# Algorithm Design and Localization Analysis in Sequential and Statistical Learning 

Yunbei Xu

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy under the Executive Committee of the Graduate School of Arts and Sciences

## COLUMBIA UNIVERSITY

© 2023

Yunbei Xu
All Rights Reserved


#### Abstract

Algorithm Design and Localization Analysis in Sequential and Statistical Learning Yunbei Xu


Learning theory is a dynamic and rapidly evolving field that aims to provide mathematical foundations for designing and understanding the behavior of algorithms and procedures that can learn from data automatically. At the heart of this field lies the interplay between algorithm design and statistical complexity analysis, with sharp statistical complexity characterizations often requiring localization analysis. This dissertation aims to advance the fields of machine learning and decision making by contributing to two key directions: principled algorithm design and localized statistical complexity. Our research develops novel algorithmic techniques and analytical frameworks to build more effective and robust learning systems. Specifically, we focus on studying uniform convergence and localization in statistical learning theory, developing efficient algorithms using the optimism principle for contextual bandits, and creating Bayesian design principles for bandit and reinforcement learning problems. The thesis comprises three parts with the following main contents:

In Chapter 1, we study problem-dependent rates, i.e., generalization errors that scale nearoptimally with the variance, the effective loss, or the gradient norms evaluated at the "best hypothesis." We introduce a principled framework dubbed "uniform localized convergence," and characterize sharp problem-dependent rates for central statistical learning problems. From a methodological viewpoint, our framework resolves several fundamental limitations of existing uniform convergence and localization analysis approaches. It also provides improvements and some level
of unification in the study of localized complexities, one-sided uniform inequalities, and samplebased iterative algorithms. In the so-called "slow rate" regime, we provides the first (momentpenalized) estimator that achieves the optimal variance-dependent rate for general "rich" classes; we also establish improved loss-dependent rate for standard empirical risk minimization. In the "fast rate" regime, we establish finite-sample problem-dependent bounds that are comparable to precise asymptotics. In addition, we show that iterative algorithms like gradient descent and firstorder Expectation-Maximization can achieve optimal generalization error in several representative problems across the areas of non-convex learning, stochastic optimization, and learning with missing data.

In Chapter 2, we propose a simple and generic principle to design optimistic algorithms for contextual bandits, dubbed "Upper Counterfactual Confidence Bounds" (UCCB). While existing optimistic algorithms (primarily UCB and its variants) are widely used and successful in multi-armed bandits and reinforcement learning, they often struggle with large context spaces. We demonstrate that our proposed algorithms based on the UCCB principle are provably optimal and efficient in the presence of large context spaces. Key components of UCCB include: 1) a systematic analysis of confidence bounds in policy space rather than in action space; and 2) the potential function perspective that is used to express the power of optimism in the contextual setting. We further show how the UCCB principle can be extended to infinite action spaces, by constructing confidence bounds via the newly introduced notion of "counterfactual action divergence."

In Chapter 3, we develop a general theory to optimize the frequentist regret for sequential learning problems, where efficient bandit and reinforcement learning algorithms can be derived from unified Bayesian principles. We propose a novel optimization approach to create "algorithmic beliefs" at each round, and use Bayesian posteriors to make decisions. This is the first approach to make Bayesian-type algorithms prior-free and applicable to adversarial settings, in a generic and optimal manner. Moreover, the algorithms are simple and often efficient to implement. As a major application, we present a novel algorithm for multi-armed bandits that achieves the "best-of-allworlds" empirical performance in stochastic, adversarial, and non-stationary environments. And
we illustrate how these principles can be used in linear bandits, bandit convex optimization, and reinforcement learning.

## Table of Contents

Acknowledgments ..... ix
Chapter 1: Towards Optimal Problem Dependent Generalization Error Bounds in Statisti- cal Learning Theory ..... 1
1.1 Introduction ..... 1
1.1.1 Background ..... 1
1.1.2 Contributions and organization of the paper ..... 4
1.2 The "uniform localized convergence" procedure ..... 5
1.2.1 The current blueprint ..... 5
1.2.2 The "uniform localized convergence" principle ..... 7
1.2.3 Merits of "uniform localized convergence" ..... 8
1.2.4 Unification and improvements over existing localization approaches ..... 11
1.3 Problem-dependent rates in the "slow rate" regime ..... 13
1.3.1 Preliminaries ..... 13
1.3.2 Loss-dependent rates via empirical risk minimization ..... 16
1.3.3 Variance-dependent rates via moment penalization ..... 18
1.4 Illustrative examples in the "slow rate" regime ..... 22
1.4.1 Discussion and illustrative examples ..... 22
1.4.2 Problem areas to which "localization" theory is applicable ..... 26
1.5 Problem-dependent rates in the parametric "fast rate" regime ..... 28
1.5.1 Background ