteed on the average solution, as of our knowledge. Last iterate convergence is still an on-going open research topic.

4.3.3 Max-min trade-off criteria

Finally, we consider the max-min trade-off criterion defined as

$$\max_{\theta} \min_{\lambda\in\Lambda} \Phi(V_{1:m}^{\pi_\theta}(\rho), \lambda), \tag{4.13}$$

where Λ is a subset of the m-dimensional probability simplex $\Delta([m])$. We assume $\Phi(\cdot, \lambda)$ is concave and $\Phi(v, \cdot)$ is convex. We also assume Φ is β-smooth w.r.t. the norm $\Psi(v, \lambda) = \|v\|_\infty + \|\lambda\|_1$. The max-min criterion can be represented by $\Phi(v, \lambda) = \sum_{i=1}^m v_i\lambda_i/c_i$ and $\Lambda = \Delta([m])$. Φ satisfies the concave-convex assumption and is β-smooth w.r.t. the norm Ψ with $\beta = O(m)$.

Denote $F(v) := \min_{\lambda\in\Lambda} \Phi(v, \lambda)$, which is concave but not necessarily smooth. Thus we cannot apply the ARNPG-IMD algorithm (Algorithm 4) due to the non-smoothness of F, and the subgradient-based method can only guarantee $O(1/\sqrt{T})$ convergence.

We next integrate the optimistic mirror descent ascent (OMDA) method [75] for solving minimax optimization in the ARNPG framework. Denote the gradients $\tilde{G}_k^\lambda = \nabla_\lambda \Phi(V_{1:m}^{\tilde{\pi}_k}(\rho), \tilde{\lambda}_k)$ and $\tilde{G}_k^v = \nabla_v \Phi(V_{1:m}^{\tilde{\pi}_k}(\rho), \tilde{\lambda}_k)$. It can be verified that $\|\tilde{G}_k^v\|_1 \le L$ for some L due to the smoothness of Φ. OMDA performs gradient ascent along the direction $\tilde{G}_k^v$ w.r.t. the value vector, and therefore we construct the reward in the ascent direction as $\tilde{r}_k(s, a) = \langle \tilde{G}_k^v, r_{1:m}(s, a)\rangle$. OMDA performs mirror descent along direction $\tilde{G}_k^\lambda$ w.r.t. the dual vector λ. A key ingredient of OMDA is that it updates twice in each macro step. ARNPG-OMD adopts this idea and update (π, λ) from the same anchor points (π_k, λ_k), first with ascent direction $(\tilde{r}_k, -\tilde{G}_k^\lambda) \in \mathbb{R}^{2m}$ and then a further step with direction $(\tilde{r}_{k+1}, -\tilde{G}_{k+1}^\lambda) \in \mathbb{R}^{2m}$.

We present ARNPG-OMDA in Algorithm 6, and establish the following performance guarantees:

Theorem 15. *For any $K \ge 1$, take uniform policy π_0, $\eta' \le \frac{1}{6\beta}$, $\alpha \ge \frac{6\beta}{(1-\gamma)^3}$, $\eta = \frac{1-\gamma}{\alpha}$, and*

Algorithm 6: ARNPG with Optimistic Mirror Descent Ascent Update (ARNPG-OMDA)

```
Input $\pi_0, \lambda_0, \eta', \alpha, \eta, t_{0:K-1}, K$
Initialize $\tilde{\pi}_0 = \pi_0$ and $\lambda_0, \tilde{\lambda}_0$ as uniform distribution on $[m]$
for $k = 0, 1, \ldots, K-1$ do
    Update $\tilde{\pi}_{k+1}$ ←InnerLoop($\pi_k, \tilde{r}_k, \alpha, \eta, t_k$), $\tilde{\lambda}_{k+1} \leftarrow \arg\min_{\lambda \in \Lambda}\{\langle \tilde{G}^{\lambda}_{k}, \lambda \rangle + \frac{D(\lambda \| \lambda_k)}{\eta'}\}$
    Update $\pi_{k+1}$ ←InnerLoop($\pi_k, \tilde{r}_{k+1}, \alpha, \eta, t_k$),
      $\lambda_{k+1} \leftarrow \arg\min_{\lambda \in \Lambda}\{\langle \tilde{G}^{\lambda}_{k+1}, \lambda \rangle + \frac{D(\lambda \| \lambda_k)}{\eta'}\}$
Return: a policy randomly chosen from $\{\tilde{\pi}_k\}_{k=1}^{K}$
```

$t_k = \lceil \frac{1}{1-\gamma} \log(\frac{5LK}{6\beta \log(|\mathcal{A}|)}) + 1 \rceil$. *The ARNPG-OMDA algorithm (Algorithm 6) satisfies*

$$F(V_{1:m}^{\pi^*}(\rho)) - F\left(\frac{1}{K}\sum_{k=1}^{K} V_{1:m}^{\tilde{\pi}_k}(\rho)\right) \leq \frac{3\alpha \log(|\mathcal{A}|)}{(1-\gamma)K} + \frac{\log(m)}{\eta' K}. \tag{4.14}$$

Similar to the discussion after Corollary 7, Theorem 15 provides a performance guarantee on the average value vector $F(\frac{1}{K}\sum_{k=1}^{K} V_{1:m}^{\tilde{\pi}_k}(\rho))$, which is inherited from the OMDA methods. Denote the total number of iterations by $T := \sum_{k=0}^{K-1} 2t_k$.

Corollary 8. *Under the same conditions as in Theorem 15, ARNPG-OMDA satisfies* $F\left(V_{1:m}^{\pi^*}(\rho)\right) - F\left(\frac{1}{K}\sum_{k=1}^{K} V_{1:m}^{\pi_k}(\rho)\right) = O\left(\frac{\beta \log(T)}{(1-\gamma)^5 T}\right)$.

4.4 Empirical Evaluation and Application

We compare the performance of the proposed ARNPG-EPD algorithm (Algorithm 5) with two benchmarks: NPG-PD [15] and CRPO [16]. The tabular CMDP, for both exact-gradient scenario (Chapter 4.4.1) and sample-based scenario (Chapter 4.4.2), follows the same experimental setting as in [15]. The MDP with $m = 2$ objectives represented by $(\mathcal{S}, \mathcal{A}, P, \rho, \gamma, r_{1:2})$ (as the system model in Chapter 4.1) is randomly generated, where $|\mathcal{S}| = 20$, $|\mathcal{A}| = 10$, ρ is uniform distribution, and $\gamma = 0.8$. For each $(s, a) \in \mathcal{S} \times \mathcal{A}$, $P(\cdot|s, a) \in \Delta(\mathcal{S})$ is generated by normalizing a random vector $\sim Unif([0, 1]^{\mathcal{S}})$, and independent rewards $r_1(s, a), r_2(s, a) \sim Unif([0, 1])$. Choosing the constraint coefficient $b_2 = 3$, the experiments are performed on the CMDP

$$\max_{\theta \in \Theta} V_{r_1}^{\pi_\theta}(\rho) \quad \text{s.t.} \ V_{r_2}^{\pi_\theta}(\rho) \geq b_2, \tag{4.15}$$

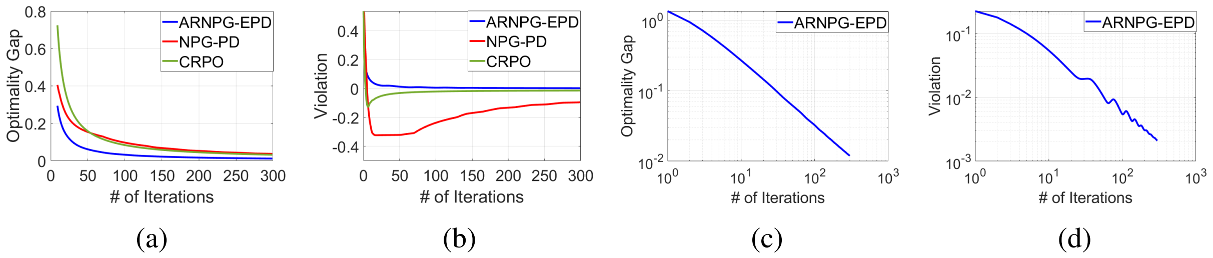
with the softmax policy class.

For both the exact-gradient scenario (Chapter 4.4.1) and the sample-based scenario (Chapter 4.4.2), we choose $\eta = 1$ and $\eta' = 1$ for ARNPG-EPD and NPG-PD (following the same hyperparameter selection as in [15]), since both rely on a primal-dual framework. Additionally, we fix $t_k = 1, \forall k = 0, 1, \dots, K-1$ and select $\alpha = \frac{1-\gamma}{\eta} = 0.2$ for ARNPG-EPD. As for CRPO with exact gradients, we first fix the tolerance parameter as 0.01 and then choose the best learning rate 0.4 from the set $\{0.1, 0.2, \dots, 0.9, 1.0\}$, which enjoys the smallest average optimality gap after 300 iterations. For sample-based CRPO, we select the best learning rate 1.0 from the set $\{0.1, 0.5, 1, 2, 5\}$, which leads to the largest reward value after 300 iterations.

4.4.1 Tabular CMDP with exact gradients

Recall that under softmax policy with exact gradients, Corollary 7 (Theorem 14) guarantees $\tilde{O}(1/T)$ convergence of both performance measures: average optimality gap and average constraint violation. We compare the proposed ARNPG-EPD with the benchmarks NPG-PD and CRPO under both performance measures on a randomly generated CMDP with a single constraint, which are illustrated in Figure 4.1. The horizontal axis is the total number of iterations, i.e., including the micro steps in InnerLoop of ARNPG-EPD.

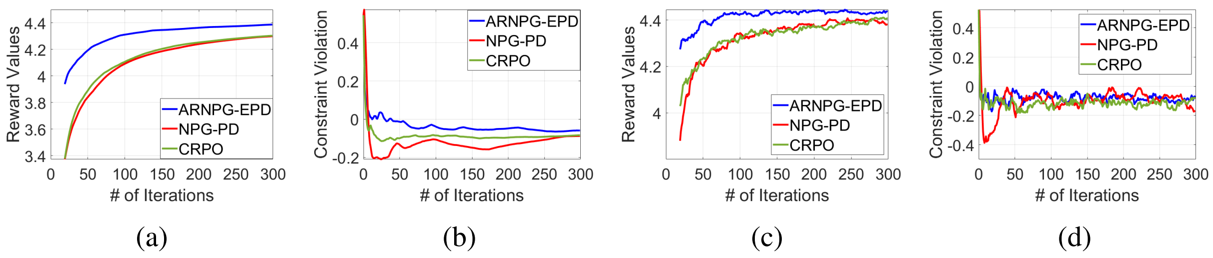
Figures 4.1(a) and 4.1(b) show that both the average optimality gap and the average constraint violation of the ARNPG-EPD algorithm converge faster than those of NPG-PD. Since the CRPO focuses on the violated constraint, the policy becomes feasible quickly, though at the cost of an initially slower convergence for the optimality gap. As illustrated in Figures 4.1(c) and 4.1(d), the slopes of both the optimality gap and the constraint violation of the ARNPG-EPD algorithm in the log-log plots are approximately between -0.9 and -1, indicating a converge rate of $\tilde{O}(1/T)$.



(a) (b) (c) (d)

Figure 4.1: The average optimality gap and the average constraint violation versus the total number of iterations, for ARNPG-EPD, NPG-PD, and CRPO on a randomly generated CMDP.

4.4.2 Sample-based tabular CMDP



(a) (b) (c) (d)

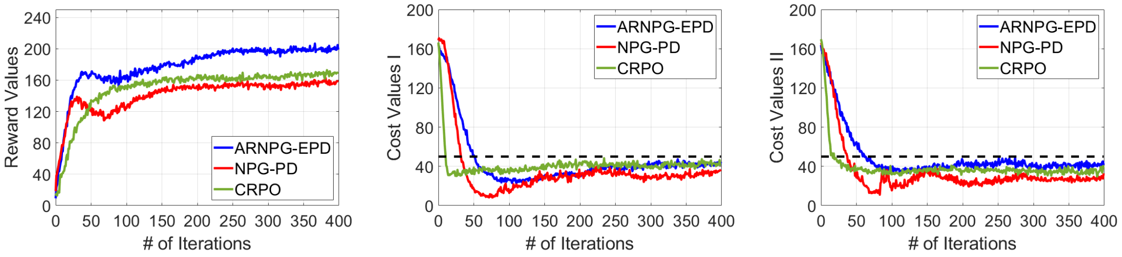
Figure 4.2: The reward values and the constraint violation with respect to the total number of iterations, for sample-based ARNPG-EPD, NPG-PD, and CRPO on a randomly generated CMDP.

We next consider the same tabular CMDP described in Chapter 4.4.1 without exact policy gradients. Instead, policy gradients are estimated by samples from a generative model that can generate independent trajectories starting from any state and action pair. The assumption of such a generative model is common [15, 16, 81].

The performances of CRPO, NPG-PD, and ARNPG-EPD in the sample-based scenario are shown in Figure 4.2. Figures 4.2(a) and 4.2(b) display the averaged performance, while Figures 4.2(c) and 4.2(d) display the performance of the current iterate (a.k.a. last-iterate in optimization literature). It shows that in this sample-based scenario, ARNPG-EPD achieves higher reward values with faster convergence, while all three algorithms satisfy the constraint after a few iterations.

4.4.3 Acrobot-v1

To demonstrate the efficacy of ARNPG-EPD on complex tasks, we have conducted experiments on the Acrobot-v1 environment from OpenAI Gym [82]. We follow the same experiment setup in [16], where there is a reward value to maximize, and two cost values to be constrained below some thresholds. The superior performance of ARNPG-EPD is shown in Figure 4.3.



(a) (b) (c)

Figure 4.3: Last-iterate performance for sample-based ARNPG-EPD, NPG-PD, CRPO averaged over 10 random seeds. The black dashed lines in (b) and (c) represent given thresholds.

Figure 4.3(a) shows that ARNPG-EPD achieves a higher reward value compared to NPG-PD and CRPO, while Figures 4.3(b) and 4.3(c) demonstrate that the cost values of all three algorithms are below the thresholds after a few initial iterations. We believe the superiority is due to the new primal-dual design inspired by [83] (discussed in Chapter 4.3.2) and the flexibility of choosing t_k in the InnerLoop in the framework.

4.5 Conclusion

We propose an ARNPG framework to systematically integrate well-performing first-order methods into the design of policy gradient-based algorithms for multi-objective MDPs. The designed algorithms achieve a global $\tilde{O}(1/T)$ convergence rate under the softmax parameterization with exact gradients and empirically have satisfactory performance beyond tabular and exact gradient settings. We believe that ARNPG has potential applications in other scenarios since the general and flexible framework allows integration with more advanced first-order methods, currently and in the future.

Theoretically, a natural future direction is to extend the results in an exact-gradient tabular setting to more general settings. For example, without having access to the gradient, the policy gradients are estimated from trajectory data in the sample-based setting. The theoretical results in the sample-based setting should incorporate the estimation error induced by the gradient estimate and quantify its impact on the convergence. Due to the KL-divergence regularization in the ARNPG framework, the policy update may not depart too much away from the anchor policy in the inner loop. The number of samples needed can be reduced by performing off-policy gradient estimation [84]. In addition, it would further reduce the sample complexity by leveraging variance-reduced techniques [85, 86]. Besides extending to the sample-based setting, it is also meaningful to consider the function approximation setting, which can handle the MDP with large state-action space.

5. SUMMARY

Information-theoretic methods have been playing important roles in many fields, among which machine learning has recently attracted a lot of attention. Lying in the heart of the information-theoretic methods are the *information-theoretic measures*, such as mutual information, Shannon's entropy, KL divergence, etc, and we have witnessed many success stories of utilizing the information-theoretic measures in communication, data storage, and statistics.

In this dissertation, we delve into the usage of information-theoretic measures in machine learning problems. Our research showcases the efficacy of information-theoretic measures in three fundamental aspects: providing insightful interpretations, enabling precise characterizations, and fostering constructive intuition. Specifically, in Chapter 2, we use information-theoretic measures to reason and interpret the generalization error in machine learning algorithms by developing generalization error upper bounds, where the information-theoretic measures depict the relationship between algorithm, data, and the generalization performance. Moreover, we demonstrate that information-theoretic measures can be used for the precise characterization of algorithmic behaviors and the fundamental limits of sample complexities for estimation problems in Chapter 2.5 and bandits learning in Chapter 3. Furthermore, in Chapter 4, we illustrate that information-theoretic measures not only motivate algorithm design but also facilitate the development of effective learning strategies for addressing the challenges of policy optimization in multi-objective reinforcement learning.

Broadly speaking, the findings presented in this dissertation emphasize the vital role that information theorists can play in pushing the boundaries of machine learning by fostering innovation, deriving accurate and constructive analysis, and driving the development of more robust and efficient learning algorithms. By leveraging the power of information-theoretic measures, we can unlock new possibilities and pave the way for future advancements in the field of machine learning.

REFERENCES

[1] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[2] D. Russo and J. Zou, “Controlling bias in adaptive data analysis using information theory,” in *Artificial Intelligence and Statistics*, pp. 1232–1240, 2016.

[3] A. Xu and M. Raginsky, “Information-theoretic analysis of generalization capability of learning algorithms,” in *Advances in Neural Information Processing Systems*, pp. 2524–2533, 2017.

[4] A. Asadi, E. Abbe, and S. Verdú, “Chaining mutual information and tightening generalization bounds,” in *Advances in Neural Information Processing Systems*, pp. 7234–7243, 2018.

[5] I. Issa, A. R. Esposito, and M. Gastpar, “Strengthened information-theoretic bounds on the generalization error,” in *2019 IEEE International Symposium on Information Theory (ISIT)*, pp. 582–586, IEEE, 2019.

[6] J. Negrea, M. Haghifam, G. K. Dziugaite, A. Khisti, and D. M. Roy, “Information-theoretic generalization bounds for SGLD via data-dependent estimates,” in *Advances in Neural Information Processing Systems*, pp. 11015–11025, 2019.

[7] X. Wu, J. H. Manton, U. Aickelin, and J. Zhu, “Information-theoretic analysis for transfer learning,” in *2020 IEEE International Symposium on Information Theory (ISIT)*, pp. 2819–2824, IEEE, 2020.

[8] Y. Bu, S. Zou, and V. V. Veeravalli, “Tightening mutual information based bounds on generalization error,” *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 121–130, 2020.

[9] T. Steinke and L. Zakynthinou, “Reasoning about generalization via conditional mutual information,” in *Conference on Learning Theory*, pp. 3437–3452, PMLR, 2020.

[10] M. Haghifam, J. Negrea, A. Khisti, D. M. Roy, and G. K. Dziugaite, “Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 9925–9935, 2020.

[11] F. Hellström and G. Durisi, “Generalization error bounds via m-th central moments of the information density,” in *2020 IEEE International Symposium on Information Theory (ISIT)*, pp. 2741–2746, IEEE, 2020.

[12] A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan, “On the theory of policy gradient methods: Optimality, approximation, and distribution shift,” *Journal of Machine Learning Research*, vol. 22, no. 98, pp. 1–76, 2021.

[13] J. Mei, C. Xiao, C. Szepesvari, and D. Schuurmans, “On the global convergence rates of softmax policy gradient methods,” in *International Conference on Machine Learning*, pp. 6820–6829, PMLR, 2020.

[14] Q. Bai, M. Agarwal, and V. Aggarwal, “Joint optimization of multi-objective reinforcement learning with policy gradient based algorithm,” *arXiv preprint arXiv:2105.14125*, 2021.

[15] D. Ding, K. Zhang, T. Basar, and M. Jovanovic, “Natural policy gradient primal-dual method for constrained markov decision processes,” *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[16] T. Xu, Y. Liang, and G. Lan, “Crpo: A new approach for safe reinforcement learning with convergence guarantee,” in *International Conference on Machine Learning*, pp. 11480–11491, PMLR, 2021.

[17] J. Zhang, A. Koppel, A. S. Bedi, C. Szepesvari, and M. Wang, “Variational policy gradient method for reinforcement learning with general utilities,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 4572–4583, 2020.

[18] D. Ying, Y. Ding, and J. Lavaei, "A dual approach to constrained markov decision processes with entropy regularization," in *International Conference on Artificial Intelligence and Statistics*, pp. 1887–1909, PMLR, 2022.

[19] T. Li, Z. Guan, S. Zou, T. Xu, Y. Liang, and G. Lan, "Faster algorithm and sharper analysis for constrained markov decision process," *arXiv preprint arXiv:2110.10351*, 2021.

[20] R. Zhou, C. Tian, and T. Liu, "Individually conditional individual mutual information bound on generalization error," *IEEE Transactions on Information Theory*, vol. 68, no. 5, pp. 3304–3316, 2022.

[21] R. Zhou, C. Tian, and T. Liu, "Stochastic chaining and strengthened information-theoretic generalization bounds," *Journal of the Franklin Institute*, vol. 360, no. 6, pp. 4114–4134, 2023.

[22] R. Zhou, C. Tian, and T. Liu, "Exactly tight information-theoretic generalization error bound for the quadratic gaussian problem," 2023.

[23] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[24] B. Rodríguez-Gálvez, G. Bassi, R. Thobaben, and M. Skoglund, "On random subset generalization error bounds and the stochastic gradient langevin dynamics algorithm," in *2020 IEEE Information Theory Workshop (ITW)*, pp. 1–5, IEEE, 2021.

[25] W. Gao, S. Oh, and P. Viswanath, "Demystifying fixed k-nearest neighbor information estimators," *IEEE Transactions on Information Theory*, vol. 64, no. 8, pp. 5629–5661, 2018.

[26] B. Rimoldi, "Successive refinement of information: Characterization of the achievable rates," *IEEE Transactions on Information Theory*, vol. 40, no. 1, pp. 253–259, 1994.

[27] W. H. Equitz and T. M. Cover, "Successive refinement of information," *IEEE Transactions on Information Theory*, vol. 37, no. 2, pp. 269–275, 1991.

[28] Y. Wu, "Lecture notes on information-theoretic methods for high-dimensional statistics," *Lecture Notes for ECE598YW (UIUC)*, vol. 16, 2017.

[29] H. Hafez-Kolahi, Z. Golgooni, S. Kasaei, and M. Soleymani, "Conditioning and processing: Techniques to improve information-theoretic generalization bounds," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[30] G. K. Dziugaite and D. M. Roy, "Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data," *Uncertainty in Artificial Intelligence*, 2017.

[31] F. Hellström and G. Durisi, "Generalization bounds via information density and conditional information density," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 3, pp. 824–839, 2020.

[32] B. Rodríguez Gálvez, G. Bassi, R. Thobaben, and M. Skoglund, "Tighter expected generalization error bounds via wasserstein distance," *Advances in Neural Information Processing Systems*, vol. 34, pp. 19109–19121, 2021.

[33] R. Zhou and C. Tian, "Approximate top-m arm identification with heterogeneous reward variances," in *International Conference on Artificial Intelligence and Statistics*, pp. 7483–7504, PMLR, 2022.

[34] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in applied mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

[35] S. Bubeck, N. Cesa-Bianchi, *et al.*, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Foundations and Trends® in Machine Learning*, vol. 5, no. 1, pp. 1–122, 2012.

[36] A. Slivkins, "Introduction to multi-armed bandits," *arXiv preprint arXiv:1904.07272*, 2019.

[37] T. Lattimore and C. Szepesvári, *Bandit algorithms*. Cambridge University Press, 2020.

[38] J.-Y. Audibert, S. Bubeck, and R. Munos, "Best arm identification in multi-armed bandits.," in *COLT*, pp. 41–53, 2010.

[39] S. Bubeck, T. Wang, and N. Viswanathan, "Multiple identifications in multi-armed bandits," in *International Conference on Machine Learning*, pp. 258–265, PMLR, 2013.

[40] A. Carpentier and A. Locatelli, "Tight (lower) bounds for the fixed budget best arm identification bandit problem," in *Conference on Learning Theory*, pp. 590–604, PMLR, 2016.

[41] K. Jamieson and R. Nowak, "Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting," in *2014 48th Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–6, IEEE, 2014.

[42] E. Even-Dar, S. Mannor, and Y. Mansour, "PAC bounds for multi-armed bandit and markov decision processes," in *International Conference on Computational Learning Theory*, pp. 255–270, Springer, 2002.

[43] E. Even-Dar, S. Mannor, and Y. Mansour, "Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems," *Journal of machine learning research*, vol. 7, no. Jun, pp. 1079–1105, 2006.

[44] S. Mannor and J. N. Tsitsiklis, "The sample complexity of exploration in the multi-armed bandit problem," *Journal of Machine Learning Research*, vol. 5, no. Jun, pp. 623–648, 2004.

[45] A. Hassidim, R. Kupfer, and Y. Singer, "An optimal elimination algorithm for learning a best arm," *Advances in Neural Information Processing Systems*, vol. 33, pp. 10788–10798, 2020.

[46] J. Katz-Samuels and K. Jamieson, "The true sample complexity of identifying good arms," in *International Conference on Artificial Intelligence and Statistics*, pp. 1781–1791, PMLR, 2020.

[47] Z. Karnin, T. Koren, and O. Somekh, "Almost optimal exploration in multi-armed bandits," in *International Conference on Machine Learning*, pp. 1238–1246, 2013.

[48] K. Jamieson, M. Malloy, R. Nowak, and S. Bubeck, “lil’ucb: An optimal exploration algorithm for multi-armed bandits,” in *Conference on Learning Theory*, pp. 423–439, 2014.

[49] L. Chen and J. Li, “On the optimal sample complexity for best arm identification,” *arXiv preprint arXiv:1511.03774*, 2015.

[50] A. Garivier and E. Kaufmann, “Optimal best arm identification with fixed confidence,” in *Conference on Learning Theory*, pp. 998–1027, 2016.

[51] E. Kaufmann, O. Cappé, and A. Garivier, “On the complexity of best-arm identification in multi-armed bandit models,” *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1–42, 2016.

[52] Y. Zhou, X. Chen, and J. Li, “Optimal pac multiple arm identification with applications to crowdsourcing,” in *International Conference on Machine Learning*, pp. 217–225, PMLR, 2014.

[53] C. Shen, “Universal best arm identification,” *IEEE Transactions on Signal Processing*, vol. 67, no. 17, pp. 4464–4478, 2019.

[54] T. Jin, J. Shi, X. Xiao, and E. Chen, “Efficient pure exploration in adaptive round model,” *Advances in Neural Information Processing Systems*, vol. 32, pp. 6609–6618, 2019.

[55] S. Assadi and C. Wang, “Exploration with limited memory: streaming algorithms for coin tossing, noisy comparisons, and multi-armed bandits,” in *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1237–1250, 2020.

[56] A. R. Chaudhuri and S. Kalyanakrishnan, “Pac identification of many good arms in stochastic multi-armed bandits,” in *International Conference on Machine Learning*, pp. 991–1000, PMLR, 2019.

[57] S. Kalyanakrishnan and P. Stone, “Efficient selection of multiple bandit arms: Theory and practice.,” in *ICML*, vol. 10, pp. 511–518, 2010.

[58] S. Kalyanakrishnan, A. Tewari, P. Auer, and P. Stone, "PAC subset selection in stochastic multi-armed bandits," in *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pp. 227–234, 2012.

[59] V. Gabillon, M. Ghavamzadeh, and A. Lazaric, "Best arm identification: A unified approach to fixed budget and fixed confidence," *Advances in Neural Information Processing Systems*, vol. 25, pp. 3212–3220, 2012.

[60] E. Kaufmann and S. Kalyanakrishnan, "Information complexity in bandit subset selection," in *Conference on Learning Theory*, pp. 228–251, PMLR, 2013.

[61] L. Chen, J. Li, and M. Qiao, "Nearly instance optimal sample complexity bounds for top-k arm selection," in *Artificial Intelligence and Statistics*, pp. 101–110, PMLR, 2017.

[62] M. Simchowitz, K. Jamieson, and B. Recht, "The simulator: Understanding adaptive sampling in the moderate-confidence regime," in *Conference on Learning Theory*, pp. 1794–1834, PMLR, 2017.

[63] R. E. Bechhofer, "A single-sample multiple decision procedure for ranking means of normal populations with known variances," *The Annals of Mathematical Statistics*, pp. 16–39, 1954.

[64] M. Faella, A. Finzi, and L. Sauro, "Rapidly finding the best arm using variance," in *Proc. of ECAI, 24th European Conference of Artificial Intelligence*, 2020.

[65] V. Gabillon, M. Ghavamzadeh, A. Lazaric, and S. Bubeck, "Multi-bandit best arm identification," in *Advances in Neural Information Processing Systems*, pp. 2222–2230, 2011.

[66] P. Lu, C. Tao, and X. Zhang, "Variance-dependent best arm identification," in *Uncertainty in Artificial Intelligence*, pp. 1120–1129, PMLR, 2021.

[67] T. M. Cover, *Elements of information theory*. John Wiley & Sons, 1999.

[68] R. Zhou, T. Liu, D. Kalathil, P. Kumar, and C. Tian, "Anchor-changing regularized natural policy gradient for multi-objective reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 13584–13596, 2022.

[69] F. P. Kelly, A. K. Maulloo, and D. K. H. Tan, “Rate control for communication networks: shadow prices, proportional fairness and stability,” *Journal of the Operational Research society*, vol. 49, no. 3, pp. 237–252, 1998.

[70] E. Altman, *Constrained Markov decision processes*, vol. 7. CRC Press, 1999.

[71] D. Bertsekas and R. Gallager, *Data networks*. Athena Scientific, 2021.

[72] A. Nemirovski, “Prox-method with rate of convergence O(1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems,” *SIAM Journal on Optimization*, vol. 15, no. 1, pp. 229–251, 2004.

[73] A. Beck and M. Teboulle, “Mirror descent and nonlinear projected subgradient methods for convex optimization,” *Operations Research Letters*, vol. 31, no. 3, pp. 167–175, 2003.

[74] S. Rakhlin and K. Sridharan, “Optimization, learning, and games with predictable sequences,” *Advances in Neural Information Processing Systems*, vol. 26, pp. 3066–3074, 2013.

[75] C.-Y. Wei, C.-W. Lee, M. Zhang, and H. Luo, “Linear last-iterate convergence in constrained saddle-point optimization,” in *International Conference on Learning Representations*, 2020.

[76] G. Lan, *First-order and Stochastic Optimization Methods for Machine Learning*. Springer Nature, 2020.

[77] S. Cen, C. Cheng, Y. Chen, Y. Wei, and Y. Chi, “Fast global convergence of natural policy gradient methods with entropy regularization,” *Operations Research*, 2021.

[78] H. Yu and M. J. Neely, “A simple parallel algorithm with an O(1/t) convergence rate for general convex programs,” *SIAM Journal on Optimization*, vol. 27, no. 2, pp. 759–783, 2017.

[79] A. Jain, S. Vaswani, R. Babanezhad, C. Szepesvari, and D. Precup, “Towards painless policy optimization for constrained mdps,” in *Uncertainty in Artificial Intelligence*, pp. 895–905, PMLR, 2022.

[80] T. Liu, R. Zhou, D. Kalathil, P. Kumar, and C. Tian, "Learning policies with zero or bounded constraint violation for constrained mdps," *Advances in Neural Information Processing Systems*, vol. 34, pp. 17183–17193, 2021.

[81] G. Lan, "Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes," *Mathematical programming*, vol. 198, no. 1, pp. 1059–1106, 2023.

[82] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "Openai gym," 2016.

[83] H. Yu and M. J. Neely, "A primal-dual parallel method with $o(1/\epsilon)$ convergence for constrained composite convex programs," *arXiv preprint arXiv:1708.00322*, 2017.

[84] C. Ni, R. Zhang, X. Ji, X. Zhang, and M. Wang, "Optimal estimation of policy gradient via double fitted iteration," in *International Conference on Machine Learning*, pp. 16724–16783, PMLR, 2022.

[85] Y. Liu, K. Zhang, T. Basar, and W. Yin, "An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7624–7636, 2020.

[86] J. Zhang, C. Ni, C. Szepesvari, M. Wang, *et al.*, "On the convergence and sample efficiency of variance-reduced policy gradient method," *Advances in Neural Information Processing Systems*, vol. 34, pp. 2228–2240, 2021.

[87] J. V. Michalowicz, J. M. Nichols, and F. Bucholtz, "Calculation of differential entropy for a mixed Gaussian distribution," *Entropy*, vol. 10, no. 3, pp. 200–206, 2008.

[88] X. Wei, H. Yu, and M. J. Neely, "Online primal-dual mirror descent under stochastic constraints," *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, vol. 4, no. 2, pp. 1–36, 2020.

APPENDIX A

PROOFS FOR CHAPTER 2

Proof of Lemma 1. The definition of conditional CGF implies that

$$\Psi_{\tilde{F}|U}(\lambda) = \ln \mathbb{E}\left[e^{\lambda \tilde{F}}|U\right] - \mathbb{E}[\lambda \tilde{F}|U]. \tag{A.1}$$

By the Donsker–Varadhan variational representation of KL divergence, for any $\lambda \in \mathbb{R}$

$$\begin{aligned} \mathbb{E}[\lambda F|U] - \ln \mathbb{E}\left[e^{\lambda \tilde{F}}|U\right] &\leq D(P_{X,Y|U}||P_{\tilde{X},\tilde{Y}|U}) && \text{(A.2)} \\ &= I_U(X;Y), && \text{(A.3)} \end{aligned}$$

where the equality is due to (2.27). It follows that for $\lambda > 0$

$$\begin{aligned} \mathbb{E}[F|U] - \mathbb{E}[\tilde{F}|U] &\leq \inf_{\lambda>0} \frac{I_U(X;F) + \Psi_{\tilde{F}|U}(\lambda)}{\lambda} && \text{(A.4)} \\ &= \Psi^{*-1}_{\tilde{F}|U}\left(I_U(X;Y)\right). && \text{(A.5)} \end{aligned}$$

Moreover

$$\begin{aligned} \mathbb{E}[F] - \mathbb{E}[\tilde{F}] &\leq \mathbb{E}\left[\Psi^{*-1}_{\tilde{F}|U}\left(I_U(X;Y)\right)\right] && \text{(A.6)} \\ &= \mathbb{E}\left[\inf_{\lambda>0} \frac{I_U(X;F) + \Psi_{\tilde{F}|U}(\lambda)}{\lambda}\right] && \text{(A.7)} \\ &\leq \inf_{\lambda>0} \frac{I(X;F|U) + \mathbb{E}\left[\Psi_{\tilde{F}|U}(\lambda)\right]}{\lambda} && \text{(A.8)} \\ &= \bar{\psi}^{*-1}_{\tilde{F}|U}\left(I(X;Y|U)\right), && \text{(A.9)} \end{aligned}$$

where the last inequality is by exchanging the order of expectation and infimum. The proof is thus complete. □

Proof of Lemma 2. By the independence of R_i and $Z^{\pm}_{[n]}$, we have

$$I(W; R_i|Z^{\pm}_{[n]}) = H(R_i) - H(R_i|W, Z^{\pm}_{[n]}), \quad I(W; R_i|Z^{\pm}_i) = H(R_i) - H(R_i|W, Z^{\pm}_i).$$

It follows that $I(W; R_i|Z^{\pm}_{[n]}) - I(W; R_i|Z^{\pm}_i) = I(R_i; Z^{\pm}_{[n]}|W, Z^{\pm}_i) \geq 0$, which concludes the proof. □

Proof of Lemma 3. First Z_i and $Z^{R_i}_i$ are both the i^{th} training sample for the input of the algorithm, thus $I(W; Z_i) = I(W; Z^{R_i}_i)$. Then since $Z^{-R_i}_i$, R_i and W are independent given $Z^{R_i}_i$,

$$I(W; Z^{\pm}_i, R_i) = I(W; Z^{R_i}_i, Z^{-R_i}_i, R_i) = I(W; Z^{R_i}_i) + I(W; Z^{-R_i}_i, R_i|Z^{R_i}_i) = I(W; Z^{R_i}_i). \tag{A.10}$$

It follows that

$$I(W; Z_i) = I(W; Z^{\pm}_i, R_i) \geq I(W; R_i|Z^{\pm}_i), \tag{A.11}$$

which concludes the proof. □

Proof of Proposition 2. Note that $I_{Z^{\pm}_i}(W; R_i) = I_{Z^{\pm}_i}\left(W - \frac{Z^-_i + Z^+_i}{2}; R_i\right)$ since $Z^{\pm}_i$ is known. Without loss of generality, Z^-_i is assumed to be greater than Z^+_i. Then $W - \frac{Z^-_i + Z^+_i}{2}$ is a mixture of two Gaussian distributions with the form of V in Lemma 4 where $\nu = (Z^-_i - Z^+_i)/2$. By Lemma 4, we have

$$I_{Z^{\pm}_i}(W; R_i) = \frac{(Z^-_i - Z^+_i)^2}{8\sigma^2}\frac{1}{n-1} + o\left(\frac{1}{n}\right). \tag{A.12}$$

Then Theorem 5 and Lemma 5 imply

$$\text{gen}(\xi, P_{W|Z_{[n]}}) \leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\Psi^{*-1}_{\tilde{G}_i|Z^{\pm}_i}\left(I_{Z^{\pm}_i}(W; R_i)\right)\right] \tag{A.13}$$

$$\leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\frac{(Z_i^- - Z_i^+)^2|Z_i^- + Z_i^+|}{2\sigma\sqrt{n-1}} + o\left(\frac{1}{\sqrt{n}}\right)\right] \tag{A.14}$$

$$= \frac{2\sigma^2}{\sqrt{\pi}}\sqrt{\frac{1}{n-1}} + o\left(\frac{1}{\sqrt{n}}\right), \tag{A.15}$$

which proves the proposition. □

Proof of Proposition 3. Similar to [8], we apply the ICIMI bound conditioned on the random sample path $V_{([T])}$, i.e.,

$$\begin{aligned}\mathrm{gen}(\xi, P_{W|Z_{[n]}}) &= \mathbb{E}_{V_{([T])}}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{Z_i^\pm}\left[\mathbb{E}_{W,R_i}\left[R_i\left(\ell(W,Z_i^-) - \ell(W,Z_i^+)\right)|Z_i^\pm, V_{([T])}\right]\middle|V_{([T])}\right]\right]\\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{Z_i^\pm,V_{([T])}}\Psi^{*-1}_{\tilde{F}_i|Z_i^\pm,V_{([T])}}\left(I_{Z_i^\pm,V_{([T])}}(W;R_i)\right).\end{aligned} \tag{A.16}$$

In order to bound the mutual information term, for any sample path $V_{([T])} = v_{([T])}$ and any samples $Z_i^\pm = z_i^\pm$ consider

$$\begin{aligned}I_{z_i^\pm,v_{([T])}}(W;R_i) &\leq I_{z_i^\pm,v_{([T])}}(W_{([T])};R_i)\\ &= \sum_{\tau=1}^{T} I_{z_i^\pm,v_{([T])}}(W_{(\tau)};R_i|W_{([\tau-1])})\\ &= \sum_{\tau\in\mathcal{T}_i(v_{([T])})} I_{z_i^\pm,v_{([T])}}(W_{(\tau)};R_i|W_{([\tau-1])}),\end{aligned} \tag{A.17}$$

where $\mathcal{T}_i(v_{([T])})$ is the realization of $\mathcal{T}_i$ given $V_{([T])} = v_{([T])}$ and the last equality is because for fixed $v_{([T])}$ and $z_i^\pm$, R_i is independent of $W_{(\tau)}$ given $W_{([\tau-1])}$ for the iterations when index i is not selected. We can then continue to write

$$\begin{aligned}I_{z_i^\pm,v_{([T])}}(W_{(\tau)};R_i|W_{([\tau-1])}) = h(W_{(\tau)}|W_{([\tau-1])}, Z_i^\pm = z_i^\pm, V_{([T])} = v_{([T])})\\ - h(W_{(\tau)}|R_i, W_{([\tau-1])}, Z_i^\pm = z_i^\pm, V_{([T])} = v_{([T])}).\end{aligned} \tag{A.18}$$

Let us consider the first term for fixed $W_{[(\tau-1)]} = w_{[(\tau-1)]}$, and for simplicity, denote the condition

$$W_{[(\tau-1)]} = w_{[(\tau-1)]}, \quad Z_i^{\pm} = z_i^{\pm}, \quad V_{([T])} = v_{([T])} \tag{A.19}$$

as $C_{i,(\tau)}$. Thus we have

$$\begin{aligned}
&h(W_{(\tau)}|W_{([\tau-1])} = w_{([\tau-1])}, Z_i^{\pm} = z_i^{\pm}, V_{([T])} = v_{([T])}) \\
&= h\left(\eta_{(\tau)}\nabla\ell(w_{(\tau-1)}, z_i^{R_i}) + \sigma_{(\tau)}\epsilon_{(\tau)}|C_{i,(\tau)}\right) \\
&= h\Bigg(\eta_{(\tau)}\nabla\ell(w_{(\tau-1)}, z_i^{R_i}) - \pi_{i,(\tau)}\eta_{(\tau)}\nabla\ell(w_{(\tau-1)}, z_i^{+}) \\
&\qquad - (1-\pi_{i,(\tau)})\eta_{(\tau)}\nabla\ell(w_{(\tau-1)}, z_i^{-}) + \sigma_{(\tau)}\epsilon_{(\tau)}\Bigg|C_{i,(\tau)}\Bigg).
\end{aligned} \tag{A.20}$$

Clearly the term

$$\nabla_{i,(\tau)} := \nabla\ell(w_{(\tau-1)}, z_i^{R_i}) - \pi_{i,(\tau)}\nabla\ell(w_{(\tau-1)}, z_i^{+}) - (1-\pi_{i,(\tau)})\nabla\ell(w_{(\tau-1)}, z_i^{-}) \tag{A.21}$$

has probability masses only on two elements, and its only source of randomness is due to R_i. Its covariance is thus given by

$$\begin{aligned}
\mathbb{E}\left[\nabla_{i,(\tau)}\nabla_{i,(\tau)}^{\top}|C_{i,(\tau)}\right] &= \mathbb{E}\left[\left(\frac{R_i+1}{2} - \pi_{i,(\tau)}\right)^2 \Bigg|C_{i,(\tau)}\right]\zeta_{(\tau)}(z_i^{\pm})\zeta_{(\tau)}(z_i^{\pm})^{\top} \\
&= \Theta_{i,(\tau)}(w_{([\tau-1])}, z_i^{\pm}, v_{([T])})\zeta_{(\tau)}(z_i^{\pm})\zeta_{(\tau)}(z_i^{\pm})^{\top}.
\end{aligned} \tag{A.22}$$

Since $\epsilon_{(\tau)}$ is independent of $\nabla_{i,(\tau)}$ under condition $C_{i,(\tau)}$, the covariance matrix of $\eta_{(\tau)}\nabla_{i,(\tau)} + \sigma_{(\tau)}\epsilon_{(\tau)}$ is thus

$$\eta_{(\tau)}^2\Theta_{i,(\tau)}(w_{([\tau-1])}, z_i^{\pm}, v_{([T])})\zeta_{(\tau)}(z_i^{\pm})\zeta_{(\tau)}(z_i^{\pm})^{\top} + \sigma_{(\tau)}^2\mathrm{I}_d, \tag{A.23}$$

where I_d is the $d \times d$ identity matrix. Consequently, the determinant of the conditional covariance matrix is

$$\left(\sigma_{(\tau)}^2 + \eta_{(\tau)}^2 \Theta_{i,(\tau)}(w_{([\tau-1])}, z_i^{\pm}, v_{([T])}) \|\zeta_{(\tau)}(z_i^{\pm})\|_2^2\right) \sigma_{(\tau)}^{2(d-1)}. \tag{A.24}$$

Since Gaussian distributions maximizes the differential entropy for random vectors with a given covariance matrix, the conditional differential entropy in (A.20) is upper bounded by

$$\frac{1}{2}\log\left(1 + \frac{\eta_{(\tau)}^2 \Theta_{i,(\tau)}(w_{([\tau-1])}, z_i^{\pm}, v_{([T])}) \|\zeta_{(\tau)}(z_i^{\pm})\|_2^2}{\sigma_{(\tau)}^2}\right) + \frac{d}{2}\log\left(2\pi e \sigma_{(\tau)}^2\right). \tag{A.25}$$

Recall that $S_{i,\tau} = \frac{\eta_{(\tau)}^2 \mathbb{E}\left[\Theta_{i,(\tau)}\|\zeta_{(\tau)}(Z_i^{\pm})\|_2^2 \,\middle|\, Z_i^{\pm}, V_{([T])}\right]}{2\sigma_{(\tau)}^2}$. The upper bound holds for any realizations of condition $C_{i,(\tau)}$, and it follows that the first term in (A.18) can be bounded as

$$\begin{aligned} & h(W_{(\tau)}|W_{([\tau-1])}, Z_i^{\pm} = z_i^{\pm}, V_{([T])} = v_{([T])}) && \text{(A.26)} \\ & \leq \frac{1}{2}\log\left(1 + \mathbb{E}\left[2S_{i,\tau} \,\middle|\, Z_i^{\pm} = z_i^{\pm}, V_{([T])} = v_{([T])}\right]\right) + \frac{d}{2}\log\left(2\pi e \sigma_{(\tau)}^2\right). && \text{(A.27)} \end{aligned}$$

The second term in (A.18) can be straightforwardly calculated as $\frac{d}{2}\log(2\pi e \sigma_{(\tau)}^2)$, and we thus have

$$\begin{aligned} & I_{z_i^{\pm}, v_{([T])}}(W_{(\tau)}; R_i | W_{([\tau-1])}) \\ & \leq \frac{1}{2}\log\left(1 + \mathbb{E}\left[2S_{i,\tau} \,\middle|\, Z_i^{\pm} = z_i^{\pm}, V_{([T])} = v_{([T])}\right]\right) \\ & \leq \mathbb{E}\left[S_{i,\tau} \,\middle|\, Z_i^{\pm} = z_i^{\pm}, V_{([T])} = v_{([T])}\right], \end{aligned} \tag{A.28}$$

since $\log(1+x) \leq x$. Combining (A.16) and (A.28), we arrive at

$$\text{gen}(\xi, P_{W|Z_{[n]}}) \leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\Psi_{\tilde{F}_i|Z_i^{\pm}, V_{([T])}}^{*-1}\left(\sum_{\tau \in \mathcal{T}_i} S_{i,\tau}\right)\right],$$

which is the desired result. □

Proof of Proposition 1. For the special case $n = 1$, i.e., there is only one training sample, the

CMI based bound and CIMI based bound, i.e., (2.37), (2.38), are equal. It is straightforward to verify that conditioned on $Z_1^{\pm}$, $\tilde{E} = \tilde{E}_1$ and $\tilde{E}$ takes $(Z_1^- - Z_1^+)^2$ with probability $\frac{1}{2}$ and takes $-(Z_1^- - Z_1^+)^2$ with probability $\frac{1}{2}$. Then we have

$$\Psi_{\tilde{E}|Z_{[1]}^{\pm}}(\lambda) = \ln\cosh\left((Z_1^- - Z_1^+)^2\lambda\right). \tag{A.29}$$

Their inverse Fenchel conjugate functions are equal and by the lower bound of $\ln\cosh(\cdot)$ function in Lemma 13,

$$\begin{aligned}\Psi^{*-1}_{\tilde{E}|Z_{[1]}^{\pm}}(\eta) = \inf_{\lambda>0}\frac{\eta+\Psi_{\tilde{E}|Z_{[1]}^{\pm}}(\lambda)}{\lambda} &\geq \inf_{\lambda>0}\frac{\eta+\min\left(\frac{(Z_1^- - Z_1^+)^2\lambda}{2}, \frac{(Z_1^- - Z_1^+)^4\lambda^2}{4}\right)}{\lambda} &(A.30)\\ &= \min\left(\frac{1}{2}, \sqrt{\eta}\right)(Z_1^- - Z_1^+)^2. &(A.31)\end{aligned}$$

Since $I_{Z_1^{\pm}}(W; R_1) = 1/\log e, a.s.,$, we have

$$\mathbb{E}\left[\Psi^{*-1}_{\tilde{E}|Z_{[1]}^{\pm}}\left(I_{Z_1^{\pm}}(W; R_1)\right)\right] \geq \sigma^2 > \frac{\sigma^2}{\pi\sqrt{\log e}}. \tag{A.32}$$

For $n \geq 2$, denote the mean of $Z_{[n]}^{\pm}$ as $\bar{Z}$, from which we have

$$\bar{Z} = \mathbb{E}\left[\tilde{W}|Z_{[n]}^{\pm}\right]. \tag{A.33}$$

For each $i = 1, \ldots, n$, let $\Delta_i = \ell(\bar{Z}, Z_i^-) - \ell(\bar{Z}, Z_i^+)$. It follows that

$$\begin{aligned}\Delta_i &= \left(Z_i^- - Z_i^+\right)\left(Z_i^- + Z_i^+ - 2\bar{Z}\right) &(A.34)\\ &= \left(1-\frac{1}{n}\right)\left(\left(Z_i^-\right)^2 - \left(Z_i^+\right)^2\right) - \frac{\sum_{j\neq i}(Z_j^- + Z_j^+)}{n}(Z_i^- - Z_i^+). &(A.35)\end{aligned}$$

Thus

$$\mathbb{E}[|\Delta_i|] = \mathbb{E}\left[\mathbb{E}\left[|\Delta_i||Z_i^{\pm}\right]\right] \geq \mathbb{E}\left[\left|\mathbb{E}\left[\Delta_i|Z_i^{\pm}\right]\right|\right]$$

$$= \left(1 - \frac{1}{n}\right) \mathbb{E}\left[\left| \left(Z_i^-\right)^2 - \left(Z_i^+\right)^2 \right| \right] \tag{A.36}$$

$$\geq \frac{1}{2}\mathbb{E}\left[|Z_i^- - Z_i^+|\right] \mathbb{E}\left[|Z_i^- + Z_i^+|\right] = \frac{2\sigma^2}{\pi}, \tag{A.37}$$

where the first inequality is by applying Jensen's inequality with respect to convex function $|\cdot|$; the last inequality is because $n \geq 2$ and $Z_i^- - Z_i^+$ and $Z_i^- + Z_i^+$ are independent; the last equality is calculated using the fact that $|Z_i^- - Z_i^+|$ and $|Z_i^- + Z_i^+|$ follow the folded Gaussian distribution. In addition, we can write

$$\begin{aligned} &\mathbb{E}\left[\ell(\tilde{W}, Z_i^-) - \ell(\tilde{W}, Z_i^+) | Z_{[n]}^{\pm}\right] = \mathbb{E}\left[\left(Z_i^- - Z_i^+\right)\left(Z_i^- + Z_i^+ - 2\tilde{W}\right) | Z_{[n]}^{\pm}\right] \\ &= \left(Z_i^- - Z_i^+\right)\left(Z_i^- + Z_i^+ - 2\mathbb{E}\left[\tilde{W} | Z_{[n]}^{\pm}\right]\right) = \left(Z_i^- - Z_i^+\right)\left(Z_i^- + Z_i^+ - 2\bar{Z}\right) = \Delta_i, \end{aligned} \tag{A.38}$$

where the last equality is by the representation of $\bar{Z}$ in (A.33).

We can then lower-bound the CMI based bound (2.37) for this problem. The function $\Psi_{\tilde{E}|Z_{[n]}^{\pm}}(\lambda)$ satisfies

$$\Psi_{\tilde{E}|Z_{[n]}^{\pm}}(\lambda) = \ln \mathbb{E}\left[\exp\left(\lambda\tilde{E} - \lambda\mathbb{E}[\tilde{E}]\right) \middle| Z_i^{\pm}\right] \tag{A.39}$$

$$= \ln \mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sum_{i=1}^{n} \tilde{R}_i(\ell(\tilde{W}, Z_i^-) - \ell(\tilde{W}, Z_i^+))\right) \middle| Z_{[n]}^{\pm}\right] \tag{A.40}$$

$$= \ln \mathbb{E}\left[\mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sum_{i=1}^{n} \tilde{R}_i(\ell(\tilde{W}, Z_i^-) - \ell(\tilde{W}, Z_i^+))\right) \middle| Z_{[n]}^{\pm}, \tilde{R}_{[n]}\right] \middle| Z_{[n]}^{\pm}\right] \tag{A.41}$$

$$\geq \ln \mathbb{E}\left[\exp\left(\frac{\lambda}{n}\mathbb{E}\left[\sum_{i=1}^{n} \tilde{R}_i(\ell(\tilde{W}, Z_i^-) - \ell(\tilde{W}, Z_i^+)) \middle| Z_{[n]}^{\pm}, \tilde{R}_{[n]}\right]\right) \middle| Z_{[n]}^{\pm}\right] \tag{A.42}$$

$$= \ln \mathbb{E}\left[\prod_{i=1}^{n} \exp\left(\frac{\lambda}{n}\tilde{R}_i\Delta_i\right) \middle| Z_{[n]}^{\pm}\right] \tag{A.43}$$

$$= \sum_{i=1}^{n} \ln \mathbb{E}\left[\exp\left(\frac{\lambda}{n}\tilde{R}_i\Delta_i\right) \middle| Z_{[n]}^{\pm}\right] \tag{A.44}$$

$$= \sum_{i=1}^{n} \ln\cosh\left(\frac{\lambda}{n}\Delta_i\right) \geq \sum_{i=1}^{n} \min\left(1, \frac{\lambda|\Delta_i|}{2n}\right)\frac{\lambda|\Delta_i|}{2n}. \tag{A.45}$$

The first equality (A.39) is the definition of $\Psi_{\tilde{E}|Z^{\pm}_{[n]}}(\lambda)$; the second equality (A.40) is by $\mathbb{E}[\tilde{E}] = 0$; the third equality (A.41) is by the law of total expectation; the first inequality (A.42) is by Jensen's inequality with respect to convex function $\exp(\cdot)$; the fourth equality (A.43) is by (A.38); the fifth equality (A.44) is by the independence of $\tilde{R}_{[n]}$ conditioned on $Z^{\pm}_{n}$; and the last inequality is due to Lemma 13. Its inverse Fenchel conjugate function can thus be lower bounded as follows.

$$\Psi^{*-1}_{\tilde{E}|Z^{\pm}_{[n]}}(\eta) = \inf_{\lambda>0} \frac{\eta + \Psi_{\tilde{E}|Z^{\pm}_{[n]}}(\lambda)}{\lambda} \geq \inf_{\lambda>0} \sum_{i=1}^{n} \frac{\frac{1}{n}\eta + \min\left(1, \frac{\lambda|\Delta_i|}{2n}\right)\frac{\lambda|\Delta_i|}{2n}}{\lambda} \tag{A.46}$$

$$\geq \sum_{i=1}^{n} \inf_{\lambda>0} \frac{\frac{1}{n}\eta + \min\left(1, \frac{\lambda|\Delta_i|}{2n}\right)\frac{\lambda|\Delta_i|}{2n}}{\lambda} \geq \sum_{i=1}^{n} \min\left(\frac{|\Delta_i|}{2n}, \frac{\sqrt{\eta}|\Delta_i|}{n^{3/2}}\right). \tag{A.47}$$

Then since $I_{Z^{\pm}_{n}}(W; R_{[n]}) = n/\log e, a.s.$ and $\Psi^{*-1}_{\tilde{E}|Z^{\pm}_{[n]}}$ is non-negative, the CMI based bound satisfies

$$\mathbb{E}\left[\Psi^{*-1}_{\tilde{E}|Z^{\pm}_{[n]}}(I_{Z^{\pm}_{[n]}}(W; R_{[n]}))\right] \geq \sum_{i=1}^{n} \mathbb{E}\left[\min\left(\frac{|\Delta_i|}{2n}, \frac{|\Delta_i|}{\sqrt{\log en}}\right)\right] \tag{A.48}$$

$$\geq \sum_{i=1}^{n} \mathbb{E}\left[\frac{|\Delta_i|}{2\sqrt{\log en}}\right] \geq \frac{\sigma^2}{\pi\sqrt{\log e}}, \tag{A.49}$$

where the last equality is by (A.37).

Similarly, we can lower-bound the CIMI based bound (2.38). The function $\Psi_{\tilde{E}_i|Z^{\pm}_{[n]}}(\lambda)$ satisfies

$$\Psi_{\tilde{E}_i|Z^{\pm}_{[n]}}(\lambda) = \ln \mathbb{E}\left[\exp\left(\lambda \tilde{R}_i(\ell(\tilde{W}, Z^{-}_{i}) - \ell(\tilde{W}, Z^{+}_{i}))\right) \middle| Z^{\pm}_{[n]}\right] \tag{A.50}$$

$$= \ln\cosh(\lambda\Delta_i) \geq \min\left(1, \frac{|\lambda\Delta_i|}{2}\right)\frac{|\lambda\Delta_i|}{2}. \tag{A.51}$$

The inverse Fenchel conjugate functions can be lower bounded as

$$\Psi^{*-1}_{\tilde{E}_i|Z^{\pm}_{[n]}}(\eta) \geq \inf_{\lambda>0} \frac{\eta + \min\left(1, \frac{|\lambda\Delta_i|}{2}\right)\frac{|\lambda\Delta_i|}{2}}{\lambda} \tag{A.52}$$

$$= \min\left(\inf_{\lambda>0} \frac{\eta + \lambda\frac{|\Delta_i|}{2}}{\lambda}, \inf_{\lambda>0} \frac{\eta + \lambda^2\frac{\Delta_i^2}{4}}{\lambda}\right) \tag{A.53}$$

$$= \min\left(\frac{1}{2}, \sqrt{\eta}\right)|\Delta_i|. \tag{A.54}$$

Since $\Psi^{*-1}_{\tilde{E}_i|Z^{\pm}_{[n]}}(\eta)$ is non-negative, and $I_{Z^{\pm}_n}(W;R_i) = 1/\log e, a.s.$, the CIMI based bound satisfies

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\Psi^{*-1}_{\tilde{E}_i|Z^{\pm}_{[n]}}(I_{Z^{\pm}_{[n]}}(W;R_i))\right] \geq \frac{1}{2\sqrt{\log en}}\sum_{i=1}^{n}\mathbb{E}[|\Delta_i|] = \frac{\sigma^2}{\pi\sqrt{\log e}}. \tag{A.55}$$

We can now conclude that the CMI and CIMI bounds in this setting are both at least $\frac{\sigma^2}{\pi\sqrt{\log e}}$. □

Proof of Lemma 4. By the representation of the differential entropy of mixed Gaussian distribution in [87], we can write

$$I(V;R) = h(V) - h(V|R) = \alpha^2 - I(\alpha), \tag{A.56}$$

where $\alpha = \frac{|\nu|}{\sigma}$ and

$$I(\alpha) = \frac{2}{\sqrt{2\pi}}e^{-\alpha^2/2}\int_0^\infty e^{-t^2/2}\text{cosh}(\alpha t)\ln(\text{cosh}(\alpha t))dt.$$

Since for any $x \in \mathbb{R}$, by the Taylor expansion,

$$1 + x^2/2 \leq \text{cosh}(x) = \frac{1}{2}(e^x + e^{-x}) \leq e^{x^2/2}, \tag{A.57}$$

it follows that for any $\alpha < 1$,

$$\frac{\frac{2}{\sqrt{2\pi}}e^{-\alpha^2/2}\int_0^\infty e^{-t^2/2}\text{cosh}(\alpha t)\ln(\text{cosh}(\alpha t))dt}{\alpha^2} \leq \frac{\frac{2}{\sqrt{2\pi}}e^{-\alpha^2/2}\int_0^\infty e^{-t^2/2}e^{\alpha^2t^2/2}\frac{\alpha^2t^2}{2}dt}{\alpha^2} \tag{A.58}$$

$$= \frac{1}{\sqrt{2\pi}}e^{-\alpha^2/2}\int_0^\infty t^2e^{-t^2(1-\alpha^2)/2}dt = \frac{1}{2\sqrt{1-\alpha^2}}e^{-\alpha^2/2}, \tag{A.59}$$

and take the limit of $\alpha^2 \to 0$ on both side,

$$\lim_{\alpha^2\to 0}\frac{\frac{2}{\sqrt{2\pi}}e^{-\alpha^2/2}\int_0^\infty e^{-t^2/2}\text{cosh}(\alpha t)\ln(\text{cosh}(\alpha t))dt}{\alpha^2} \leq \frac{1}{2}. \tag{A.60}$$

In addition,

$$\frac{\frac{2}{\sqrt{2\pi}}e^{-\alpha^2/2}\int_0^\infty e^{-t^2/2}\text{cosh}(\alpha t)\ln(\text{cosh}(\alpha t))dt}{\alpha^2} \tag{A.61}$$

$$\geq \frac{\frac{2}{\sqrt{2\pi}}e^{-\alpha^2/2}\int_0^\infty e^{-t^2/2}\left(1+\frac{\alpha^2t^2}{2}\right)\ln\left(1+\frac{\alpha^2t^2}{2}\right)dt}{\alpha^2} \tag{A.62}$$

take the limit of $\alpha^2 \to 0$ on both side,

$$\lim_{\alpha^2\to 0}\frac{\frac{2}{\sqrt{2\pi}}e^{-\alpha^2/2}\int_0^\infty e^{-t^2/2}\text{cosh}(\alpha t)\ln(\text{cosh}(\alpha t))dt}{\alpha^2} \tag{A.63}$$

$$\geq \lim_{\alpha^2\to 0}\frac{\frac{2}{\sqrt{2\pi}}e^{-\alpha^2/2}\int_0^\infty e^{-t^2/2}\left(1+\frac{\alpha^2t^2}{2}\right)\ln\left(1+\frac{\alpha^2t^2}{2}\right)dt}{\alpha^2} \tag{A.64}$$

$$= \frac{2}{\sqrt{2\pi}}\int_0^\infty e^{-t^2/2}\lim_{\alpha^2\to 0}\frac{\left(1+\frac{\alpha^2t^2}{2}\right)\ln\left(1+\frac{\alpha^2t^2}{2}\right)}{\alpha^2}dt \tag{A.65}$$

$$= \frac{1}{\sqrt{2\pi}}\int_0^\infty t^2e^{-t^2/2}dt = \frac{1}{2}, \tag{A.66}$$

where the first equality is by exchanging the limit and integral because function $\frac{(1+x)\ln(1+x)}{x}$ is monotonically increasing for $x \geq 0$ and $\lim_{\alpha^2\to 0}e^{-\alpha^2/2} = 1$. Thus the Taylor expansion of $I(\alpha)$ is

$$I(\alpha) = \frac{1}{2}\alpha^2 + o(\alpha^2), \tag{A.67}$$

plugging which in equation (A.56) completes the proof. □

Proof of Lemma 5. Given $Z_i^{\pm} = z_{\pm} \in \mathcal{Z}^2$, $\tilde{W}_i$ and W are identically distributed. Drop the index i and write $\tilde{W}_i$ as $\tilde{W}$ for simplicity. With probability $1/2$, $\tilde{W} \sim N\left(\frac{z_+}{n}, \frac{n-1}{n^2}\sigma^2\right)$, and with probability $1/2$, $\tilde{W} \sim N\left(\frac{z_-}{n}, \frac{n-1}{n^2}\sigma^2\right)$. For any $\lambda > 0$,

$$\exp\left(\Psi_{\tilde{G}_i|Z_i^{\pm}=z_{\pm}}(\lambda)\right) \tag{A.68}$$

$$=\mathbb{E}\left[\exp\left(\lambda\tilde{R}\left(\ell(\tilde{W},z_-)-\ell(\tilde{W},z_+)\right)\right)\right] \tag{A.69}$$

$$=\mathbb{E}\left[\exp\left(\lambda\tilde{R}\left(z_-^2-z_+^2+2(z_+-z_-)\tilde{W}\right)\right)\right] \tag{A.70}$$

$$
\begin{aligned}
&=\frac{1}{2}\mathbb{E}\left[\exp\left(2\lambda(z_+ - z_-)\tilde{W}\right)\right]\exp\left(\lambda(z_-^2 - z_+^2)\right)\\
&+\frac{1}{2}\mathbb{E}\left[\exp\left(2\lambda(z_- - z_+)\tilde{W}\right)\right]\exp\left(-\lambda(z_-^2 - z_+^2)\right)
\end{aligned} \tag{A.71}
$$

$$
\begin{aligned}
\leq&\Bigg(\frac{1}{2}\exp\left(2\lambda|z_+ - z_-|\frac{|z_-|}{n} + 2\lambda^2(z_+ - z_-)^2\frac{n-1}{n^2}\sigma^2\right)\\
&+\frac{1}{2}\exp\left(2\lambda|z_+ - z_-|\frac{|z_+|}{n} + 2\lambda^2(z_+ - z_-)^2\frac{n-1}{n^2}\sigma^2\right)\Bigg)\\
&\cdot\left(\frac{1}{2}\exp(\lambda(z_-^2 - z_+^2)) + \frac{1}{2}\exp(\lambda(z_+^2 - z_-^2))\right)
\end{aligned} \tag{A.72}
$$

$$
\begin{aligned}
\leq&\exp\left(2\sigma^2\lambda^2(z_+ - z_-)^2\left(\frac{1}{n} - \frac{1}{n^2}\right)\right)\\
&\cdot\exp\left(2\lambda|z_+ - z_-|\frac{\max(|z_+|, |z_-|)}{n}\right)\\
&\cdot\left(\frac{1}{2}\exp(\lambda(z_-^2 - z_+^2)) + \frac{1}{2}\exp(\lambda(z_+^2 - z_-^2))\right)
\end{aligned} \tag{A.73}
$$

$$
\begin{aligned}
\leq&\exp\left(2\sigma^2\lambda^2(z_+ - z_-)^2\frac{1}{n}\right)\\
&\cdot\exp\left(2\lambda|z_+ - z_-|\frac{\max(|z_+|, |z_-|)}{n}\right)\\
&\cdot\exp\left(\frac{\lambda^2}{2}(z_-^2 - z_+^2)^2\right),
\end{aligned} \tag{A.74}
$$

where the last inequality is from $\frac{1}{2}(e^x + e^{-x}) \leq e^{x^2/2}$. We have for any $\eta > 0$,

$$
\Psi^{*-1}_{\tilde{G}_i|Z_i^\pm}(\eta) = \inf_{\lambda>0}\left\{\frac{1}{\lambda}\left(\eta + \Psi_{\tilde{G}_i|Z_i^\pm}(\lambda)\right)\right\} \tag{A.75}
$$

$$
\leq \inf_{\lambda>0}\left\{\frac{1}{\lambda}\eta + \frac{\lambda}{2}\left(\left(Z_i^+\right)^2 - \left(Z_i^-\right)^2\right)^2 + \frac{2\sigma^2\lambda}{n}(Z_i^+ - Z_i^-)^2 + \frac{4\max(Z_i^+, Z_i^-)^2}{n}\right\}. \tag{A.76}
$$

It follows that if $|Z_i^+| \neq |Z_i^-|$, take $\lambda = \frac{\sqrt{2\eta}}{|\left(Z_i^+\right)^2 - \left(Z_i^-\right)^2|}$

$$
\Psi^{*-1}_{\tilde{G}_i|Z_i^\pm}(\eta) \leq B_{Z_i^\pm, n}(\eta), \tag{A.77}
$$

and if $Z_i^+ = Z_i^-$, take $\lambda \to +\infty$,

$$\Psi^{*-1}_{\tilde{G}_i|Z_i^\pm}(\eta) \leq \frac{4\max\left(\left(Z_1^+\right)^2, \left(Z_i^-\right)^2\right)}{n}, \tag{A.78}$$

and if $Z_i^+ = -Z_i^- \neq 0$, take $\lambda = \frac{1}{2\sigma|Z_i^+|}\sqrt{\frac{n\eta}{2}}$,

$$\Psi^{*-1}_{\tilde{G}_i|Z_i^\pm}(\eta) \leq 4\sigma\sqrt{\frac{2\eta}{n}}|Z_i^+| + \frac{4\left(Z_1^+\right)^2}{n}. \tag{A.79}$$

□

Lemma 13. *The function* $\ln\cosh(x)$ *is lower bounded as*

$$\ln\cosh(x) \geq \min\left(1, \frac{|x|}{2}\right)\frac{|x|}{2}. \tag{A.80}$$

Proof. When $|x| \geq 2$, $\frac{1}{2}\left(e^x + e^{-x}\right) > \frac{1}{2}e^{|x|} = \frac{e^{|x|/2}}{2}e^{|x|/2} > e^{|x|/2}$. Take $\ln$ on both, $\ln\cosh(x) \geq \frac{|x|}{2}, \quad |x| \in [2, \infty)$. When $|x| \leq 2$, it is straightforward to verify by calculating derivatives that the function $\tanh(x) - \frac{x}{2}$ for $x \geq 0$ is increasing then decreasing. Since $\tanh(0) = 0$, $\ln\cosh(x) - \frac{x^2}{4}$, whose derivative is $\tanh(x) - \frac{x}{2}$, for $x \geq 0$ is increasing (then decreasing but is not needed here). Since $\ln\cosh(0) = 0$ and $\ln\cosh(2) - 1 > 0$, by the fact that $\ln\cosh(x) - \frac{x^2}{4}$ is an even function, we know $\ln\cosh(x) \geq \frac{x^2}{4}$ for any $|x| \in [0, 2]$. We then conclude the proof by combining both results. □

APPENDIX B

PROOFS FOR CHAPTER 3

B.1 Proofs for Chapter 3.2

We will need the following well known inequality frequently.

Lemma 14 (Hoeffding's inequality)**.** *Let $X_{1:n}$ be n independent random variables follow some σ^2-sub-Gaussian distribution with mean μ. Let $\hat{\mu}$ be their sample mean. Then the following inequalities hold*

$$\mathbb{P}\left(\hat{\mu} - \mu \geq \epsilon\right) \leq e^{-\frac{\epsilon^2 n}{2\sigma^2}}, \quad \mathbb{P}\left(\hat{\mu} - \mu \leq -\epsilon\right) \leq e^{-\frac{\epsilon^2 n}{2\sigma^2}}. \tag{B.1}$$

Lemma 15 (Restate Lemma 6)**.** *For any $m \geq 2$,* $\mathrm{Ent}(\sigma^2_{G^r}) \leq 8\ln(m)$.

Proof of Lemma 6. For any choice of $\sigma^2_{1:n}$. Let $s_j = \sum_{i \in G'_j} \sigma_i^2$ for each $i = 1, \ldots, k$. By the grouping property of entropy, we have

$$\mathrm{Ent}(\sigma^2_{G^r}) = \mathrm{Ent}(s_{1:k}) + \sum_{j=1}^{k} \frac{s_j}{\sum_{i=1}^{k} s_i} \mathrm{Ent}(\sigma^2_{G'_j}) \tag{B.2}$$

$$\leq \mathrm{Ent}(s_{1:k}) + \ln(2m), \tag{B.3}$$

where the inequality is due to the principal of maximum entropy.

For $j = 1, \ldots, k$, if $|G'_j| > 0$, we have $2^{j-1} \leq s_j/\underline{\sigma}^2 < 2m2^j$, otherwise $s_j = 0$. Without loss of generality, assume $\underline{\sigma}^2 = 1$ and $s_k > 0$. Let $s_{1:k}$ be the assignment with the largest entropy $\mathrm{Ent}(s_{1:k})$. If there are only $2m$ non-zero $s_{1:k}$, we have $\mathrm{Ent}(s_{1:k}) \leq \ln(2m)$ and the lemma is already proved. When there are more than $2m$ non-zero $s_{1:k}$, we have

$$\sum_{j=1}^{k-2m+1} s_j \leq 2m \sum_{j=1}^{k-2m+1} 2^j = 4m(2^{k-2m+1} - 1) < 4m2^{k-2m+1}, \tag{B.4}$$

and $s_k \geq 2^{k-1}$. It follows that

$$\sum_{j=1}^{k-2m+1} s_j = \frac{\sum_{j=1}^{k-2m+1} s_j}{\sum_{i=k-2m+2}^{k} s_i + \sum_{j=1}^{k-2m+1} s_j} \sum_{j=1}^{k} s_j \tag{B.5}$$

$$\leq \frac{\sum_{j=1}^{k-2m+1} s_j}{s_k + \sum_{j=1}^{k-2m+1} s_j} \sum_{j=1}^{k} s_j < \frac{4m2^{k-2m+1}}{2^{k-1} + 4m2^{k-2m+1}} \sum_{j=1}^{k} s_j \tag{B.6}$$

$$= \frac{4m2^{-2m+2}}{1 + 4m2^{-2m+2}} \sum_{j=1}^{k} s_j. \tag{B.7}$$

We can then write

$$\mathrm{Ent}(s_{1:k}) = \mathrm{Ent}(\sum_{j=1}^{k-2m+1} s_j, s_{k-2m+2:k}) + \frac{\sum_{j=1}^{k-2m+1} s_j}{\sum_{j=1}^{k} s_j} \mathrm{Ent}(s_{1:k-2m+1}) \tag{B.8}$$

$$\leq \ln(2m) + \frac{4m2^{-2m+2}}{1 + 4m2^{-2m+2}} \mathrm{Ent}(s_{1:k}), \tag{B.9}$$

where the equality is by the grouping property of entropy function, and the inequality is by $\mathrm{Ent}(s_{1:k-2m+1}) \leq \mathrm{Ent}(s_{1:k})$ since $s_{1:k}$ is the optimal assignment in terms of the largest entropy with k subsets, thus assignment $s_{1:k-2m+1}$ has smaller entropy. It implies $\mathrm{Ent}(s_{1:k}) \leq (1 + 4m2^{-2m+2}) \ln(2m) \leq 3\ln(2m)$. We thus have $\mathrm{Ent}(\sigma^2_{G^r}) \leq 4\ln(2m) \leq 8\ln(m)$. □

B.2 Proofs for Chapter 3.3

Lemma 16 (Restate Lemma 7). *Let $\omega_i = \delta \frac{\sigma_i^2}{\sum_{j=1}^n \sigma_j^2}$, the weighted naive elimination algorithm takes*

$$8 \sum_{i \in [n]} \frac{\sigma_i^2}{\epsilon^2} \left(\ln \frac{1}{\delta} + \mathrm{Ent}(\sigma^2_{1:n}) \right) \tag{B.10}$$

samples, and solves the (ϵ, δ) top-m arm identification problem for any $\epsilon > 0$ and $0 < \delta < 1$.

Proof of Lemma 7. The stopping time is clearly

$$\sum_{i=1}^{n} \frac{2\sigma_i^2}{(\epsilon/2)^2} \ln \frac{1}{\omega_i} = 8 \frac{\sum_{i=1}^n \sigma_i^2}{\epsilon^2} \left(\ln \frac{1}{\delta} + \mathrm{Ent}(\sigma^2_{1:n}) \right). \tag{B.11}$$

After the arms have been pulled and the reward observations collected, by Hoeffding's inequality (Lemma 14), we have $\mathbb{P}(\hat{\mu}_i \leq \mu_i - \epsilon/2) \leq \omega_i$ for any $i \in [m]$ and $\mathbb{P}(\hat{\mu}_j \geq \mu_j + \epsilon/2) \leq \omega_j$ for any $j \in [n] \setminus [m]$. Since $\sum_{i \in [n]} \omega_j = \delta$, the union bound implies that the event $\mathcal{E} = \{\hat{\mu}_i > \mu_i - \epsilon/2, \forall i \in [m]\} \cap \{\hat{\mu}_j < \mu_j + \epsilon/2, \forall j \in [n] \setminus [m]\}$ occurs with probability at least $1 - \delta$.

Suppose event $\mathcal{E}$ occurs. Consider a threshold $\mu_m - \epsilon/2$. Firstly, for any $i \in [m]$, $\hat{\mu}_i > \mu_i - \epsilon/2 \geq \mu_m - \epsilon/2$. In addition, any $j \in [n]/[m]$ with $\hat{\mu}_j > \mu_m - \epsilon/2$ must satisfy $\mu_j + \epsilon/2 > \hat{\mu}_j > \mu_m - \epsilon/2$, which implies $\mu_j > \mu_m - \epsilon$, i.e., the j-th arm is ϵ-approximate top-m. In other words, any arm with a sample mean greater than the threshold $\mu_m - \epsilon/2$ must be ϵ-approximate top-m. Since there are at least m arms with sample means greater than $\mu_m - \epsilon/2$, the m selected arms must be ϵ-approximate top-m. □

Lemma 17 (Restate Lemma 8). *For any $\sigma^2_{1:n}$, if $\max_{i \in [n]} \sigma_i^2 / \min_{j \in [n]} \sigma_j^2 \leq 2$, the* MedElim *algorithm has an expected stopping time*

$$O\left(\frac{\sum_{i \in [n]} \sigma_i^2}{\epsilon^2}\left(\ln\frac{1}{\delta} + \ln(m)\right)\right). \tag{B.12}$$

Moreover, for any $m' \leq m$, the MedElim algorithm satisfies the $(\epsilon, \frac{m'}{m}\delta)$ top-m' condition.

Proof of Lemma 8. We study the stopping time and accuracy separately.

Stopping time analysis: Recall that $\overline{r} = \frac{\max_{i \in [n]} \sigma_i^2}{\min_{j \in [n]} \sigma_j^2}$. It is clear that the size of the candidate set $\mathcal{S}_\ell$ decreases as $|\mathcal{S}_\ell| \leq \frac{n}{2^{\ell-1}}$. The sum of variances in the candidate set $\mathcal{S}_\ell$ decreases as follows

$$\frac{\sum_{i \in \mathcal{S}_\ell} \sigma_i^2}{\sum_{j \in [n]} \sigma_j^2} \leq \frac{\sum_{i \in \mathcal{S}_\ell} \overline{r}\underline{\sigma}^2}{\sum_{j \in [n]} \underline{\sigma}^2} \leq \overline{r}\frac{|\mathcal{S}_\ell|}{n} \leq \frac{\overline{r}}{2^{\ell-1}}. \tag{B.13}$$

This implies that

$$\frac{\sum_{i \in S_\ell} \sigma_i^2}{(\epsilon_\ell/2)^2} = 36\frac{16^\ell}{9^\ell}\frac{\sum_{i \in S_\ell} \sigma_i^2}{\epsilon^2} \leq 72\overline{r}\frac{8^\ell}{9^\ell}\frac{\sum_{i=1}^n \sigma_i^2}{\epsilon^2}. \tag{B.14}$$

The (random) total number of samples is thus upper bounded by

$$\sum_{\ell=1}^{\infty}\sum_{i\in S_\ell} t_{i,\ell} = \sum_{\ell=1}^{\infty} \frac{2\sum_{i\in S_\ell}\sigma_i^2}{(\epsilon_\ell/2)^2}\ln\left(\frac{m}{\delta_\ell}\right) \tag{B.15}$$

$$\le \bar{r}\frac{144\sum_{i=1}^{n}\sigma_i^2}{\epsilon^2}\sum_{\ell=1}^{\infty}\frac{8^\ell}{9^\ell}\left(\ell\ln(2)+\ln\frac{4m}{\delta}\right) \tag{B.16}$$

$$= O\left(\bar{r}\frac{\sum_{i=1}^{n}\sigma_i^2}{\epsilon^2}\left(\ln\frac{1}{\delta}+\ln(m)\right)\right), \tag{B.17}$$

with probability one. Thus the expected stopping time is of order $O\left(\tilde{r}\frac{\sum_{i\in[n]}\sigma_i^2}{\epsilon^2}\left(\ln\frac{1}{\delta}+\ln(m)\right)\right)$.

Accuracy analysis. Take an arbitrary $\ell \ge 1$ with $|\mathcal{S}_\ell| > 2m$. Fix some $m' \le m$. Let $1_\ell, 2_\ell, \dots, m'_\ell$ be the indices of the top-m' arms in S_ℓ obtained in iteration-$(\ell-1)$. For any $i \in [m']$, by Hoeffding's inequality (Lemma 14), we have $\mathbb{P}(\hat{\mu}_{i_\ell,\ell} > \mu_{i_\ell} - \epsilon_\ell/2) \ge 1 - \frac{1}{m}\delta_\ell$. Define the event $\mathcal{E}_\ell = \{\forall i \in [m'],\ \hat{\mu}_{i_\ell,\ell} > \mu_{i_\ell} - \epsilon_\ell/2\}$. By applying the union bound over $i \in [m']$, it is straightforward to verify that $\mathbb{P}(\mathcal{E}_\ell) \ge 1 - \frac{m'}{m}\delta_\ell$.

Conditioned on event $\mathcal{E}_\ell$ occurring, consider a threshold $\mu_{m'_\ell} - \epsilon_\ell/2$. It is clear that for any $i \in [m']$, $\hat{\mu}_{i_\ell,\ell} > \mu_{i_\ell} - \epsilon/2 \ge \mu_{m'_\ell} - \epsilon/2$. Thus any arm in $\{1_\ell, \dots, m'_\ell\}$ has an empirical mean greater than the threshold $\mu_{m'_\ell} - \epsilon_\ell/2$. In iteration-$\ell$, $|\mathcal{S}_{\ell+1}|$ arms with the largest empirical means are selected from set $\mathcal{S}_\ell$.

- If the selected arm with the smallest sample mean $\min\{\hat{\mu}_{i,\ell} :\ i \in \mathcal{S}_{\ell+1}\}$ is less than or equal to the threshold, then all the arms in $\{1_\ell, \dots, m'_\ell\}$ must be selected and they are still the top-m' arms within $\mathcal{S}_{\ell+1}$. It implies that $\mu_{m'_{\ell+1}} = \mu_{m'_\ell} > \mu_{m'_\ell} - \epsilon_\ell$.

- On the other hand, if the selected arm with the smallest sample mean is greater than the threshold, some arms in $\{1_\ell, \dots, m'_\ell\}$ may not be selected. Define the set of bad arms $B_\ell := \{i \in \mathcal{S}_\ell :\ \mu_i < \mu_{m'_\ell} - \epsilon_\ell\}$. A bad arm will be selected only if its empirical mean is greater than the threshold. Denote the set of bad arms with such overestimated sample means as $N_{m',\ell} = \{j \in B_\ell :\ \hat{\mu}_{j,\ell} > \mu_{m'_\ell} - \epsilon_\ell/2\}$. Then there are at most $|N_{m',\ell}|$ bad arms in $\mathcal{S}_{\ell+1}$. If $|N_{m',\ell}| \le |\mathcal{S}_{\ell+1}| - m'$, at least m' good arms remain in $\mathcal{S}_{\ell+1}$, which guarantees $\mu_{m'_{\ell+1}} \ge \mu_{m'_\ell} - \epsilon_\ell$.

These two situations indicate that conditioned on $\mathcal{E}_\ell$, $|N_{m',\ell}| \leq |\mathcal{S}_{\ell+1}| - m'$ implies $\mu_{m'_{\ell+1}} \geq \mu_{m'_\ell} - \epsilon_\ell$. It follows that

$$\begin{aligned}\mathbb{P}\left(\mu_{m'_{\ell+1}} < \mu_{m'_\ell} - \epsilon_\ell | \mathcal{E}_\ell\right) &\leq \mathbb{P}\left(|N_{m',\ell}| \geq |S_{\ell+1}| - m' + 1 | \mathcal{E}_\ell\right) \\ &\leq \frac{\mathbb{E}[|N_{m',\ell}||\mathcal{E}_\ell]}{|S_{\ell+1}| - i + 1}.\end{aligned}$$

where the second inequality is due to Markov inequality. The expectation can be bounded by

$$\begin{aligned}\mathbb{E}[|N_{m',\ell}||\mathcal{E}_\ell] &= \sum_{j\in B_\ell} \mathbb{P}\left(\hat{\mu}_{j,\ell} > \mu_{m'_\ell} - \epsilon_\ell/2|\mathcal{E}_\ell\right) \\ &= \sum_{j\in B_\ell} \mathbb{P}\left(\hat{\mu}_{j,\ell} > \mu_{m'_\ell} - \epsilon_\ell/2\right) \\ &\leq \sum_{j\in B_\ell} \mathbb{P}\left(\hat{\mu}_{j,\ell} > \mu_j + \epsilon_\ell/2\right) \\ &\leq (|S_\ell| - m')\frac{\delta_\ell}{m},\end{aligned}$$

where the equality is because $\mathcal{E}_\ell$ is defined by the samples of arms in $[1_\ell, \ldots, m'_\ell]$ which are independent from the samples of arms in B_ℓ, the first inequality is by $\mu_{m'_\ell} > \mu_j$ for $j \in B_\ell$, and the last inequality is by applying Hoeffding's inequality to each $\hat{\mu}_{j,l}, j \in B_\ell$ and $|B_\ell| \leq |\mathcal{S}_\ell| - m'$. We thus have

$$\begin{aligned}\mathbb{P}\left(\mu_{m'_{\ell+1}} < \mu_{m'_\ell} - \epsilon_\ell|\mathcal{E}_\ell\right) &\leq \frac{\delta_\ell}{m}\frac{|S_\ell| - m'}{|S_{\ell+1}| - m' + 1} \\ &\leq \frac{\delta_\ell}{m}\frac{|\mathcal{S}_\ell| - m}{|\mathcal{S}_{\ell+1}| - m + 1} && \text{by } m' \leq m \\ &\leq \frac{\delta_\ell}{m}\frac{2|\mathcal{S}_{\ell+1}| + 1 - m}{|\mathcal{S}_{\ell+1}| - m + 1} && \text{by } |\mathcal{S}_\ell| \leq 2|\mathcal{S}_{\ell+1}| + 1 \\ &= \frac{\delta_\ell}{m}\left(2 + \frac{m-1}{|\mathcal{S}_{\ell+1}| - m + 1}\right) \\ &\leq \frac{\delta_\ell}{m}\left(2 + \frac{m-1}{2m - m + 1}\right) && \text{by } |\mathcal{S}_{\ell+1}| \geq 2m \\ &< \frac{3\delta_\ell}{m}.\end{aligned}$$

It follows that

$$\begin{aligned}\mathbb{P}\left(\mu_{m'_{\ell+1}} < \mu_{m'_{\ell}} - \epsilon_\ell\right) &= \mathbb{P}(\mathcal{E})\mathbb{P}\left(\mu_{m'_{\ell+1}} < \mu_{m'_{\ell}} - \epsilon_\ell|\mathcal{E}\right) + \mathbb{P}(\mathcal{E}^c)\mathbb{P}\left(\mu_{m'_{\ell+1}} < \mu_{m'_{\ell}} - \epsilon_\ell|\mathcal{E}^c\right)\\ &\le \mathbb{P}\left(\mu_{m'_{\ell+1}} < \mu_{m'_{\ell}} - \epsilon_\ell|\mathcal{E}\right) + \mathbb{P}(\mathcal{E}^c)\\ &\le \frac{3\delta_\ell}{m} + \frac{m'\delta_\ell}{m} \le \frac{4m'}{m}\delta_\ell.\end{aligned}$$

The argument above holds for any $\ell \geq 1$ with $|S_\ell| > 2m$. The parameters satisfy

$$\sum_{\ell=1}^{\infty} \epsilon_\ell = \frac{\epsilon}{3}\sum_{\ell=1}^{\infty}(3/4)^\ell = \epsilon, \qquad \sum_{\ell=1}^{\infty} 4\delta_\ell = \delta\sum_{\ell=1}^{\infty}(1/2)^\ell = \delta.$$

The returned arm set is $R = \mathcal{S}_{\ell^*}$ for certain ℓ^*, and thus with probability at least $1 - \frac{m'}{m}\delta$, the final returned arm set R satisfies

$$\begin{aligned}\max{}^{m'}_{i\in R}\mu_i &= \max{}^{m'}_{i\in\mathcal{S}_{\ell^*}}\mu_i\\ &\ge \max{}^{m'}_{i\in\mathcal{S}_{\ell^*-1}}\mu_i - \epsilon_{\ell^*-1}\\ &\ge \cdots\\ &\ge \max{}^{m'}_{i\in\mathcal{S}_1}\mu_i - \sum_{\ell=1}^{\ell^*-1}\epsilon_\ell\\ &> \max{}^{m'}_{i\in[n]}\mu_i - \epsilon.\end{aligned}$$

The proof is thus complete. □

Calculation in the illustrative example. Recall the illustrative example, where $\log(m) = k$ and $\log(n) = k^2$ for some integer $k \geq 2$ and $\ell = \lceil\log(k)\rceil$. Among these n arms, there are 2^i arms with the same variance 2^{-i} for each $i = 0, 1, \ldots, \ell - 1$, and the rest $n - \sum_{i=0}^{\ell-1} 2^i = 2^{k^2} - 2^\ell + 1$ arms have the same variance $2^{-k^2}\ell/k$. Then G^m is the set of arms with variances $2^{-k^2}\ell/k$, and G^l

is the set of arms with variances 2^{-i} for $i = 0, 1, \dots, \ell - 1$. It is seen that

$$\sum_{j \in G^m} \sigma_j^2 = (2^{k^2} - 2^{\ell} + 1) 2^{-k^2} \ell / k = \Theta(\ell / k), \tag{B.18}$$

$$\sum_{j \in G^l} \sigma_j^2 = \sum_{i=0}^{\ell-1} 2^i 2^{-i} = \ell = \Theta(\log(k)), \tag{B.19}$$

which implies $\sum_{j \in [n]} \sigma_j^2 = \Theta(\log(k))$. Furthermore, we can calculate that

$$\mathrm{Ent}(\sigma_{G^l}^2) = \sum_{i=0}^{\ell-1} \frac{2^i 2^{-i}}{\ell} \ln(2^i) = \frac{\ln(2)}{2}(\ell - 1) = \Theta(\ell) = \Theta(\log(k)). \tag{B.20}$$

Furthermore, we can calculate that

$$\sum_{j \in G^r} \sigma_j^2 = 2m 2^{-k^2} \ell / k + \sum_{j \in G^l} \sigma_j^2 = 2^{-k^2+1} \ell + \ell = \Theta(\ell) = \Theta(\log(k)). \tag{B.21}$$

Then the entropy values can be calculated as

$$\begin{aligned}
\mathrm{Ent}(\sigma_{G^r}^2) &= \frac{\sum_{j \in G^r / G^l} \sigma_j^2}{\sum_{j \in G^r} \sigma_j^2} \mathrm{Ent}(\sigma_{G^r/G^l}^2) + \frac{\sum_{j \in G^l} \sigma_j^2}{\sum_{j \in G^r} \sigma_j^2} \mathrm{Ent}(\sigma_{G^l}^2) && \text{(B.22)} \\
&= \frac{2^{-k^2+1} \ell}{\sum_{j \in G^r} \sigma_j^2} \ln(2m) + \frac{\ell}{\sum_{j \in G^r} \sigma_j^2} \mathrm{Ent}(\sigma_{G^l}^2) && \text{(B.23)} \\
&= \Theta\left(2^{-k^2} k + \mathrm{Ent}(\sigma_{G^l}^2)\right) && \text{(B.24)} \\
&= \Theta(\mathrm{Ent}(\sigma_{G^l}^2)) = \Theta(\log(k)), && \text{(B.25)}
\end{aligned}$$

and $\mathrm{Ent}(\sigma_{G^m}^2) = \Theta(k^2)$ implies

$$\begin{aligned}
\mathrm{Ent}(\sigma_{1:n}^2) &= \frac{\sum_{j \in G^m} \sigma_j^2}{\sum_{j \in [n]} \sigma_j^2} \mathrm{Ent}(\sigma_{G^m}^2) + \frac{\sum_{j \in G^l} \sigma_j^2}{\sum_{j \in [n]} \sigma_j^2} \mathrm{Ent}(\sigma_{G^l}^2) && \text{(B.26)} \\
&= \Theta\left(\frac{\ell / k}{\log(k)} k^2 + \log(k)\right) = \Theta(k). && \text{(B.27)}
\end{aligned}$$

B.3 A More Adaptive Median Elimination Algorithm

Let us sort $\sigma^2_{1:n}$ in decreasing order, and denote the sorted variances as $\tilde{\sigma}^2_{1:n}$. For each $\ell \geq 1$, define $h_\ell := \max\{j \geq m : \sum_{i\in[j]} \tilde{\sigma}^2_i \leq \frac{1}{2^{\ell-1}} \sum_{i\in[n]} \sigma^2_i\}$ if the set is not empty, otherwise $h_\ell = m$. Let $\ell^* := \min\{\ell \geq 1 : h_\ell = m\}$.

Define a ratio

$$\underline{r} := \min_{j\in[\ell^*-1]} \frac{h_{j+1}}{h_j} \tag{B.28}$$

```
Algorithm 7: Adapted-MedElim($\sigma^2_{1:n}, m, [n], \epsilon, \delta$)
sInitialize $S_1 = [n]$, $\ell = 1$ and $\epsilon_\ell = (\epsilon/3)\frac{3^\ell}{4^\ell},\ \delta_\ell = \frac{\underline{r}\delta}{2^\ell}$
for $\ell = 1, 2, \ldots, \ell^* - 1$ do
    Pull arm-$i$ $t_{i,\ell} = \frac{2\sigma_i^2}{(\epsilon_\ell/2)^2} \ln \frac{m}{\delta_\ell}$ times and calculate their sample mean $\hat{\mu}_{i,\ell}$ for each $i \in S_\ell$
    Update candidate set $S_{\ell+1} = \arg\max_{i\in S_\ell}^{1:h_{\ell+1}} \hat{\mu}_{i,\ell}$
Return: $S_{\ell^*}$
```

In the homogeneous setting, the MedElim algorithm halves the complexity of the problem if the candidate set is halved. However, it should be noted that in the heterogeneous setting, simply halving the candidate set may not be efficient since the complexity would depend on the sum of the variances, instead of the number of the candidate arms. We can instead aim to half the sum of the variances of the candidate set. This discrepancy is less pronounced when the heterogeneity is low, and thus the MedElim algorithm performs reasonably well in such cases.

Lemma 18. *The algorithm is valid and has an expected stopping time*

$$O\left(\sum_{i\in[n]} \frac{\sigma_i^2}{\epsilon^2}\left(\ln\frac{1}{\delta} + \ln(m) + \ln\frac{1}{\underline{r}}\right)\right). \tag{B.29}$$

Proof of Lemma 18. We study the stopping time and accuracy separately.

Stopping time analysis: First, notice the sum of variances in the candidate set decreases as follows:

$$\sum_{i\in\mathcal{S}_\ell}\sigma_i^2 = \frac{\sum_{i\in\mathcal{S}_\ell}\sigma_i^2}{\sum_{i\in[n]}\sigma_i^2}\sum_{i\in[n]}\sigma_i^2 \le \frac{\sum_{i\in[h_\ell]}\tilde{\sigma}_i^2}{\sum_{i\in[n]}\sigma_i^2}\sum_{i\in[n]}\sigma_i^2 \le \frac{1}{2^{\ell-1}}\sum_{i\in[n]}^{n}\sigma_i^2. \tag{B.30}$$

This implies that

$$\frac{\sum_{i\in S_\ell}\sigma_i^2}{(\epsilon_\ell/2)^2} = 36\frac{16^\ell}{9^\ell}\frac{\sum_{i\in S_\ell}\sigma_i^2}{\epsilon^2} \le 72\overline{r}\frac{8^\ell}{9^\ell}\frac{\sum_{i=1}^n\sigma_i^2}{\epsilon^2}. \tag{B.31}$$

The stopping time is thus upper bounded by

$$\sum_{\ell=1}^{\infty}\sum_{i\in S_\ell} t_{i,\ell} = \sum_{\ell=1}^{\infty}\frac{2\sum_{i\in S_\ell}\sigma_i^2}{(\epsilon_\ell/2)^2}\ln\left(\frac{m}{\delta_\ell}\right) \tag{B.32}$$

$$\le \frac{144\sum_{i=1}^n\sigma_i^2}{\epsilon^2}\sum_{\ell=1}^{\infty}\frac{8^\ell}{9^\ell}\left(\ell\ln(2)+\ln\frac{m}{\delta}+\ln\frac{1}{\underline{r}}\right) \tag{B.33}$$

$$= O\left(\frac{\sum_{i=1}^n\sigma_i^2}{\epsilon^2}\left(\ln\frac{1}{\delta}+\ln(m)+\ln\frac{1}{\underline{r}}\right)\right). \tag{B.34}$$

The expected stopping time is of order $O\left(\frac{\sum_{i=1}^n\sigma_i^2}{\epsilon^2}\left(\ln\frac{1}{\delta}+\ln(m)+\ln\frac{1}{\underline{r}}\right)\right)$.

Accuracy analysis. Take an arbitrary $\ell \in [\ell^*-1]$, and it is clear that $|\mathcal{S}_\ell| = h_\ell > m$. Fix some $m' \le m$. Let $1_\ell, 2_\ell, \dots, m'_\ell$ be the indices of the top-m' arms in S_ℓ, respectively. For any $i \in [m']$, by Hoeffding's inequality (Lemma 14), we have $\mathbb{P}(\hat{\mu}_{i_\ell,\ell} > \mu_{i_\ell} - \epsilon_\ell/2) \ge 1 - \frac{1}{m}\delta_\ell$. Define the event $\mathcal{E}_\ell = \{\forall i \in [m'],\ \hat{\mu}_{i_\ell,\ell} > \mu_{i_\ell} - \epsilon_\ell/2\}$. By applying the union bound over $i \in [m']$, it is straightforward to verify that $\mathbb{P}(\mathcal{E}_\ell) \ge 1 - \frac{m'}{m}\delta_\ell$.

Conditioned on the event $\mathcal{E}_\ell$ occurring, consider a threshold $\mu_{m'_\ell} - \epsilon_\ell/2$. It is clear that for any $i \in [m']$, $\hat{\mu}_{i_\ell,\ell} > \mu_{i_\ell} - \epsilon/2 \ge \mu_{m'_\ell} - \epsilon/2$. Thus any arm in $\{1_\ell, \dots, m'_\ell\}$ has empirical mean greater than the threshold $\mu_{m'_\ell} - \epsilon_\ell/2$. $|\mathcal{S}_{\ell+1}| = h_{\ell+1}$ arms with the largest sample means are selected from set $\mathcal{S}_\ell$.

- If the smallest selected sample mean $\min\{\hat{\mu}_{i,\ell} : i \in \mathcal{S}_{\ell+1}\}$ is less or equal to the threshold, all arms in $\{1_\ell, \dots, m'_\ell\}$ must be selected and they are still top-m' arms within $\mathcal{S}_{\ell+1}$. It implies that $\mu_{m'_{\ell+1}} = \mu_{m'_\ell} > \mu_{m'_\ell} - \epsilon_\ell$.

- On the other hand, if the smallest selected sample mean is greater than the threshold, some arms in $\{1_\ell, \ldots, m'_\ell\}$ may not be selected. Define the set of bad arms $B_\ell := \{i \in \mathcal{S}_\ell : \ \mu_i < \mu_{m'_\ell} - \epsilon_\ell\}$. A bad arm can be selected only if its empirical mean is greater than the threshold. Define the set of such overestimated bad arms as $N_{m',\ell} = \{j \in B_\ell : \ \hat{\mu}_{j,\ell} > \mu_{m'_\ell} - \epsilon_\ell/2\}$. Then there are at most $|N_{m',\ell}|$ bad arms in $\mathcal{S}_{\ell+1}$. If $|N_{m',\ell}| \leq |\mathcal{S}_{\ell+1}| - m'$, at least m' good arms remain in $\mathcal{S}_{\ell+1}$, which guarantees $\mu_{m'_{\ell+1}} \geq \mu_{m'_\ell} - \epsilon_\ell$.

These two situations indicate that $|N_{m',\ell}| \leq |\mathcal{S}_{\ell+1}| - m'$ implies $\mu_{m'_{\ell+1}} \geq \mu_{m'_\ell} - \epsilon_\ell$ conditioned on $\mathcal{E}_\ell$. It follows that

$$\begin{aligned}\mathbb{P}\left(\mu_{m'_{\ell+1}} < \mu_{m'_\ell} - \epsilon_\ell | \mathcal{E}_\ell\right) &\leq \mathbb{P}\left(|N_{m',\ell}| \geq |S_{\ell+1}| - m' + 1 | \mathcal{E}_\ell\right) \\ &\leq \frac{\mathbb{E}[|N_{m',\ell}||\mathcal{E}_\ell]}{|S_{\ell+1}| - i + 1}.\end{aligned}$$

where the second inequality is by Markov inequality. The expectation can be bounded by

$$\mathbb{E}[|N_{m',\ell}||\mathcal{E}_\ell] = \sum_{j \in B_\ell} \mathbb{P}\left(\hat{\mu}_{j,\ell} > \mu_{m'_\ell} - \epsilon_\ell/2 | \mathcal{E}_\ell\right) \leq (|S_\ell| - m')\frac{\delta_\ell}{m},$$

where the inequality is by Hoeffding's inequality and $|B_\ell| \leq |\mathcal{S}_\ell| - m'$. We thus have

$$\begin{aligned}\mathbb{P}\left(\mu_{m'_{\ell+1}} < \mu_{m'_\ell} - \epsilon_\ell | \mathcal{E}_\ell\right) &\leq \frac{\delta_\ell}{m} \frac{|S_\ell| - m'}{|S_{\ell+1}| - m' + 1} \\ &= \frac{\delta_\ell}{m} \frac{h_\ell - m'}{h_{\ell+1} - m' + 1} \\ &\leq \frac{\delta_\ell}{m} \frac{h_{\ell+1}/\underline{r} - m'}{h_{\ell+1} - m' + 1} \qquad \text{by } h_\ell \leq \frac{1}{\underline{r}} h_{\ell+1} \\ &= \frac{\delta_\ell}{m}\left(\frac{1}{\underline{r}} + \frac{(1/\underline{r} - 1)m' - 1/\underline{r}}{h_{\ell+1} - m' + 1}\right) \\ &\leq \frac{\delta_\ell}{m}\left(1/\underline{r} + (1/\underline{r} - 1)m' - 1/\underline{r}\right) \qquad \text{by } h_{\ell+1} \geq m \geq m' \\ &= \frac{m'\delta_\ell}{m}(1/\underline{r} - 1).\end{aligned}$$

It follows that

$$\begin{aligned}\mathbb{P}\left(\mu_{m'_{\ell+1}} < \mu_{m'_\ell} - \epsilon_\ell\right) &= \mathbb{P}(\mathcal{E})\mathbb{P}\left(\mu_{m'_{\ell+1}} < \mu_{m'_\ell} - \epsilon_\ell | \mathcal{E}\right) + \mathbb{P}(\mathcal{E}^c)\mathbb{P}\left(\mu_{m'_{\ell+1}} < \mu_{m'_\ell} - \epsilon_\ell | \mathcal{E}^c\right) \\ &\leq \mathbb{P}\left(\mu_{m'_{\ell+1}} < \mu_{m'_\ell} - \epsilon_\ell | \mathcal{E}\right) + \mathbb{P}(\mathcal{E}^c) \\ &\leq \frac{m'\delta_\ell}{m}(1/\underline{r} - 1) + \frac{m'\delta_\ell}{m} = \frac{1}{\underline{r}}\frac{m'}{m}\delta_\ell.\end{aligned}$$

The argument above holds for any $\ell \geq 1$ with $|S_\ell| > 2m$. The parameters satisfy

$$\sum_{\ell=1}^{\infty} \epsilon_\ell = \frac{\epsilon}{3}\sum_{\ell=1}^{\infty}(3/4)^\ell = \epsilon, \qquad \sum_{\ell=1}^{\infty}\frac{1}{\underline{r}}\delta_\ell = \delta\sum_{\ell=1}^{\infty}(1/2)^\ell = \delta.$$

The returned arm set $R = \mathcal{S}_{\ell^*}$ for some ℓ^*. With probability at least $1 - \frac{m'}{m}\delta$, the final returned arm set R satisfies

$$\begin{aligned}\max{}_{i\in R}^{m'}\mu_i &= \max{}_{i\in\mathcal{S}_{\ell^*}}^{m'}\mu_i \\ &\geq \max{}_{i\in\mathcal{S}_{\ell^*-1}}^{m'}\mu_i - \epsilon_{\ell^*-1} \\ &\geq \cdots \\ &\geq \max{}_{i\in\mathcal{S}_1}^{m'}\mu_i - \sum_{\ell=1}^{\ell^*-1}\epsilon_\ell \\ &> \max{}_{i\in[n]}^{m'}\mu_i - \epsilon.\end{aligned}$$

□

B.4 Lower Bound Proofs for Chapter 3.4

Define $\mathcal{I}(\sigma_{1:n}^2) := \{(\mu_{1:n}, \sigma_{1:n}^2) : \ \mu_{1:n} \in \mathbb{R}^n\}$. When $\sigma_{1:n}^2$ is obvious in the context, we simply write $\mathcal{I}(\sigma_{1:n}^2)$ as $\mathcal{I}$. The sample complexity of the approximate top-m identification problem under algorithm inputs $(\epsilon, \delta, m, [n], \sigma_{1:n}^2)$ is

$$\mathrm{SC}(\epsilon, \delta, m, [n], \sigma_{1:n}^2) := \inf_{\mathsf{A}} \sup_{I\in\mathcal{I}(\sigma_{1:n}^2)} \mathbb{E}_I[T^{\mathsf{A}}], \tag{B.35}$$

where the infimum is taken over all valid algorithms, the supreme is taken over the instance class $\mathcal{I}(\sigma^2_{1:n}) := \{(\mu_{1:n}, \sigma^2_{1:n}) : \ \mu_{1:n} \in \mathbb{R}^n\}$, and the subscript I in the expectation $\mathbb{E}_I[\cdot]$ indicates that it is with respect to bandit model I.

Lemma 19 (Restate Lemma 9)**.** *For any two probability measure P, Q on the same measurable space $(\Omega, \mathcal{F})$, if $\mathcal{E} \in \mathcal{F}$ with $P(\mathcal{E}) \geq 1-\delta > Q(\mathcal{E})$, we have*

$$Q(\mathcal{E}) \geq B(\delta) e^{-\frac{D(P||Q)}{1-\delta}}, \tag{B.36}$$

where $D(\cdot||\cdot)$ is the Kullback-Leibler divergence and $B(\delta) = e^{-\frac{\mathrm{Ent}(\delta, 1-\delta)}{1-\delta}}$ is a strictly decreasing function with $B(0.1) > 0.69$.

Proof of Lemma 9. Let $D_b(p,q) = p \ln \frac{p}{q} + (1-p) \ln \frac{1-p}{1-q}$ be the binary KL-divergence. Since $P(\mathcal{E}) \geq 1-\delta$, by the data processing inequality for the KL-divergence, we have

$$\begin{aligned} D(P||Q) &\geq D_b(P(\mathcal{E}), Q(\mathcal{E})) \geq D_b(1-\delta, Q(\mathcal{E})) && \text{(B.37)} \\ &> (1-\delta) \ln \frac{1-\delta}{Q(\mathcal{E})} + \delta \ln \delta \geq (1-\delta) \ln \frac{B(\delta)}{Q(\mathcal{E})}, && \text{(B.38)} \end{aligned}$$

where the second inequality is due to $P(\mathcal{E}) \geq 1-\delta > Q(\mathcal{E})$, and the fact that $D_b(p,q)$ is monotonically increasing in p in the range $[q, 1]$ for any fixed q. We thus concludes that $Q(\mathcal{E}) \geq B(\delta) e^{-\frac{D(P||Q)}{1-\delta}}$. □

Lemma 20 (Restate Lemma 10)**.** *For $\epsilon > 0$, $\delta < 0.25$, $m < n/2$, $(\sigma_i^2)_{i \in [n]}$, $\mathrm{SC}(\epsilon, \delta, m, [n], \sigma^2_{1:n}) \geq \frac{1-\delta}{2\epsilon^2} v^*$, where v^* is the optimal value of the following optimization problem:*

$$\textit{maximize:} \quad \sum_{M \subset [n] : |M| = m} \left(\sum_{l \in M} \eta_{M \setminus \{l\}} \sigma_l^2 \right) \left(\ln \frac{B(\delta)}{\delta} + \mathrm{Ent}(\{\eta_{M \setminus \{l\}} \sigma_l^2\}_{l \in M}) \right) \tag{B.39}$$

$$\textit{subject to:} \quad \sum_{F \subset [n] : |F| = m-1} \eta_F = 1, \quad \eta_F \geq 0, \ \forall F \subset [n], |F| = m-1. \tag{B.40}$$

Proof of Lemma 10. We have shown in Chapter 3.4 that $\mathrm{SC}(\epsilon, \delta, m, [n], \sigma^2_{1:n})$ is lower bounded by the optimal value of the following optimization problem:

$$\text{minimize:} \qquad \max_{F\subset[n]:|F|=m-1,\ l\notin F} \sum_{j\notin F\cup\{l\}} t_{l,F,j} \tag{B.41}$$

$$\text{subject to:} \qquad \sum_{i\in M} \exp\left(-t_{l,M\setminus\{i\},i}/\theta_i\right) \le \delta', \quad \forall M\subset[n], |M|=m,\ \forall l\notin M, \tag{B.42}$$

where $\theta_i = \frac{(1-\delta)\sigma_i^2}{2\epsilon^2}, \forall i \in [n]$ and $\delta' = \frac{\delta}{B(\delta)}$. This problem is equivalent to the following convex optimization.

$$\min_{t,\tau} \qquad \tau \tag{B.43}$$

$$\text{s.t.} \qquad \sum_{j\notin F\cup\{l\}} t_{l,F,j} \le \tau, \quad \forall F\subset[n]\setminus\{l\} : |F| = m-1, \forall l\in[n] \tag{B.44}$$

$$\sum_{i\in M} \exp\left(-t_{l,M\setminus\{i\},i}/\theta_i\right) \le \delta', \quad \forall M\subset[n]\setminus\{l\} : |M| = m, \forall l\in[n]. \tag{B.45}$$

For simplicity, we use notation $\sum_{l,F}$ and $\sum_{l,M}$ to indicate $\sum_{l\in[n]}\sum_{F\subset[n]\setminus\{i\}:|F|=m-1}$ and $\sum_{l\in[n]}\sum_{M\subset[n]\setminus\{i\}:|M|=m}$, respectively. The Lagrangian of the optimization problem above is

$$L(t,\tau,\eta,\lambda) = \tau + \sum_{l,F}\eta_{l,F}\left(\sum_{j\notin F\cup\{l\}} t_{l,F,j} - \tau\right) + \sum_{l,M}\lambda_{l,M}\left(\sum_{i\in M}\exp\left(-t_{l,M\setminus\{i\},i}/\theta_i\right) - \delta'\right) \tag{B.46}$$

It is straightforward to check the optimization problem satisfies Slater's condition by assigning large enough $t_{l,F,j}$ and τ values. Since the optimization problem is convex, the optimal value equals to $\sup_{\eta,\lambda}\inf_{t,\tau} L(t,\tau,\eta,\lambda)$ according to the strong duality. For the saddle point, we must have $\sum_{l,F}\eta_{l,F} = 1$, or else $\inf_{t,\tau} L(t,\tau,\eta,\lambda) = -\infty$. Decision variable τ can thus be omitted. Let $L(t,\eta,\lambda) = L(t,\tau,\eta,\lambda)$ by restricting $\sum_{l,F}\eta_{l,F} = 1$. The derivative can be calculated that

$$\frac{\mathrm{d}L(t,\eta,\lambda)}{\mathrm{d}t_{l,F,i}} = \eta_{l,F} - \frac{\lambda_{l,F\cup\{i\}}}{\theta_i}\exp(-t_{l,F,i}/\theta_i). \tag{B.47}$$

It implies that when $\eta_{l,F} > 0$ and $\lambda_{l,F\cup\{i\}} > 0$, $t_{l,F,i} = \theta_i \ln \frac{\lambda_{l,F\cup\{i\}}}{\eta_{l,F}\theta_i}$. Define $\ln(0) = -\infty$ and let $0 \cdot \infty = 0$. The extended real valued function $g(\eta, \lambda)$ for $\sum_{l,F} \eta_{l,F} = 1$, $\eta_{l,F} \geq 0$ and $\lambda_{l,M} \geq 0$, is

$$\begin{aligned} g(\eta,\lambda) := \inf_t L(t,\eta,\lambda) = \sum_{l,F} \eta_{l,F} \sum_{i \notin F\cup\{l\}} \theta_i \ln \lambda_{l,F\cup\{i\}} - \sum_{l,F} \eta_{l,F} \sum_{i\notin F\cup\{l\}} \theta_i \ln(\eta_{l,F}\theta_i) \\ + \sum_{l,M} \left(\sum_{i\in M} \eta_{l,M\setminus\{i\}}\theta_i - \delta' \lambda_{l,M} \right). \end{aligned} \tag{B.48}$$

This dual function has two set of variables, however one of them can be eliminated explicitly as follows. For fixed η's with $\sum_{l,F} \eta_{l,F} = 1$ and $\eta_{l,F} \geq 0$, the function is separable with respect to λ's, and thus we can maximize $g(\eta, \lambda)$ by optimizing each individual $\lambda_{l,M}$ separately. It is straightforward to verify that $\lambda_{l,M} = \left(\sum_{F,i:F\cup\{i\}=M} \eta_{l,F}\theta_i\right)/\delta'$.

Since η's, θ's and δ' are positive, the assignments of λ's are also positive, which satisfy the constraints in the dual program. Plug it into $g(\eta, \lambda)$, we have the induced objective as

$$g(\eta) = \sum_{l,F} \eta_{l,F} \sum_{i\notin F\cup\{l\}} \theta_i \ln \frac{\sum_{F',i':F'\cup\{i'\}=F\cup\{i\}} \eta_{l,F}\theta_i}{\eta_{l,F}\theta_i\delta'} \tag{B.49}$$

$$= \sum_F \sum_{i\notin F} \sum_{l\notin F\cup\{i\}} \eta_{l,F}\theta_i \ln \frac{\sum_{F',i':F'\cup\{i'\}=F\cup\{i\}} \eta_{l,F}\theta_i}{\eta_{l,F}\theta_i\delta'}. \tag{B.50}$$

and the dual variables η's lie in a probability simplex.

Further constraining the problem by requiring $\eta_F := (n-m)\eta_{l,F}$ for all $l \notin F$ reduces the number of dual variables, but does not change the fact that any valid assignment of η_F's will provide a lower bound to the original primal problem. The following restricted objective will be considered:

$$g(\eta) = \sum_F \sum_{i\notin F} \sum_{l\notin F\cup\{i\}} \frac{\eta_F}{n-m}\theta_i \ln \frac{\sum_{F',i':F'\cup\{i'\}=F\cup\{i\}} \eta_F\theta_i}{\eta_F\theta_i\delta'} \tag{B.51}$$

$$= \sum_F \sum_{i\notin F} \eta_F\theta_i \ln \frac{\sum_{F',i':F'\cup\{i'\}=F\cup\{i\}} \eta_F\theta_i}{\eta_F\theta_i\delta'}. \tag{B.52}$$

The optimal value of the optimization above is lower bounded by

$$\begin{aligned}
&\text{maximize} && \sum_{M\subset[n],|M|=m}\left(\left(\sum_{j\in M}\eta_{M\setminus\{j\}}\theta_j\right)\left(\mathrm{Ent}(\{\eta_{M\setminus\{j\}}\sigma_j^2\}_{j\in M})+\ln\frac{B(\delta)}{\delta}\right)\right) && \text{(B.53)}\\
&\text{subject to} && \sum_{F\subset[n]:|F|=m-1}\eta_F=1,\quad \eta_F\geq 0,\ \forall F\subset[n],|F|=m-1. && \text{(B.54)}
\end{aligned}$$

The lemma is proved. □

Recall that the optimization in (3.25) is

$$\begin{aligned}
&\text{maximize:} && \sum_{M\subset[n]:|M|=m}\left(\sum_{l\in M}\eta_{M\setminus\{l\}}\sigma_l^2\right)\mathrm{Ent}(\{\eta_{M\setminus\{l\}}\sigma_l^2\}_{l\in M}) && \text{(B.55)}\\
&\text{subject to:} && \sum_{F\subset[n]:|F|=m-1}\eta_F=1,\quad \eta_F\geq 0,\ \forall F\subset[n],|F|=m-1. && \text{(B.56)}
\end{aligned}$$

Lemma 21 (Restate Lemma 11)**.** *The optimal value of the optimization (B.55) is lower-bounded by* $\frac{1}{3}\sum_{j\in G^m}\sigma_j^2\ln(m)$.

Proof of Lemma 11. The objective function of equation (B.55) can be written as

$$\begin{aligned}
&\sum_{F\subset[n]:|F|=m-1}\sum_{i\notin F}\eta_F\sigma_i^2\ln\left(\frac{\sum_{F'\cup\{j\}=F\cup\{i\}}\eta_{F'}\sigma_j^2}{\eta_F\sigma_i}\right)\\
&=\sum_{i\in[n]}\sum_{F:i\notin F}\eta_F\sigma_i^2\ln\left(\frac{\sum_{F'\cup\{j\}=F\cup\{i\}}\eta_{F'}\sigma_j^2}{\eta_F\sigma_i}\right) \qquad \text{(B.57)}
\end{aligned}$$

For any $F\subset G^m$ with $|F|=m-1$, let $\eta_F=\frac{\prod_{i\in F}\sigma_i^2}{\sum_{F'\subset G^m:|F'|=m-1}\prod_{j\in F}\sigma_i^2}$; and for any $F\not\subset G^m$ with $|F|=m-1$, set $\eta_F=0$. In the following analysis F indicates subset of G^m with $|F|=m-1$ and E indicates subset of G^m with $|E|=m-2$. Items in (B.57) can be lower bounded as follows.

$$\begin{aligned}
&\ln(m)\sum_{i\in G^m}\sigma_i^2\sum_{F:i\notin F}\eta_F && \text{(B.58)}\\
&=\ln(m)\sum_{i\in G^m}\sigma_i^2\frac{\sum_{F:i\notin F}\prod_{l\in F}\sigma_l^2}{\sum_{F:i\in F}\prod_{l\in F}\sigma_l^2+\sum_{F:i\notin F}\prod_{l\in F}\sigma_l^2} && \text{(B.59)}
\end{aligned}$$

$$= \ln(m) \sum_i \sigma_i^2 \left(\frac{\sum_{F:i\in F} \prod_{l\in F} \sigma_l^2}{\sum_{F:i\notin F} \prod_{l\in F} \sigma_l^2} + 1 \right)^{-1} \tag{B.60}$$

$$= \ln(m) \sum_{i\in G^m} \sigma_i^2 \left((m-1) \frac{\sum_{F:i\in F} \prod_{l\in F} \sigma_l^2}{(m-1)\sum_{F:i\notin F} \prod_{l\in F} \sigma_l^2} + 1 \right)^{-1} \tag{B.61}$$

$$= \ln(m) \sum_{i\in G^m} \sigma_i^2 \left((m-1)\sigma_i^2 \frac{\sum_{E:i\notin E} \prod_{l\in E} \sigma_l^2}{\sum_{E:i\notin E} \prod_{l\in E} \sigma_l^2 \left(\sum_{j\in G^m \setminus E \setminus \{i\}} \sigma_j^2 \right)} + 1 \right)^{-1} \tag{B.62}$$

$$\geq \ln(m) \sum_{i\in G^m} \sigma_i^2 \left((m-1)\sigma_i^2 \frac{\sum_{E} \prod_{l\in E} \sigma_l^2}{\sum_{E} \prod_{l\in E} \sigma_l^2 \left(\min_{F:i\in F} \sum_{j\in G^m \setminus F} \sigma_j^2 \right)} + 1 \right)^{-1} \tag{B.63}$$

$$= \ln(m) \sum_{i\in G^m} \sigma_i^2 \left(\frac{(m-1)\sigma_i^2}{\sum_{j\in G^m} \sigma_j^2 - \max_{F:i\in F} \sum_{l\in F} \sigma_l^2} + 1 \right)^{-1}, \tag{B.64}$$

where the last inequality is by $\sum_{j\in G^m\setminus E\setminus\{i\}} \sigma_j^2 \geq \min_{F:i\in F} \sum_{j\in G^m\setminus F} \sigma_j^2$ for any $\mathcal{E}$. Recall the definition of G^m: there are $G_{1:k}$ groups partitioning $[n]$ and $G^m = \cup_{j:|G_j|>2m} G_j$. Consider the group $G_{k'} \subset G^m$ with the largest index $k' \leq k$. Since the heterogeneity within group $G_{k'}$ is at most 2, we have $\max_{i\in G^m} \sigma_i^2 \leq 2\sigma_j^2$ for any $j \in G_{k'}$. Then for any $F \subset G^m$ and any $i \in G^m$,

$$\frac{(m-1)\sigma_i^2}{\sum_{j\in G^m} \sigma_j^2 - \sum_{l\in F} \sigma_l^2} = \frac{(m-1)\sigma_i^2}{\sum_{j\in G^m\setminus F} \sigma_j^2} \leq \frac{(m-1)\sigma_i^2}{\sum_{j\in G_{k'}\setminus F} \sigma_j^2} \leq \frac{2(m-1)}{|G_{k'} \setminus F|} \leq \frac{2(m-1)}{m+1} < 2, \tag{B.65}$$

where the first inequality is by $G_{k'} \subset G^m$, the second inequality is by $\sigma_i^2 \leq 2\sigma_j^2$ for any $j \in G_{k'}$, and the third inequality is by $|G_{k'}| > 2m$. It follows that

$$\ln(m) \sum_i \sigma_i^2 \left(\frac{(m-1)\sigma_i^2}{\sum_{j\in G^m} \sigma_j^2 - \max_{F:i\in F} \sum_{l\in F} \sigma_l^2} + 1 \right)^{-1} \tag{B.66}$$

$$\geq \ln(m) \sum_i \sigma_i^2 \left(2+1\right)^{-1} = \frac{1}{3} \ln(m) \sum_j \sigma_j^2. \tag{B.67}$$

□

Lemma 22. *There exists some constant* $0 < c' < 1$, *that for any choices of* $\sigma_{1:n}^2$, $\mathrm{Ent}(\sigma_L^2) \geq c'\mathrm{Ent}(\sigma_{G^r}^2) - c' \ln(2)$.

Proof of Lemma 22. By the grouping property of entropy, we have

$$\begin{aligned}
\mathrm{Ent}(\sigma^2_{G^r}) &= \mathrm{Ent}(\sum_{j\in L}\sigma_j^2, \sum_{i\in G^r\setminus L}\sigma_i^2) \tag{B.68}\\
&\quad + \frac{\sum_{i\in L}\sigma_i^2}{\sum_{j\in G^r}\sigma_j^2}\mathrm{Ent}(\sigma_L^2) + \left(1 - \frac{\sum_{i\in L}\sigma_i^2}{\sum_{j\in G^r}\sigma_j^2}\right)\mathrm{Ent}(\sigma^2_{G^r\setminus L}) \tag{B.69}\\
&< \ln(2) + \mathrm{Ent}(\sigma_L^2) + \left(1 - \frac{\sum_{i\in L}\sigma_i^2}{\sum_{j\in G^r}\sigma_j^2}\right) 8\ln(m) \tag{B.70}\\
&\le \ln(2) + 33\mathrm{Ent}(\sigma_L^2), \tag{B.71}
\end{aligned}$$

where the first inequality is due to the principal of maximum entropy and Lemma 6, and the last inequality is by Lemma 25. □

Lemma 23 (Retate Lemma 12). *Let $\eta_F = \binom{2m}{m-1}^{-1}$ for any $F \subset L$ with $|F| = m - 1$ and $\eta_F = 0$ otherwise. There exists some constant $c' > 0.005$. The objective of optimization (3.25) is at least $c' \sum_{i\in G^l}\sigma_i^2\mathrm{Ent}(\sigma^2_{G^r}) - \ln(2)\sum_{i\in L}\sigma_i^2$.*

Proof. Recall $L \subset G^r$ with $|L| = 2m$ is the subset of arms with largest variances within G^r. For any $M \subset L$ with $|M| = m$, by the grouping property of entropy function we have

$$\begin{aligned}
\mathrm{Ent}(\sigma_L^2) &= \mathrm{Ent}(\sum_{i\in M}\sigma_i^2, \sum_{j\in L\setminus M}\sigma_j^2) + \frac{\sum_{i\in M}\sigma_i^2}{\sum_{j\in L}\sigma_j^2}\mathrm{Ent}(\sigma_M^2) + \frac{\sum_{i\in L\setminus M}\sigma_i^2}{\sum_{j\in L}\sigma_j^2}\mathrm{Ent}(\sigma^2_{L\setminus M}) \tag{B.72}\\
&\le \ln(2) + \frac{\sum_{i\in M}\sigma_i^2}{\sum_{j\in L}\sigma_j^2}\mathrm{Ent}(\sigma_M^2) + \frac{\sum_{i\in L\setminus M}\sigma_i^2}{\sum_{j\in L}\sigma_j^2}\mathrm{Ent}(\sigma^2_{L\setminus M}), \tag{B.73}
\end{aligned}$$

where the inequality is by the principal of maximum entropy. Multiply $\sum_{j\in L}\sigma_j^2$ on both side, and we have

$$\sum_{i\in M}\sigma_j^2\mathrm{Ent}(\sigma_M^2) + \sum_{i\in L\setminus M}\sigma_i^2\mathrm{Ent}(\sigma^2_{L\setminus M}) \ge \sum_{j\in L}\sigma_j^2(\mathrm{Ent}(\sigma_l^2) - \ln(2)). \tag{B.74}$$

Since $|M| = |L \setminus M| = m$, summing the inequality above for each $M \subset L$ with $|M| = m$ and multiplying by $\frac{1}{2\binom{2m}{m-1}}$ gives us

$$\sum_{M\subset L:|M|=m}\frac{1}{\binom{2m}{m-1}}\sum_{i\in M}\sigma_i^2\mathrm{Ent}(\sigma_M^2)\geq\frac{\binom{2m}{m}}{2\binom{2m}{m-1}}(\mathrm{Ent}(\sigma_L^2)-\ln(2))\sum_{i\in L}\sigma_i^2 \tag{B.75}$$

$$\geq\frac{1}{2}(\mathrm{Ent}(\sigma_L^2)-\ln(2))\sum_{i\in L}\sigma_i^2=\frac{1}{2}\mathrm{Ent}(\sigma_L^2)\sum_{i\in L}\sigma_i^2-\frac{\ln(2)}{2}\sum_{i\in L}\sigma_i^2 \tag{B.76}$$

$$\geq\frac{1}{2}\sum_{i\in L}\sigma_i^2\frac{\mathrm{Ent}(\sigma_{G^r}^2)-\ln(2)}{33}-\frac{\ln(2)}{2}\sum_{i\in L}\sigma_i^2 \tag{B.77}$$

$$\geq\frac{1}{6}\sum_{i\in G^r}\sigma_i^2\frac{\mathrm{Ent}(\sigma_{G^r}^2)}{33}-\frac{1}{2}\sum_{i\in L}\sigma_i^2\frac{\ln(2)}{33}-\frac{\ln(2)}{2}\sum_{i\in L}\sigma_i^2 \tag{B.78}$$

$$\geq\frac{1}{174}\sum_{i\in G^r}\sigma_i^2\mathrm{Ent}(\sigma_{G^r}^2)-\ln(2)\sum_{i\in L}\sigma_i^2, \tag{B.79}$$

where the second inequality is by $\frac{\binom{2m}{m}}{\binom{2m}{m-1}}\geq 1$, the third and forth inequalities are by Lemma 22. □

B.5 Supporting Lemmas

Lemma 24 (Lemma 5.1 in [37]). *Given two bandit instances $I=(\mu_{1:n},\sigma_{1:n}^2)$ and $I'=(\mu'_{1:n},\sigma'^2_{1:n})$, and let P_I and $P_{I'}$ be the probability measure associated with the bandit instances, respectively. Then for any algorithm A with the number of pulling for each arm-i as T_i^{A}, which is a random variable, let τ^{A} be the bandit process and let $\mathbb{P}_{I,\pi}$ and $\mathbb{P}_{I',\pi}$ be the probability measures induced by τ^{A} on instance I and I', respectively. We have*

$$D(\mathbb{P}_{I,\mathsf{A}}||\mathbb{P}_{I',\mathsf{A}})=\sum_{i=1}^{n}\mathbb{E}_I[T_i^{\mathsf{A}}]D\left(\mathcal{N}(\mu_i,\sigma_i^2)||\mathcal{N}(\mu'_i,\sigma_i'^2)\right). \tag{B.80}$$

Lemma 25. *For any $\sigma_{1:n}^2$, we have $\frac{\sum_{i\in L}\sigma_i^2}{\sum_{j\in G^r}\sigma_j^2}\geq\frac{1}{3}$. In addition,*

$$\left(1-\frac{\sum_{i\in L}\sigma_i^2}{\sum_{j\in G^r}\sigma_j^2}\right)\ln(m)\leq 4\mathrm{Ent}(\sigma_L^2), \tag{B.81}$$

for some constant $c>0$.

Proof of Lemma 25. Suppose the minimum variance in σ_L^2 is $\tilde{\sigma}^2$. Let $\alpha=\frac{2m\tilde{\sigma}^2}{\sum_{i\in L}\sigma_i^2}\in(0,1]$, which implies $\sum_{i\in L} \sigma_i^2 = 2m\tilde{\sigma}^2/\alpha$. In addition, $\sum_{j\in G^r\setminus L} \sigma_j^2 \leq 2m\tilde{\sigma}^2 \sum_{i=0}^{\infty} 2^{-i} = 4m\tilde{\sigma}^2$. It is straightforward to verify that

$$\frac{\sum_{i\in L} \sigma_i^2}{\sum_{j\in G^r} \sigma_j^2} = \frac{2m\tilde{\sigma}^2/\alpha}{\sum_{j\in G^r\setminus L} \sigma_j^2 + 2m\tilde{\sigma}^2/\alpha} \geq \frac{2m\tilde{\sigma}^2/\alpha}{4m\tilde{\sigma}^2 + 2m\tilde{\sigma}^2/\alpha} = \frac{1/\alpha}{2+1/\alpha} \geq \frac{1}{3}, \tag{B.82}$$

which proves the first statement. It follows that

$$1 - \frac{\sum_{i\in L} \sigma_i^2}{\sum_{j\in G^r} \sigma_j^2} \leq 1 - \frac{1/\alpha}{2+1/\alpha} = \frac{2}{2+1/\alpha} < \frac{2}{1+1/\alpha}. \tag{B.83}$$

By concavity of entropy function, $\mathrm{Ent}(\sigma_L^2) \geq \mathrm{Ent}\left(1 - \frac{2m-1}{2m}\alpha, \frac{\alpha}{2m}, \frac{\alpha}{2m}, \cdots, \frac{\alpha}{2m}\right)$. It implies that

$$(1+1/\alpha)\mathrm{Ent}(\sigma_L^2) \tag{B.84}$$

$$\geq (1+1/\alpha)\left(-(1-\frac{2m-1}{m}\alpha)\ln\left(1-\frac{2m-1}{2m}\alpha\right) + \frac{2m-1}{2m}\alpha\ln\frac{2m}{\alpha}\right) \tag{B.85}$$

$$\geq \frac{2m-1}{2m}\ln(2m) + \frac{2m-1}{2m}\ln\frac{1}{\alpha} - \left(\frac{1}{\alpha} - \frac{2m-1}{2m}\right)\ln\left(1-\frac{2m-1}{2m}\alpha\right) \tag{B.86}$$

$$\geq \frac{1}{2}\ln(2m) - \frac{1}{2}\ln(\alpha) - (1/\alpha - 1)\ln(1-\alpha) \tag{B.87}$$

$$\geq \frac{1}{2}\ln(2m) > \frac{1}{2}\ln(m). \tag{B.88}$$

We thus have

$$4\mathrm{Ent}(\sigma_L^2) > \frac{2}{1+1/\alpha}\ln(m) > \left(1 - \frac{\sum_{i\in L} \sigma_i^2}{\sum_{j\in G^r} \sigma_j^2}\right)\ln(m). \tag{B.89}$$

□

APPENDIX C

PROOFS FOR CHAPTER 4

C.1 Supporting Lemmas

Before delving into detailed proofs for the proposition and theorems, we introduce some supporting lemmas.

Recall that the Bregman divergence generated by a convex differentiable function $h(\cdot)$ is

$$B_h(x, y) := h(x) - h(y) - \langle \nabla h(y), x - y \rangle.$$

The fundamental inequality (4.2) associated with mirror ascent is formally presented in the following lemma.

Lemma 26. *Let $B_h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a Bregman divergence function, $\mathcal{X} \subset \mathbb{R}^n$ be a compact convex set, and $g \in \mathbb{R}^n$. Suppose $x' = \arg\max_{y \in \mathcal{X}}\{\langle g, y \rangle - \alpha B_h(y||x)\}$ for a fixed $x \in \mathcal{X}$ and $\alpha > 0$. Then for any $y \in \mathcal{X}$,*

$$\langle g, x' \rangle - \alpha B_h(x'||x) \geq \langle g, y \rangle - \alpha B_h(y||x) + \alpha B_h(y||x').$$

Inequalities of the same form have appeared in many previous works, e.g., Lemma 3.4 in [76] and a case of $\mathcal{X}$ being a probability simplex (Lemma 2.1 in [88]). For completeness, we provide a proof of Lemma 26.

Proof of Lemma 26. Since h is proper and convex, $x' := \arg\max_{y \in \mathcal{X}}\{\langle g, y \rangle - \alpha B_h(y||x)\}$ exists and satisfies the first order condition

$$\langle g - \alpha \nabla h(x') + \alpha \nabla h(x), x' - y \rangle = \langle g - \alpha \nabla_{x'} B_h(x'||x), x' - y \rangle \geq 0, \quad \forall y \in \mathcal{X},$$

which implies $\langle g, x' - y \rangle \geq \alpha \langle \nabla h(x) - \nabla h(x'), x' - y \rangle$. It can be verified that

$$\langle \nabla h(x) - \nabla h(x'), x' - y \rangle = B_h(x' || x) - B_h(y || x) + B_h(y || x').$$

We can conclude the proof by substituting the equation into the previous inequality. □

The following lemma draws a connection between the ℓ_1 difference of state-action visitation distributions and averaged KL-divergence.

Lemma 27. *Let $d_\rho^{\pi'}, d_\rho^{\pi}$ be two discounted state-action visitation distributions corresponding to policies π' and π. Then*

$$\|d_\rho^{\pi'} - d_\rho^{\pi}\|_1 \leq \frac{\gamma\sqrt{2}}{1-\gamma} \sqrt{\min\left(D_{d_\rho^{\pi'}}(\pi' || \pi), D_{d_\rho^{\pi'}}(\pi || \pi'), D_{d_\rho^{\pi}}(\pi' || \pi), D_{d_\rho^{\pi}}(\pi || \pi')\right)}.$$

Proof. Let $d_{\rho,h}^{\pi}(\cdot,\cdot)$ be the state-action visitation distribution at step h, which implies $\frac{1}{1-\gamma} d_\rho^{\pi}(\cdot,\cdot) = \sum_{h\geq 0} \gamma^h d_{\rho,h}^{\pi}(\cdot,\cdot)$. Denote $\tilde{\pi}_h$ as the policy that implements policy π for the first h steps and then commits to policy π' thereafter. Denote its corresponding discounted state-action visitation distribution by $d_\rho^{\tilde{\pi}_h}$. It follows that

$$\begin{aligned}
\frac{1}{1-\gamma}\|d_\rho^{\pi'} - d_\rho^{\pi}\|_1 &\overset{(a)}{=} \frac{1}{1-\gamma}\left\|\sum_{h=0}^{\infty}(d_\rho^{\tilde{\pi}_h} - d_\rho^{\tilde{\pi}_{h+1}})\right\|_1 \overset{(b)}{\leq} \frac{1}{1-\gamma}\sum_{h=0}^{\infty}\|d_\rho^{\tilde{\pi}_h} - d_\rho^{\tilde{\pi}_{h+1}}\|_1 \\
&= \sum_{h=0}^{\infty}\left\|\sum_{t=0}^{\infty}\gamma^t(d_{\rho,t}^{\tilde{\pi}_h} - d_{\rho,t}^{\tilde{\pi}_{h+1}})\right\|_1 = \sum_{h=0}^{\infty}\left\|\sum_{t=h+1}^{\infty}\gamma^t(d_{\rho,t}^{\tilde{\pi}_h} - d_{\rho,t}^{\tilde{\pi}_{h+1}})\right\|_1 \\
&\overset{(c)}{\leq} \sum_{h=0}^{\infty}\sum_{t\geq h+1}^{\infty}\gamma^t\|d_{\rho,t}^{\tilde{\pi}_h} - d_{\rho,t}^{\tilde{\pi}_{h+1}}\|_1 \overset{(d)}{\leq} \sum_{h=0}^{\infty}\sum_{t\geq h+1}^{\infty}\gamma^t\|d_{\rho,h}^{\tilde{\pi}_h} - d_{\rho,h}^{\tilde{\pi}_{h+1}}\|_1 \\
&= \frac{\gamma}{1-\gamma}\sum_{h=0}^{\infty}\gamma^h\mathbb{E}_{s\sim d_{\rho,h}^{\pi}}\|\pi(\cdot|s) - \pi'(\cdot|s)\|_1 \\
&\overset{(e)}{\leq} \frac{\gamma}{1-\gamma}\sqrt{\left(\sum_{h\geq 0}\gamma^h\right)\left(\sum_{h=0}^{\infty}\gamma^h\mathbb{E}_{s\sim d_{\rho,h}^{\pi}}\|\pi(\cdot|s) - \pi'(\cdot|s)\|_1^2\right)} \\
&= \frac{\gamma}{(1-\gamma)^2}\sqrt{\mathbb{E}_{s\sim d_\rho^{\pi}}\|\pi(\cdot|s) - \pi'(\cdot|s)\|_1^2}\,.
\end{aligned}$$

Above, (a) holds by telescoping, (b) and (c) hold due to the triangle inequality of ℓ_1-norm and the definition of $\tilde{\pi}_h$, (d) hold owing to the data processing inequality for f-divergence $\|\cdot\|_1$, and (e) holds due to the Cauchy-Schwarz inequality. Due to the symmetry between π and π', it can be similarly derived

$$\|d_\rho^{\pi'} - d_\rho^{\pi}\|_1 \leq \frac{\gamma}{1-\gamma}\sqrt{\mathbb{E}_{s\sim d_\rho^{\pi'}}\|\pi(\cdot|s) - \pi'(\cdot|s)\|_1^2}\,.$$

We can conclude the proof by further applying Pinsker's inequality. □

An application of Lemma 27 gives an upper bound on the difference between value function vectors as follows.

Lemma 28. *For any* $k = 0, 1, \ldots, K-1$,

$$\frac{1}{2}\|V_{1:m}^{\pi_k}(\rho) - V_{1:m}^{\pi_{k+1}}(\rho)\|_\infty^2 \leq \frac{\gamma^2}{(1-\gamma)^4} D_{d_\rho^{\pi_{k+1}}}(\pi_{k+1}||\pi_k).$$

Proof. For any $i = 1, 2, \ldots, m$, we have

$$\begin{aligned}\left|V_i^{\pi_k}(\rho) - V_i^{\pi_{k+1}}(\rho)\right| &= \frac{1}{1-\gamma}\left|\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} r_i(s,a)(d_\rho^{\pi_k}(s,a) - d_\rho^{\pi_{k+1}}(s,a))\right| \\ &\leq \frac{1}{1-\gamma}\|d_\rho^{\pi_k} - d_\rho^{\pi_{k+1}}\|_1 \leq \frac{\gamma\sqrt{2}}{(1-\gamma)^2}\sqrt{D_{d_\rho^{\pi_{k+1}}}(\pi_{k+1}||\pi_k)},\end{aligned}$$

where the last inequality is due to Lemma 27. □

C.2 Proofs for Chapter 4.2

This section presents the formal proof of Proposition 4. We begin by presenting some properties of InnerLoop. We shall omit θ in π_θ, since the policies are under softmax parameterization.

C.2.1 Linear convergence of InnerLoop

InnerLoop($\tilde{r}_k, \pi_k, \alpha, \eta, t_k$) approximately solves the following KL-regularized MDP via natural policy gradient. Note that

$$\tilde{V}_{k,\alpha}^{\pi}(s) = \mathbb{E}\left[\sum_{t\geq 0} \gamma^t \left(\tilde{r}_k(s_t, a_t) - \alpha \log \pi_k(a_t|s_t) + \alpha \log \pi(a_t|s_t)\right)|s_0 = s, \pi\right], \tag{C.1}$$

which can be viewed as an entropy regularized value with reward function $\tilde{r}_k(s, a) - \alpha \log \pi_k(a|s)$. The entropy-regularized state-action value function is then defined as [77]

$$\tilde{Q}_{k,\alpha}^{\pi}(s, a) = \tilde{r}_k(s, a) - \alpha \log \pi_k(a|s) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[\tilde{V}_{k,\alpha}^{\pi}(s')]. \tag{C.2}$$

The convergence of NPG in entropy-regularized MDP has been well-studied in [77], with the key results summarized in the following lemma.

Lemma 29 (Linear convergence of entropy-regularized NPG, Theorem 1 in [77])**.** *For any learning rate $0 < \eta \leq (1-\gamma)/\alpha$ and any $k = 0, 1, \ldots, K-1$, the entropy-regularized NPG updates satisfy*

$$\left\|\tilde{Q}_{k,\alpha}^{\pi_k^*} - \tilde{Q}_{k,\alpha}^{\pi_k^{(t+1)}}\right\|_\infty \leq C_k \gamma (1-\eta\alpha)^t,$$
$$\left\|\log \pi_k^* - \log \pi_k^{(t+1)}\right\|_\infty \leq 2C_k \alpha^{-1} (1-\eta\alpha)^t,$$
$$\left\|\tilde{V}_{k,\alpha}^{\pi_k^*} - \tilde{V}_{k,\alpha}^{\pi_k^{(t+1)}}\right\|_\infty \leq 3C_k (1-\eta\alpha)^t,$$

for all $t \geq 0$, where C_k satisfies $C_k \geq \left\|\tilde{Q}_{k,\alpha}^{\pi_k^*} - \tilde{Q}_{k,\alpha}^{\pi_k^{(0)}}\right\|_\infty + 2\alpha\left(1 - \frac{\eta\alpha}{1-\gamma}\right)\left\|\log \pi_k^* - \log \pi_k^{(0)}\right\|_\infty$.

Remark. There is a typographical mistake in the inequality " $\|\tilde{V}_{k,\alpha}^{\pi_k^*} - \tilde{V}_{k,\alpha}^{\pi_k^{(t+1)}}\|_\infty \leq 3\gamma C_k(1 - \eta\alpha)^t$ " in [77], and it has been corrected here. It is not hard to verify that the proofs of the inequalities in Lemma 29 [77] hold without the assumption that $0 \leq r(s, a) \leq 1$.

Denote $\tilde{V}_k^{\pi}(s) := V_{\tilde{r}_k}^{\pi}(s)$. For the regularized MDP, its optimal policy is uniformly optimal, i.e., for any state $s \in \mathcal{S}$,

$$\frac{1}{1-\gamma}\|\tilde{r}_k\|_\infty \geq \tilde{V}_k^{\pi_k^*}(s) \geq \tilde{V}_k^{\pi_k^*}(s) - \frac{\alpha}{1-\gamma} D_{d_s^{\pi^*}}(\pi^*||\pi_k) = \tilde{V}_{k,\alpha}^{\pi_k^*}(s) \geq \tilde{V}_{k,\alpha}^{\pi_k}(s) = \tilde{V}_k^{\pi_k}(s). \tag{C.3}$$

It follows that $\forall (s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\left|\tilde{Q}_{k,\alpha}^{\pi_k^*}(s,a) - \tilde{Q}_{k,\alpha}^{\pi_k}(s,a)\right| = \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) \left|\tilde{V}_{k,\alpha}^{\pi_k^*}(s') - \tilde{V}_{k,\alpha}^{\pi_k}(s')\right| \overset{(a)}{\leq} \frac{\gamma\|\tilde{r}_k\|_\infty}{1-\gamma},$$

where (a) holds due to the relation in (C.3). It implies $\|\tilde{Q}_{k,\alpha}^{\pi_k^*} - \tilde{Q}_{k,\alpha}^{\pi_k}\|_\infty \leq \frac{\gamma\|\tilde{r}_k\|_\infty}{1-\gamma}$. Since $1 - \frac{\eta\alpha}{1-\gamma} = 0$ when $\eta = \frac{1-\gamma}{\alpha}$, we can apply results in Lemma 29 with $C_k = \frac{\gamma\|\tilde{r}_k\|_\infty}{1-\gamma}$, which gives

$$\tilde{V}_k^{\pi_{k+1}}(\rho) + \frac{\alpha}{1-\gamma} D_{d_\rho^{\pi_{k+1}}}(\pi_{k+1}||\pi_k) \leq -\tilde{V}_k^{\pi_k^*}(\rho) + \frac{\alpha}{1-\gamma} D_{d_\rho^{\pi_k^*}}(\pi_k^*||\pi_k) + 3C_k(1-\eta\alpha)^{t_k}. \tag{C.4}$$

C.2.2 Hidden convexity in state-action visitation distribution

Noting that the class of softmax policies is almost complete in the sense that its closure contains all stationary policies, we will omit the parameter θ in π_θ. The set of achievable state-action visitations is $\mathcal{D} = \{d \in \Delta(\mathcal{S} \times \mathcal{A}) : \gamma \sum_{s',a'} P(s|s',a')d(s',a') + (1-\gamma)\rho(s) = \sum_a d(s,a),\ \forall s \in \mathcal{S}\}$, which is a convex compact set.

For any policies π, π', define a pseudo KL-divergence between $d_\rho^\pi, d_\rho^{\pi'} \in \mathcal{D}_\rho$ by

$$\tilde{D}(d_\rho^\pi||d_\rho^{\pi'}) := \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} d_\rho^\pi(s,a) \log \frac{d_\rho^\pi(s,a)/d_\rho^\pi(s)}{d_\rho^{\pi'}(s,a)/d_\rho^{\pi'}(s)}. \tag{C.5}$$

It is not hard to verify that

$$\begin{aligned} D_{d_\rho^\pi}(\pi||\pi') &= \sum_{s \in \mathcal{S}} d_\rho^\pi(s) \sum_{a \in \mathcal{A}} \pi(a|s) \log \frac{\pi(a|s)}{\pi'(a|s)} = \sum_{s \in \mathcal{S}} d_\rho^\pi(s) \sum_{a \in \mathcal{A}} \frac{d_\rho^\pi(s,a)}{d_\rho^\pi(s)} \log \frac{d_\rho^\pi(s,a)/d_\rho^\pi(s)}{d_\rho^{\pi'}(s,a)/d_\rho^{\pi'}(s)} \\ &= \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} d_\rho^\pi(s,a) \log \frac{d_\rho^\pi(s,a)/d_\rho^\pi(s)}{d_\rho^{\pi'}(s,a)/d_\rho^{\pi'}(s)} = \tilde{D}(d_\rho^\pi||d_\rho^{\pi'}). \end{aligned} \tag{C.6}$$

This equation bridges the state-action visitation space and the policy space. The following lemma shows that the pseudo KL-divergence defined in (C.5) is actually a Bregman divergence between state-action visitation distributions.

Lemma 30. *The pseudo KL-divergence $\tilde{D}(d_\rho^\pi || d_\rho^{\pi'})$ defined in (C.5) is a Bregman divergence $B_h(d_\rho^\pi || d_\rho^{\pi'})$ generated by the convex function*

$$h(d_\rho^\pi) = \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_\rho^\pi(s,a)\log d_\rho^\pi(s,a) - \sum_{s\in\mathcal{S}} d_\rho^\pi(s)\log d_\rho^\pi(s).$$

Proof of Lemma 30. It can be verified by elementary algebera that

$$\tilde{D}(d_\rho^\pi || d_\rho^{\pi'}) = h(d_\rho^\pi) - h(d_\rho^{\pi'}) - \langle \nabla h(d_\rho^{\pi'}), d_\rho^\pi - d_\rho^{\pi'}\rangle,$$

where $\nabla_{(s,a)} h(d_\rho^{\pi'}) = \log d_\rho^{\pi'}(s,a) - \log d_\rho^{\pi'}(s)$. Hence we only need to show that $h(d_\rho^\pi)$ is convex. The Hessian matrix of function $h(d_\rho^\pi)$ can be calculated as $\operatorname{diag}\left(H_1, H_2, \ldots, H_{|\mathcal{S}|}\right)$, where $H_s = \frac{1}{d_\rho^\pi(s)}\left(\operatorname{diag}(d_\rho^\pi(s)/d_\rho^\pi(s,\cdot)) - \mathbf{1}\mathbf{1}^T\right)$ is an $|\mathcal{A}| \times |\mathcal{A}|$ matrix corresponding to state s. For each H_s, we know for any $x_{1:|\mathcal{A}|} \in \mathbb{R}^{|\mathcal{A}|}$,

$$\begin{aligned} x^T H_s x &= \frac{1}{d_\rho^\pi(s)}\left(\sum_{a\in\mathcal{A}} \frac{d_\rho^\pi(s)}{d_\rho^\pi(s,a)} x_a^2 - \left(\sum_{a\in\mathcal{A}} x_a\right)^2\right) \\ &= \frac{1}{d_\rho^\pi(s)}\left(\left(\sum_{a\in\mathcal{A}} \frac{d_\rho^\pi(s,a)}{d_\rho^\pi(s)}\right)\left(\sum_{a\in\mathcal{A}} \frac{d_\rho^\pi(s)}{d_\rho^\pi(s,a)} x_a^2\right) - \left(\sum_{a\in\mathcal{A}} x_a\right)^2\right) \\ &\overset{(a)}{\geq} \frac{1}{d_\rho^\pi(s)}\left(\left(\sum_{a\in\mathcal{A}} |x_a|\right)^2 - \left(\sum_{a\in\mathcal{A}} x_a\right)^2\right) \geq 0, \end{aligned}$$

where (a) is due to the Cauchy-Schwarz inequality. Thus the Hessian matrix of $h(d_\rho^\pi)$ is positive semi-definite, which implies that $h(d_\rho^\pi)$ is convex. □

InnerLoop of the ARNPG framework is solving a KL-regularized MDP with value as in (4.4),

$$\tilde{V}_{k,\alpha}^{\pi_\theta}(\rho) = V_{\tilde{r}_k}^{\pi_\theta}(\rho) - \alpha \frac{D_{d_\rho^{\pi_\theta}}(\pi_\theta || \pi_{\theta_k})}{1-\gamma}.$$

This optimization can be equivalently represented by viewing state-action visitation as the decision variables:

$$\max_\pi V_{\tilde{r}_k}^{\pi}(\rho) - \alpha \frac{D_{d_\rho^\pi}(\pi || \pi_k)}{1-\gamma} \quad \Leftrightarrow \quad \max_{d \in \mathcal{D}} \langle \tilde{r}_k, d \rangle - \alpha \tilde{D}(d || d_\rho^{\pi_k}). \tag{C.7}$$

Here $\Leftrightarrow$ means that they are equivalent in the sense that the optimal policy solution π_k^* for the former optimization and the optimal visitation solution d_k^* for the latter satisfy $d_k^* = d_\rho^{\pi_k^*}$. Note that $\tilde{V}_{\tilde{r}_k}^{\pi}(\rho) = \frac{1}{1-\gamma}\langle \tilde{r}_k, d_\rho^\pi \rangle$ is a linear function of d_ρ^π, $\tilde{D}(\cdot||\cdot)$ is a Bregman divergence, and $\mathcal{D}$ is compact. We can apply Lemma 26 on the latter optimization and have

$$\langle \tilde{r}_k, d_k^* \rangle - \alpha \tilde{D}(d_k^* || d_\rho^{\pi_k}) \geq \langle \tilde{r}_k, d \rangle - \alpha \tilde{D}(d || d_\rho^{\pi_k}) + \alpha \tilde{D}(d || d_k^*), \quad \forall d \in \mathcal{D}. \tag{C.8}$$

Since the policy class and the state-action visitation class are both complete, the inequality above implies that

$$V_{\tilde{r}_k}^{\pi_{k+1}}(\rho) - \alpha \frac{D_{d_\rho^{\pi_k^*}}(\pi_k^* || \pi_k)}{1-\gamma} \geq V_{\tilde{r}_k}^{\pi}(\rho) - \alpha \frac{D_{d_\rho^\pi}(\pi || \pi_k) - D_{d_\rho^\pi}(\pi || \pi_k^*)}{1-\gamma}, \quad \forall \pi. \tag{C.9}$$

InnerLoop does not seek to find the precise solution π_k^* but approximates it with $\pi_{k+1} = \pi_k^{(t_k)}$ via t_k micro-step iterations. Proposition 4 provides a quantitative bound regarding the approximation error of π_{k+1}.

C.2.3 Proof of Proposition 4

Proof of Proposition 4. Combining (C.4) and (C.9) gives

$$-\tilde{V}_k^{\pi_{k+1}}(\rho) + \alpha \frac{D_{d_\rho^{\pi_{k+1}}}(\pi_{k+1} || \pi_k)}{1-\gamma} \leq -\tilde{V}_k^{\pi}(\rho) + \alpha \frac{D_{d_\rho^\pi}(\pi || \pi_k) - D_{d_\rho^\pi}(\pi || \pi_{k+1})}{1-\gamma}$$

$$+ 3C_k(1-\eta\alpha)^{t_k} + \alpha \frac{D_{d_\rho^\pi}(\pi||\pi_{k+1}) - D_{d_\rho^\pi}(\pi||\pi_k^*)}{1-\gamma}.$$

Note that

$$\begin{aligned}
&D_{d_\rho^\pi}(\pi||\pi_{k+1}) - D_{d_\rho^\pi}(\pi||\pi_k^*) = \left\langle d_\rho^\pi(\cdot,\cdot), \log \frac{\pi_k^*(\cdot,\cdot)}{\pi_{k+1}(\cdot,\cdot)} \right\rangle \\
\leq &\|d_\rho^\pi\|_1 \| \log \pi_k^* - \log \pi_{k+1}\|_\infty = \| \log \pi_k^* - \log \pi_{k+1}\|_\infty \leq 2C_k \alpha^{-1}(1-\eta\alpha)^{t_k},
\end{aligned}$$

where the first inequality follows from Cauchy-Schwartz, and the last inequality is due to Lemma 29. We thus have

$$-\tilde{V}_k^{\pi_{k+1}}(\rho) + \alpha \frac{D_{d_\rho^{\pi_{k+1}}}(\pi_{k+1}||\pi_k)}{1-\gamma} \leq -\tilde{V}_k^{\pi}(\rho) + \alpha \frac{D_{d_\rho^\pi}(\pi||\pi_k) - D_{d_\rho^\pi}(\pi||\pi_{k+1})}{1-\gamma} + \frac{5C_k(1-\eta\alpha)^{t_k}}{1-\gamma}.$$

We then conclude the proposition, since $\frac{5C_k(1-\eta\alpha)^t}{1-\gamma} \leq \epsilon_k$ can be guaranteed by $t_k \geq \frac{1}{1-\gamma}\log(\frac{5\gamma\|\tilde{r}_k\|_\infty}{(1-\gamma)^2\epsilon_k})$.

□

C.3 Proofs for Chapter 4.3

C.3.1 ARNPG-IMD for smooth scalarization

Proof of Theorem 13. By $|\tilde{r}_k(s,a)| = |\langle \tilde{G}_k, r_{1:m}(s,a)\rangle| \leq \|\tilde{G}_k\|_1 \|r_{1:m}(s,a)\|_\infty \leq L$, we know $\|\tilde{r}_k\|_\infty \leq L$. Recall $\alpha \geq \frac{\beta}{(1-\gamma)^3}$. Taking $\epsilon_k = \frac{\alpha \log(|\mathcal{A}|)}{(1-\gamma)K}$, we choose $t_k = \lceil \frac{1}{1-\gamma} \log(\frac{5LK}{\beta \log(|\mathcal{A}|)}) + 1 \rceil$. Thus by Proposition 4, for any policy π, we have the fundamental inequality

$$V_{\tilde{r}_k}^{\pi_{k+1}}(\rho) - \alpha \frac{D_{d_\rho^{\pi_{k+1}}}(\pi_{k+1}||\pi_k)}{1-\gamma} \geq V_{\tilde{r}_k}^{\pi}(\rho) - \alpha \frac{D_{d_\rho^\pi}(\pi||\pi_k) - D_{d_\rho^\pi}(\pi||\pi_{k+1})}{1-\gamma} - \epsilon_k. \qquad \text{(C.10)}$$

For the RHS of (C.10), by the concavity of F, we have

$$V_{\tilde{r}_k}^{\pi}(\rho) - V_{\tilde{r}_k}^{\pi_k}(\rho) = \langle \tilde{G}_k, V_{1:m}^{\pi}(\rho) - V_{1:m}^{\pi_k}(\rho)\rangle \geq F(V_{1:m}^{\pi}(\rho)) - F(V_{1:m}^{\pi_k}(\rho)).$$

For the LHS of (C.10), by the fact that F is β-smooth, we know

$$\begin{aligned} V_{\tilde{r}_k}^{\pi_{k+1}}(\rho) - V_{\tilde{r}_k}^{\pi_k}(\rho) &= \langle \tilde{G}_k, V_{1:m}^{\pi_{k+1}}(\rho) - V_{1:m}^{\pi_k}(\rho) \rangle \\ &\leq F(V_{1:m}^{\pi_{k+1}}(\rho)) - F(V_{1:m}^{\pi_k}(\rho)) + \frac{\beta}{2} \left\| V_{1:m}^{\pi_k}(\rho) - V_{1:m}^{\pi_{k+1}}(\rho) \right\|_\infty^2 . \end{aligned}$$

From Lemma 28 and recalling $\alpha \geq \frac{\beta}{(1-\gamma)^3}$,

$$\frac{\beta}{2} \| V_{1:m}^{\pi_k}(\rho) - V_{1:m}^{\pi_{k+1}}(\rho) \|_\infty^2 \leq \frac{\gamma^2 \beta}{(1-\gamma)^4} D_{d_\rho^{\pi_{k+1}}}(\pi_{k+1} || \pi_k) \leq \alpha \frac{D_{d_\rho^{\pi_{k+1}}}(\pi_{k+1} || \pi_k)}{1-\gamma}.$$

Substituting these three inequalities into the fundamental inequality (C.10), telescoping from $k = 0$ to $K - 1$, and selecting $\pi = \pi^*$, we can conclude that

$$\frac{1}{K} \sum_{k=1}^{K} F(V_{1:m}^{\pi_k}(\rho)) \geq F(V_{1:m}^{\pi^*}(\rho)) - \frac{\alpha D_{d_\rho^{\pi^*}}(\pi^* || \pi_0)}{(1-\gamma)K} - \frac{1}{K} \sum_{k=0}^{K-1} \epsilon_k \geq F(V_{1:m}^{\pi^*}(\rho)) - \frac{2\alpha \log(|\mathcal{A}|)}{(1-\gamma)K}.$$

□

Proof of Corollary 6. Note that $T = \sum_{k=0}^{K-1} t_k = \Theta(\frac{K}{1-\gamma} \log(K))$. It implies $\frac{K}{1-\gamma} = \Theta(T / \log(T))$. Substituting this into Theorem 13 concludes Corollary 6. □

C.3.2 ARNPG-EPD for CMDP

We first introduce the properties of the Lagrange multiplier updates (4.10) in the following lemma.

Lemma 31 (Properties of Lagrange multiplier updates)**.** *Based on the update of the Lagrange multipliers λ_k, for any $i \in [2 : m]$ we have:*

1. *At any macro step k, $\lambda_{k,i} \geq 0$.*
2. *At any macro step k, $\lambda_{k,i} + \eta'(b_i - V_i^{\pi_k}(\rho)) \geq 0$.*
3. *At macro step 0, $|\lambda_{0,i}| \leq \eta' |V_i^{\pi_0}(\rho) - b_i|$; at any macro step $k > 0$, $|\lambda_{k,i}| \geq \eta' |V_i^{\pi_k}(\rho) - b_i|$.*

Remark. The first property guarantees the feasibility of the Lagrange multipliers; the second property ensures that the Lagrangian in the inner loop can indeed maximize the constraint rewards; and the third property is a key supporting step for the analysis of the constraint violation.

Proof of Lemma 31. Taking any $i \in [2 : m]$, we prove each property respectively.

1. Note that $\lambda_{0,i} = \max\{0, \eta'(V_i^{\pi_0}(\rho) - b_i)\} \geq 0$ by initialization. Suppose $\lambda_{k,i} \geq 0$. The update is $\lambda_{k+1,i} = \max\left\{\eta'(V_i^{\pi_{k+1}}(\rho) - b_i), \lambda_{k,i} + \eta'(b_i - V_i^{\pi_{k+1}}(\rho))\right\}$.
   If $b_i - V_i^{\pi_{k+1}}(\rho) < 0$, then $\lambda_{k+1,i} \geq 0$ by the first component in the $\max\{\cdot,\cdot\}$.
   If $b_i - V_i^{\pi_{k+1}}(\rho) \geq 0$, then $\lambda_{k+1,i} \geq 0$ by the second component in the $\max\{\cdot,\cdot\}$.
   Thus, $\lambda_{k+1,i} \geq 0$, and property can be proved by induction.

2. For $k = 0$, $\lambda_{0,i} + \eta'(b_i - V_i^{\pi_0}(\rho)) = \max\{\eta'(b_i - V_i^{\pi_0}(\rho)), 0\} \geq 0$.
   The update is $\lambda_{k+1,i} = \max\left\{\eta'(V_i^{\pi_{k+1}}(\rho) - b_i), \lambda_{k,i} + \eta'(b_i - V_i^{\pi_{k+1}}(\rho))\right\}$. Thus for $k \geq 0$, $\lambda_{k+1,i} + \eta'(b_i - V_i^{\pi_{k+1}}(\rho)) = \max\left\{0, \lambda_{k,i} + 2\eta'(b_i - V_i^{\pi_{k+1}}(\rho))\right\} \geq 0$.

3. For $k = 0$, the initialization is $\lambda_{0,i} = \max\{0, \eta'(V_i^{\pi_0}(\rho) - b_i)\}$.
   If $V_i^{\pi_0}(\rho) - b_i \leq 0$, then $\lambda_{0,i} = 0$ and $|\lambda_{0,i}| \leq \eta'|V_i^{\pi_0}(\rho) - b_i|$.
   If $V_i^{\pi_0}(\rho) - b_i > 0$, then $\lambda_{0,i} = \eta'(V_i^{\pi_0}(\rho) - b_i)$ and $|\lambda_{0,i}| = \eta'|V_i^{\pi_0}(\rho) - b_i|$.
   For $k \geq 0$, the update is $\lambda_{k+1,i} = \max\left\{\eta'(V_i^{\pi_{k+1}}(\rho) - b_i), \lambda_{k,i} + \eta'(b_i - V_i^{\pi_{k+1}}(\rho))\right\}$.
   If $V_i^{\pi_{k+1}}(\rho) - b_i \leq 0$, then $\lambda_{k+1,i} = \lambda_{k,i} + \eta'(b_i - V_i^{\pi_{k+1}}(\rho))$, and $|\lambda_{k+1,i}| = \lambda_{k,i} + \eta'|V_i^{\pi_{k+1}}(\rho) - b_i| \geq \eta'|V_i^{\pi_{k+1}}(\rho) - b_i|$ by the first property that $\lambda_{k,i} \geq 0$.
   If $V_i^{\pi_{k+1}}(\rho) - b_i > 0$, then $\lambda_{k+1,i} \geq \eta'(V_i^{\pi_{k+1}}(\rho) - b_i) > 0$. Thus $|\lambda_{k+1,i}| \geq \eta'|V_i^{\pi_{k+1}}(\rho) - b_i|$.

□

We now analyze the optimality gap and constraint violation separately.

C.3.2.1 Optimality gap of ARNPG-EPD

Recall the definition of the reward in the ascent direction

$$\tilde{r}_k(s,a) = r_1(s,a) + \sum_{i=2}^{m} [\lambda_{k,i} + \eta'(b_i - V_i^{\pi_k}(\rho))] r_i(s,a). \tag{C.11}$$

Since $r_i(s,a) \leq 1$, we can verify that $|\tilde{r}_k(s,a)| \leq 1 + \frac{\eta'(m-1)}{1-\gamma} + \sum_{i=2}^m \lambda_{k,i} =: L_k$, which implies $\|\tilde{r}_k\|_\infty \leq L_k$. Taking $\epsilon_k = \frac{\alpha \log(|\mathcal{A}|)}{(1-\gamma)K}$, we choose $t_k = \lceil \frac{1}{1-\gamma} \log(\frac{5L_kK}{2\eta' m \log(|\mathcal{A}|)}) + 1 \rceil$.

Since $\lambda_{k,i} + \eta'(b_i - V_i^{\pi_k}(\rho)) \geq 0$ by the second property in Lemma 31, and $V_i^{\pi^*}(\rho) \geq b_i$ for any $i \in [2:m]$, taking $\pi = \pi^*$ in Proposition 4 gives

$$\begin{aligned} &V_1^{\pi_{k+1}}(\rho) + \sum_{i=2}^m [\lambda_{k,i} + \eta'(b_i - V_i^{\pi_k}(\rho))] \cdot [V_i^{\pi_{k+1}}(\rho) - b_i] - \alpha \frac{D_{d_\rho^{\pi_{k+1}}}(\pi_{k+1}||\pi_k)}{1-\gamma} \\ \geq &V_1^{\pi^*}(\rho) - \alpha \frac{D_{d_\rho^{\pi^*}}(\pi^*||\pi_k) - D_{d_\rho^{\pi^*}}(\pi^*||\pi_{k+1})}{1-\gamma} - \epsilon_k. \end{aligned} \tag{C.12}$$

Denote $\delta_{k,i} := b_i - V_i^{\pi_k}(\rho)$ as the constraint violation for the i-th constraint at macro step k. We thus have

$$[\lambda_{k,i} + \eta'(b_i - V_i^{\pi_k}(\rho))] \cdot (V_i^{\pi_{k+1}}(\rho) - b_i) = -\lambda_{k,i}\delta_{k+1,i} - \eta'\delta_{k,i}\delta_{k+1,i}.$$

We can then bound this two terms respectively.

- $\lambda_{k,i}\delta_{k+1,i}$: Note that $\lambda_{k+1,i} = \max\{-\eta'\delta_{k+1,i}, \lambda_{k,i} + \eta'\delta_{k+1,i}\}$.

  If $\lambda_{k+1,i} = -\eta'\delta_{k+1,i}$, then

  $$\frac{1}{2}\lambda_{k+1,i}^2 - \frac{1}{2}\lambda_{k,i}^2 - \eta'^2\delta_{k+1,i}^2 = -\frac{1}{2}\lambda_{k,i}^2 - \frac{\eta'^2}{2}\delta_{k+1,i}^2 \leq \eta'\lambda_{k,i}\delta_{k+1,i},$$

  which implies $-\lambda_{k,i}\delta_{k+1,i} \leq \frac{\lambda_{k,i}^2 - \lambda_{k+1,i}^2}{2\eta'} + \eta'\delta_{k+1,i}^2$.

  If $\lambda_{k+1,i} = \lambda_{k,i} + \eta'\delta_{k+1,i}$, then

  $$\eta'\lambda_{k,i}\delta_{k+1,i} = \frac{1}{2}(\lambda_{k,i} + \eta'\delta_{k+1,i})^2 - \frac{1}{2}\lambda_{k,i}^2 - \frac{\eta'^2}{2}\delta_{k+1,i}^2 \geq \frac{1}{2}\lambda_{k+1,i}^2 - \frac{1}{2}\lambda_{k,i}^2 - \eta'^2\delta_{k+1,i}^2,$$

  which also implies $-\lambda_{k,i}\delta_{k+1,i} \leq \frac{\lambda_{k,i}^2 - \lambda_{k+1,i}^2}{2\eta'} + \eta'\delta_{k+1,i}^2$.

- $\eta'\delta_{k,i}\delta_{k+1,i}$: Note that $\eta'\delta_{k,i}\delta_{k+1,i} = \frac{\eta'}{2}\delta_{k,i}^2 + \frac{\eta'}{2}\delta_{k+1,i} - \frac{\eta'}{2}(\delta_{k,i} - \delta_{k+1,i})^2$, and $\frac{\eta'}{2}(\delta_{k,i} - \delta_{k+1,i})^2 \leq \frac{\gamma^2\eta'}{(1-\gamma)^4} D_{d_\rho^{\pi_{k+1}}}(\pi_{k+1}||\pi_k)$. We thus have $-\eta'\delta_{k,i}\delta_{k+1,i} \leq -\frac{\eta'}{2}(\delta_{k,i}^2 + \delta_{k+1,i}^2) +$

$\frac{\gamma^2\eta'}{(1-\gamma)^4} D_{d_\rho^{\pi_{k+1}}}(\pi_{k+1}||\pi_k)$.

Substituting the above upper bounds into (C.12) leads to

$$\begin{aligned}
&V_1^{\pi_{k+1}}(\rho) + \frac{\|\lambda_k\|_2^2 - \|\lambda_{k+1}\|_2^2}{2\eta'} + \eta'\frac{\|\delta_{k+1}\|_2^2 - \|\delta_k\|_2^2}{2} + \left(\frac{\eta'\gamma^2 m}{(1-\gamma)^4} - \frac{\alpha}{1-\gamma}\right) D_{d_\rho^{\pi_{k+1}}}(\pi_{k+1}||\pi_k) \\
\geq& V_1^{\pi^*}(\rho) - \alpha\frac{D_{d_\rho^{\pi^*}}(\pi^*||\pi_k) - D_{d_\rho^{\pi^*}}(\pi^*||\pi_{k+1})}{1-\gamma} - \epsilon_k.
\end{aligned}$$

Recall $\alpha \geq \frac{2\eta' m}{(1-\gamma)^3}$, it then follows from telescoping that

$$\begin{aligned}
\sum_{k=1}^{K} V_1^{\pi_k}(\rho) \geq& K V_1^{\pi^*}(\rho) - \alpha\frac{D_{d_\rho^{\pi^*}}(\pi^*||\pi_0) - D_{d_\rho^{\pi^*}}(\pi^*||\pi_K)}{1-\gamma} - \sum_{k=0}^{K-1}\epsilon_k \\
&+ \eta'\frac{\|\delta_0\|_2^2 - \|\delta_K\|_2^2}{2} + \frac{\|\lambda_K\|_2^2 - \|\lambda_0\|_2^2}{2\eta'} \qquad \text{(C.13)} \\
=& K V_1^{\pi^*}(\rho) - \alpha\frac{D_{d_\rho^{\pi^*}}(\pi^*||\pi_0) - D_{d_\rho^{\pi^*}}(\pi^*||\pi_K)}{1-\gamma} - \sum_{k=0}^{K-1}\epsilon_k \\
&+ \left(\frac{\|\lambda_K\|_2^2}{2\eta'} - \eta'\frac{\|\delta_K\|_2^2}{2}\right) + \eta'\frac{\|\delta_0\|_2^2 - \|\lambda_0\|_2^2}{2} - \frac{1/\eta' - \eta'}{2}\|\lambda_0\|_2^2 \qquad \text{(C.14)} \\
\overset{(a)}{\geq}& K V_1^{\pi^*}(\rho) - \alpha\frac{D_{d_\rho^{\pi^*}}(\pi^*||\pi_0) - D_{d_\rho^{\pi^*}}(\pi^*||\pi_K)}{1-\gamma} - \sum_{k=0}^{K-1}\epsilon_k - \frac{1/\eta' - \eta'}{2}\|\lambda_0\|_2^2 \\
\overset{(b)}{\geq}& K V_1^{\pi^*}(\rho) - \frac{3\alpha\log(|\mathcal{A}|)}{1-\gamma}. \qquad \text{(C.15)}
\end{aligned}$$

(a) holds due to the third property of Lemma 31, and (b) holds since π_0 is the uniformly distributed policy. Thus $D_{d_\rho^{\pi^*}}(\pi^*||\pi_0) = \sum_{s\in\mathcal{S}} d_\rho^{\pi^*}(s)\sum_{a\in\mathcal{A}}\pi^*(a|s)\log(|\mathcal{A}|\pi^*(a|s)) \leq \log(|\mathcal{A}|)$, $\sum_{k=0}^{K-1}\epsilon_k = \frac{\alpha\log(|\mathcal{A}|)}{1-\gamma}$, and $\lambda_{0,i}^2 = \eta'^2[\delta_{0,i}]_+^2$ implying $\frac{1/\eta'-\eta'}{2}\|\lambda_0\|^2 \leq \frac{(\eta'-\eta'^3)\|\delta_0\|^2}{2} \leq \frac{\eta'}{2(1-\gamma)^2} \leq \frac{\alpha\log(|\mathcal{A}|)}{1-\gamma}$. We now obtain the bound (4.11), after dividing by K on both sides.

C.3.2.2 Violation gap of ARNPG-EPD

Recall that $\delta_{k,i} := b_i - V_i^{\pi_k}(\rho)$ is the constraint violation for the i-th constraint at macro step k. We aim to provide an upper bound on $\sum_{k=1}^{K}\delta_{k,i}$ to control the constraint violation.

For any $i \in [2:m]$, since $\lambda_{k,i} = \max\{-\eta'\delta_{k,i}, \lambda_{k-1,i} + \eta'\delta_{k,i}\} \geq \lambda_{k-1,i} + \eta'\delta_{k,i}$, we have

$$\sum_{k=1}^{K} \delta_{k,i} \leq \frac{\lambda_{K,i} - \lambda_{0,i}}{\eta'} \leq \frac{\lambda_{K,i}}{\eta'} \leq \frac{\|\lambda_K\|_2}{\eta'} \leq \frac{\|\lambda^*\|_2 + \|\lambda_K - \lambda^*\|_2}{\eta'}. \tag{C.16}$$

To upper bound the constraint violation, it therefore suffices to bound the dual variables.

Consider the Lagrangian with optimal dual variable $\mathcal{L}(\pi, \lambda^*) = V_1^{\pi}(\rho) + \sum_{i=2}^{m} \lambda_i^*(V_i^{\pi}(\rho) - b_i)$, whose maximum value $V_1^{\pi^*}(\rho)$ is achieved by the optimal policy π^*. We know

$$\begin{aligned}
&KV_1^{\pi^*}(\rho) \overset{(a)}{=} K\mathcal{L}(\pi^*, \lambda^*) \geq \sum_{k=1}^{K} \mathcal{L}(\pi_k, \lambda^*) = \sum_{k=1}^{K} V_1^{\pi_k}(\rho) + \sum_{i=2}^{m} \lambda_i^* \sum_{k=1}^{K} (V_i^{\pi_k}(\rho) - b_i) \\
&= \sum_{k=1}^{K} V_1^{\pi_k}(\rho) - \sum_{i=2}^{m} \lambda_i^* \sum_{k=1}^{K} \delta_{k,i} \overset{(b)}{\geq} \sum_{k=1}^{K} V_1^{\pi_k}(\rho) - \frac{1}{\eta'} \sum_{i=2}^{m} \lambda_i^* \lambda_{K,i} \\
&\overset{(c)}{\geq} KV_1^{\pi^*}(\rho) - \alpha \frac{D_{d_\rho^{\pi^*}}(\pi^*||\pi_0) - D_{d_\rho^{\pi^*}}(\pi^*||\pi_K)}{1-\gamma} + \frac{\|\lambda_K\|^2}{2\eta'} - \frac{\eta'\|\delta_K\|^2}{2} - \frac{\lambda_i^* \sum_{i=2}^{m} \lambda_{K,i}}{\eta'} - \Delta_K \\
&\geq KV_1^{\pi^*}(\rho) - \frac{\alpha \log(|\mathcal{A}|)}{1-\gamma} + \frac{\alpha D_{d_\rho^{\pi^*}}(\pi^*||\pi_K)}{1-\gamma} + \frac{\|\lambda_K\|_2^2}{2\eta'} - \frac{\eta'\|\delta_K\|_2^2}{2} - \frac{\lambda_i^* \sum_{i=2}^{m} \lambda_{K,i}}{\eta'} - \Delta_K,
\end{aligned}$$

where $\Delta_K := \frac{2\alpha \log(|\mathcal{A}|)}{1-\gamma} \geq \sum_{k=0}^{K-1} \epsilon_k + \frac{1/\eta' - \eta'}{2}\|\lambda_0\|_2^2$. Then (a) holds due to complementary slackness $\lambda_i^*(V_i^{\pi^*}(\rho) - b_i) = 0$, (b) follows from (C.16), and (c) follows from (C.14) and the third property of Lemma 31. It then follows that

$$\frac{\|\lambda_K\|_2^2}{2\eta'} - \frac{\lambda_i^* \sum_{i=2}^{m} \lambda_{K,i}}{\eta'} \leq \frac{\alpha \log(|\mathcal{A}|)}{1-\gamma} - \frac{\alpha D_{d_\rho^{\pi^*}}(\pi^*||\pi_K)}{1-\gamma} + \frac{\eta'\|\delta_K\|_2^2}{2} + \Delta_K. \tag{C.17}$$

Denoting $\delta_i^* := b_i - V_i^{\pi^*}(\rho) \leq 0$, according to Lemma 28, we have

$$\frac{\alpha D_{d_\rho^{\pi^*}}(\pi^*||\pi_K)}{1-\gamma} \geq \frac{(1-\gamma)^3\alpha}{2\gamma^2}\|\delta_K - \delta^*\|_\infty^2 \geq \frac{(1-\gamma)^3\alpha}{2\gamma^2 m}\|\delta_K - \delta^*\|_2^2. \tag{C.18}$$

We can also obtain

$$-\frac{(1-\gamma)^3\alpha}{2\gamma^2 m}\|\delta_K - \delta^*\|_2^2 + \frac{\eta'}{2}\|\delta_K\|_2^2 = \left(\frac{\eta'}{2} - \frac{\gamma^2 m \eta'^2}{2[\gamma^2 m \eta' - (1-\gamma)^3\alpha]}\right)\|\delta^*\|^2 \tag{C.19}$$

$$+\frac{\gamma^2 m\eta' - (1-\gamma)^3\alpha}{2\gamma^2 m}\left\|\delta_K - \delta^* + \frac{\gamma^2 m\eta'}{\gamma^2 m\eta' - (1-\gamma)^3\alpha}\delta^*\right\|^2,$$

by substituting $a = \frac{(1-\gamma)^3\alpha}{2\gamma^2 m}, b = \frac{\eta'}{2}, x = \delta_K - \delta^*, y = \delta^*$ into the binomial equation

$$-a\|x\|_2^2 + b\|x+y\|_2^2 = (b - \frac{b^2}{b-a})\|y\|_2^2 + (b-a)\|x + \frac{b}{b-a}y\|_2^2.$$

Recalling $\alpha \geq \frac{2\eta' m}{(1-\gamma)^3}$, we can verify that $\frac{\gamma^2 m\eta' - (1-\gamma)^3\alpha}{2\gamma^2 m} \leq 0$ and $\frac{\eta'}{2} - \frac{\gamma^2 m\eta'^2}{2[\gamma^2 m\eta' - (1-\gamma)^3\alpha]} \leq \eta'$. It follows that

$$-\frac{(1-\gamma)^3\alpha}{2\gamma^2 m}\|\delta_K - \delta^*\|_2^2 + \frac{\eta'}{2}\|\delta_K\|_2^2 \leq \eta'\|\delta^*\|_2^2. \tag{C.20}$$

Substituting (C.18) and (C.20) into (C.17) gives

$$\begin{aligned}\frac{1}{2\eta'}\|\lambda_K - \lambda^*\|_2^2 =& \frac{1}{2\eta'}\|\lambda^*\|^2 + \frac{1}{2\eta'}\|\lambda_K\|^2 - \frac{1}{\eta'}\sum_{i=1}^m \lambda_i^*\lambda_{K,i} \\ \leq& \frac{1}{2\eta'}\|\lambda^*\|_2^2 + \frac{\alpha\log(|\mathcal{A}|)}{1-\gamma} + \Delta_K + \eta'\|\delta^*\|^2 \\ \leq& \frac{1}{2\eta'}\|\lambda^*\|_2^2 + \frac{3\alpha\log(|\mathcal{A}|)}{1-\gamma} + \frac{\eta'(m-1)}{(1-\gamma)^2} \\ \leq& \frac{1}{2\eta'}\|\lambda^*\|_2^2 + \frac{4\alpha\log(|\mathcal{A}|)}{1-\gamma},\end{aligned} \tag{C.21}$$

where the last inequality follows from $\frac{\eta'(m-1)}{(1-\gamma)^2} \leq \frac{\alpha\log(|\mathcal{A}|)}{1-\gamma}$. Using the above bound in (C.16), we get

$$\begin{aligned}\sum_{k=1}^K \delta_{k,i} \leq \frac{\|\lambda^*\|_2}{\eta'} + \frac{\|\lambda_K - \lambda^*\|_2}{\eta'} \leq \frac{\|\lambda^*\|_2}{\eta'} + \sqrt{\frac{\|\lambda^*\|_2^2}{\eta'^2} + \frac{8\alpha\log(|\mathcal{A}|)}{(1-\gamma)\eta'}} \\ \leq 2\frac{\|\lambda^*\|_2}{\eta'} + 3\sqrt{\frac{\alpha\log(|\mathcal{A}|)}{(1-\gamma)\eta'}},\end{aligned} \tag{C.22}$$

from which we the constraint violation upper bound given in (4.12) follows.

Proof of Theorem 14. We can conclude Theorem 14 from the above discussion on the optimality gap and the constraint violation. □

Proof of Corollary 7. Note that the number of iterations in the inner loop depends on the value of dual variables, i.e., $t_k = \lceil \frac{1}{1-\gamma} \log(\frac{5L_k K}{2\eta' m \log(|\mathcal{A}|)}) + 1 \rceil$ with $L_k = 1 + \frac{\eta'(m-1)}{1-\gamma} + \sum_{i=2}^m \lambda_{k,i}$. It is easy to verify that

$$\frac{1}{2\eta'} \|\lambda_k - \lambda^*\|_2^2 \leq \frac{1}{2\eta'} \|\lambda^*\|_2^2 + \frac{4\alpha \log(|\mathcal{A}|)}{1-\gamma}$$

in the same manner as the proof of inequality (C.21). It then follows that

$$\begin{aligned} \sum_{i=2}^m \lambda_{k,i} = \|\lambda_k\|_1 \leq \sqrt{m}\|\lambda_k\|_2 \leq \sqrt{m}(\|\lambda_k - \lambda^*\| + \|\lambda^*\|) \\ \leq \sqrt{2m\|\lambda^*\|_2^2 + \frac{8\eta' m\alpha \log(|\mathcal{A}|)}{1-\gamma}} = O\left(\sqrt{m}\|\lambda^*\|_2 + \frac{m \log(|\mathcal{A}|)}{(1-\gamma)^2}\right). \end{aligned}$$

We then have $t_k = \Theta\left(\frac{1}{1-\gamma}\log(K)\right)$, and $T = \sum_{k=0}^{K-1} t_k = \Theta\left(\frac{K}{1-\gamma}\log(K)\right)$. We conclude the proof by $\frac{K}{1-\gamma} = \Theta(T/\log(T))$. □

C.3.3 ARNPG-OMDA for max-min trade-off

C.3.3.1 Smoothness property

Define $\mathcal{X} := \mathcal{V}_\rho \times \Delta([m]) \subset \mathbb{R}^{2m}$. Define a norm Ψ on $\mathbb{R}^{2m}$ by $\Psi(v, \lambda) = \|v\|_\infty + \|\lambda\|_1$. Its dual norm is $\Psi^*(v, \lambda) = \|v\|_1 + \|\lambda\|_\infty$.

Define $G^{v,-\lambda}(X) := (\nabla_v \Phi(X), -\nabla_\lambda \Phi(X))$ for $X \in \mathcal{X}$. Assume the function Φ is β-smooth w.r.t. the Ψ-norm over its domain $\mathcal{X}$, i.e.,

$$\Psi^*(G^{v,-\lambda}(X) - G^{v,-\lambda}(X')) \leq \beta\Psi(X - X'), \quad \forall X, X' \in \mathcal{X}. \tag{C.23}$$

Define E_k, which will be an auxiliary term for the convergence analysis, as follows:

$$E_k := \langle \tilde{G}_k^v - \tilde{G}_{k+1}^v, V_{1:m}^{\pi_{k+1}}(\rho) - V_{1:m}^{\tilde{\pi}_{k+1}}(\rho) \rangle + \alpha \frac{D_{d_\rho^{\tilde{\pi}_{k+1}}}(\tilde{\pi}_{k+1}||\pi_k) + D_{d_\rho^{\tilde{\pi}}}(\pi_{k+1}||\tilde{\pi}_{k+1})}{1-\gamma}$$

$$+ \langle \tilde{G}^{\lambda}_{k} - \tilde{G}^{\lambda}_{k+1}, \tilde{\lambda}_{k+1} - \lambda_{k+1} \rangle + \frac{D(\lambda_{k+1}||\tilde{\lambda}_{k+1}) + D(\tilde{\lambda}_{k+1}||\lambda_k)}{\eta'}. \tag{C.24}$$

Lemma 32 (Technical lemma for smoothness)**.** *When* $\alpha \geq \frac{6\beta}{(1-\gamma)^4}$ *and* $\eta' \leq \frac{1}{6\beta}$, $\sum_{k=0}^{K-1} E_k \geq 0$.

Proof of Lemma 32. Recall the definition of E_k (C.24). Let $X_k := (V^{\pi_k}_{1:m}(\rho), \lambda_k) \in \mathcal{X}$ and $\tilde{X}_k := (V^{\tilde{\pi}_k}_{1:m}(\rho), \tilde{\lambda}_k) \in \mathcal{X}$; $G^{v,-\lambda}_k := G^{v,-\lambda}(X_k)$ and $\tilde{G}^{v,-\lambda}_k := G^{v,-\lambda}(\tilde{X}_k)$. We can then rewrite E_k as

$$\begin{aligned}E_k =& \langle \tilde{G}^{v,-\lambda}_k - \tilde{G}^{v,-\lambda}_{k+1}, X_{k+1} - \tilde{X}_{k+1} \rangle + \alpha \frac{D_{d_\rho^{\tilde{\pi}_{k+1}}}(\tilde{\pi}_{k+1}||\pi_k) + D_{d_\rho^{\pi}}(\pi_{k+1}||\tilde{\pi}_{k+1})}{1-\gamma} \\ &+ \frac{D(\lambda_{k+1}||\tilde{\lambda}_{k+1}) + D(\tilde{\lambda}_{k+1}||\lambda_k)}{\eta'}.\end{aligned}$$

We can obtain

$$\begin{aligned}&\langle \tilde{G}^{v,-\lambda}_{k+1} - \tilde{G}^{v,-\lambda}_k, X_{k+1} - \tilde{X}_{k+1} \rangle \overset{(a)}{\leq} \Psi^*(\tilde{G}^{v,-\lambda}_{k+1} - \tilde{G}^{v,-\lambda}_k)\Psi(X_{k+1} - \tilde{X}_{k+1}) \\ \overset{(b)}{\leq}& \Psi^*(\tilde{G}^{v,-\lambda}_{k+1} - G^{v,-\lambda}_k)\Psi(X_{k+1} - \tilde{X}_{k+1}) + \Psi^*(G^{v,-\lambda}_k - \tilde{G}^{v,-\lambda}_k)\Psi(X_{k+1} - \tilde{X}_{k+1}) \\ \overset{(c)}{\leq}& \beta\Psi(\tilde{X}_{k+1} - X_k)\Psi(X_{k+1} - \tilde{X}_{k+1}) + \beta\Psi(X_k - \tilde{X}_k)\Psi(X_{k+1} - \tilde{X}_{k+1}) \\ \overset{(d)}{\leq}& \frac{\beta}{\sqrt{8}-2}\Psi(\tilde{X}_{k+1} - X_k)^2 + \left(\frac{\beta}{\sqrt{8}+2} + \frac{\beta}{2}\right)\Psi(X_{k+1} - \tilde{X}_{k+1})^2 + \frac{\beta}{2}\Psi(X_k - \tilde{X}_k)^2.\end{aligned}$$

Inequality (a) follows from the Cauchy-Schwarz inequality for the Ψ-norm; (b) from the triangle inequality; (c) from the smoothness of function Φ defined in (C.23); and (d) from $ac + bc \leq \frac{a^2}{\sqrt{8}-2} + \frac{c^2}{\sqrt{8}+2} + \frac{b^2}{2} + \frac{c^2}{2}$.

Since $X_0 = \tilde{X}_0$, $\frac{1}{\sqrt{8}+2} + \frac{1}{2} + \frac{1}{2} = \frac{1}{\sqrt{8}-2}$, and $\Psi(v, \lambda)^2 \leq 2\|v\|^2_\infty + 2\|\lambda\|^2_1$, we have

$$\begin{aligned}&\sum_{k=0}^{K-1} \langle \tilde{G}^{v,-\lambda}_{k+1} - \tilde{G}^{v,-\lambda}_k, X_{k+1} - \tilde{X}_{k+1} \rangle \\ \leq& \frac{\beta}{\sqrt{8}-2}\sum_{k=0}^{K-1}\Psi(\tilde{X}_{k+1} - X_k)^2 + \frac{\beta}{\sqrt{8}-2}\sum_{k=1}^{K}\Psi(X_k - \tilde{X}_k)^2 \\ \leq& \frac{2\beta}{\sqrt{8}-2}\sum_{k=0}^{K-1}\Big(\|V^{\tilde{\pi}_{k+1}}_{1:m}(\rho) - V^{\pi_k}_{1:m}(\rho)\|^2_\infty + \|\tilde{\lambda}_{k+1} - \lambda_k\|^2_1\end{aligned}$$

$$+\|V_{1:m}^{\tilde{\pi}_{k+1}}(\rho) - V_{1:m}^{\pi_{k+1}}(\rho)\|_\infty^2 + \|\tilde{\lambda}_{k+1} - \lambda_{k+1}\|_1^2\Big).$$

Noting that $\frac{2\beta}{\sqrt{8}-2} \leq 3\beta$, by Lemma 28 we have

$$\frac{2\beta}{\sqrt{8}-2}\|V_{1:m}^{\tilde{\pi}_{k+1}}(\rho) - V_{1:m}^{\pi_k}(\rho)\|_\infty^2 \leq \frac{6\gamma^2\beta}{(1-\gamma)^4} D_{d_\rho^{\tilde{\pi}_{k+1}}}(\tilde{\pi}_{k+1}||\pi_k).$$

By Pinsker's inequality, we have

$$\frac{2\beta}{\sqrt{8}-2}\|\tilde{\lambda}_{k+1} - \lambda_{k+1}\|_1^2 \leq 6\beta D(\lambda_{k+1}||\tilde{\lambda}_{k+1}).$$

Since $\alpha \geq \frac{6\beta}{(1-\gamma)^4}$ and $\eta' \leq \frac{1}{6\beta}$, we conclude that $\sum_{k=0}^{K-1} E_k \geq 0$. □

C.3.3.2 Convergence of ARNPG-OMDA

Proof of Theorem 15. By $|\tilde{r}_k(s,a)| = |\langle \tilde{G}_k^v, r_{1:m}(s,a)\rangle| \leq \|\tilde{G}_k^v\|_1 \|r_{1:m}(s,a)\|_\infty \leq L$, we know $\|\tilde{r}_k\|_\infty \leq L$. Taking $\epsilon_k = \frac{\alpha \log(|\mathcal{A}|)}{(1-\gamma)K}$, we choose $t_k = \lceil \frac{1}{1-\gamma} \log(\frac{5LK}{6\beta \log(|\mathcal{A}|)}) + 1 \rceil$.

Then by Proposition 4, for any policy π, we have two fundamental inequalities for the updates $\tilde{\pi}_{k+1}$ and π_{k+1} respectively:

$$\begin{aligned}
V_{\tilde{r}_k}^{\tilde{\pi}_{k+1}}(\rho) - \alpha \frac{D_{d_\rho^{\tilde{\pi}_{k+1}}}(\tilde{\pi}_{k+1}||\pi_k)}{1-\gamma} &\geq V_{\tilde{r}_k}^{\pi}(\rho) - \alpha \frac{D_{d_\rho^\pi}(\pi||\pi_k) - D_{d_\rho^\pi}(\pi||\tilde{\pi}_{k+1})}{1-\gamma} - \epsilon_k, \\
V_{\tilde{r}_{k+1}}^{\pi_{k+1}}(\rho) - \alpha \frac{D_{d_\rho^{\pi_{k+1}}}(\pi_{k+1}||\pi_k)}{1-\gamma} &\geq V_{\tilde{r}_{k+1}}^{\pi}(\rho) - \alpha \frac{D_{d_\rho^\pi}(\pi||\pi_k) - D_{d_\rho^\pi}(\pi||\pi_{k+1})}{1-\gamma} - \epsilon_k.
\end{aligned}$$

Note that $V_{\tilde{r}_k}^{\pi}(\rho) = \langle \tilde{G}_k^v, V_{1:m}^\pi(\rho)\rangle$. Taking $\pi = \pi_{k+1}$ in the first inequality, and summing two inequalities gives

$$\begin{aligned}
\langle \tilde{G}_{k+1}^v, V_{1:m}^{\tilde{\pi}_{k+1}}(\rho) - V_{1:m}^\pi(\rho)\rangle &\geq \alpha \frac{D_{d_\rho^\pi}(\pi||\pi_{k+1}) - D_{d_\rho^\pi}(\pi||\pi_k)}{1-\gamma} - 2\epsilon_k \\
&+ \langle \tilde{G}_k^v - \tilde{G}_{k+1}^v, V_{1:m}^{\pi_{k+1}}(\rho) - V_{1:m}^{\tilde{\pi}_{k+1}}(\rho)\rangle + \alpha \frac{D_{d_\rho^{\tilde{\pi}_{k+1}}}(\tilde{\pi}_{k+1}||\pi_k) + D_{d_\rho^\pi}(\pi_{k+1}||\tilde{\pi}_{k+1})}{1-\gamma}.
\end{aligned} \tag{C.25}$$

We can similarly get the inequality for λ that

$$\langle \tilde{G}_k^\lambda, \tilde{\lambda}_{k+1}\rangle + \frac{D(\tilde{\lambda}_{k+1}||\lambda_k)}{\eta'} \le \langle \tilde{G}_k^\lambda, \lambda\rangle + \frac{D(\lambda||\lambda_k) - D(\lambda||\tilde{\lambda}_{k+1})}{\eta'}, \tag{C.26}$$

$$\langle \tilde{G}_{k+1}^\lambda, \lambda_{k+1}\rangle + \frac{D(\lambda_{k+1}||\lambda_k)}{\eta'} \le \langle \tilde{G}_{k+1}^\lambda, \lambda\rangle + \frac{D(\lambda||\lambda_k) - D(\lambda||\lambda_{k+1})}{\eta'}. \tag{C.27}$$

Taking $\lambda = \lambda_{k+1}$ in the first inequality and summing two inequalities gives

$$\begin{aligned}\langle \tilde{G}_{k+1}^\lambda, \lambda - \tilde{\lambda}_{k+1}\rangle \ge& \frac{D(\lambda||\lambda_{k+1}) - D(\lambda||\lambda_k)}{\eta'}\\ &+ \langle \tilde{G}_k^\lambda - \tilde{G}_{k+1}^\lambda, \tilde{\lambda}_{k+1} - \lambda_{k+1}\rangle + \frac{D(\lambda_{k+1}||\tilde{\lambda}_{k+1}) + D(\tilde{\lambda}_{k+1}||\lambda_k)}{\eta'}.\end{aligned} \tag{C.28}$$

Recall the definition of E_k in (C.24) that

$$\begin{aligned}E_k =& \langle \tilde{G}_k^v - \tilde{G}_{k+1}^v, V_{1:m}^{\pi_{k+1}}(\rho) - V_{1:m}^{\tilde{\pi}_{k+1}}(\rho)\rangle + \alpha\frac{D_{d_\rho^{\tilde{\pi}_{k+1}}}(\tilde{\pi}_{k+1}||\pi_k) + D_{d_\rho^\pi}(\pi_{k+1}||\tilde{\pi}_{k+1})}{1-\gamma}\\ &+ \langle \tilde{G}_k^\lambda - \tilde{G}_{k+1}^\lambda, \tilde{\lambda}_{k+1} - \lambda_{k+1}\rangle + \frac{D(\lambda_{k+1}||\tilde{\lambda}_{k+1}) + D(\tilde{\lambda}_{k+1}||\lambda_k)}{\eta'}.\end{aligned}$$

We then have

$$\begin{aligned}&-\Phi(V_{1:m}^\pi(\rho), \tilde{\lambda}_{k+1}) + \Phi(V_{1:m}^{\tilde{\pi}_{k+1}}(\rho), \lambda)\\ =&\Phi(V_{1:m}^{\tilde{\pi}_{k+1}}(\rho), \tilde{\lambda}_{k+1}) - \Phi(V_{1:m}^\pi(\rho), \tilde{\lambda}_{k+1}) + \Phi(V_{1:m}^{\tilde{\pi}_{k+1}}(\rho), \lambda) - \Phi(V_{1:m}^{\tilde{\pi}_{k+1}}(\rho), \tilde{\lambda}_{k+1})\\ \overset{(a)}{\ge}&\langle \tilde{G}_{k+1}^v, V_{1:m}^{\tilde{\pi}_{k+1}}(\rho) - V_{1:m}^\pi(\rho)\rangle + \langle \tilde{G}_{k+1}^\lambda, \lambda - \tilde{\lambda}_{k+1}\rangle\\ \overset{(b)}{\ge}&\alpha\frac{D_{d_\rho^\pi}(\pi||\pi_{k+1}) - D_{d_\rho^\pi}(\pi||\pi_k)}{1-\gamma} + \frac{D(\lambda||\lambda_{k+1}) - D(\lambda||\lambda_k)}{\eta'} - 2\epsilon_k + E_k.\end{aligned}$$

Inequality (a) is by the concavity of $\Phi(\cdot, \tilde{\lambda}_{k+1})$ and convexity of $\Phi(V_{1:m}^{\tilde{\pi}_{k+1}}(\rho), \cdot)$. Inequality (b) is based on combining (C.25) and (C.28).

Taking $\pi = \pi^*$ and $\lambda = \arg\min_{\lambda' \in \Lambda} \Phi\left(\frac{1}{K}\sum_{k=1}^{K} V_{1:m}^{\tilde{\pi}_k}(\rho), \lambda'\right)$, we have

$$\begin{aligned}
&F\left(\frac{1}{K}\sum_{k=1}^{K} V_{1:m}^{\tilde{\pi}_k}(\rho)\right) = \Phi\left(\frac{1}{K}\sum_{k=1}^{K} V_{1:m}^{\tilde{\pi}_k}(\rho), \lambda\right) \geq \frac{1}{K}\sum_{k=1}^{K} \Phi(V_{1:m}^{\tilde{\pi}_k}(\rho), \lambda)\\
\geq&\frac{1}{K}\sum_{k=0}^{K-1} \Phi(V_{1:m}^{\pi^*}(\rho), \tilde{\lambda}_{k+1}) + \alpha\frac{D_{d_\rho^{\pi^*}}(\pi^*||\pi_K) - D_{d_\rho^{\pi^*}}(\pi^*||\pi_0)}{(1-\gamma)K} + \frac{D(\lambda||\lambda_K) - D(\lambda||\lambda_0)}{\eta' K}\\
&-\frac{2}{K}\sum_{k=0}^{K-1}\epsilon_k + \frac{1}{K}\sum_{k=0}^{K-1} E_k\\
\overset{(a)}{\geq}&F(V_{1:m}^{\pi^*}(\rho)) - \frac{3\alpha\log(|\mathcal{A}|)}{(1-\gamma)K} - \frac{\log(m)}{\eta' K}.
\end{aligned}$$

Inequality (a) is due to $D_{d_\rho^{\pi^*}}(\pi^*||\pi_0) \leq \log(|\mathcal{A}|)$ and Lemma 32. □

Proof of Corollary 8. Note that $T = \sum_{k=0}^{K-1} t_k = \Theta(\frac{K}{1-\gamma}\log(K))$. It implies $\frac{K}{1-\gamma} = \Theta(T/\log(T))$. Substituting this into Theorem 15 concludes Corollary 8. □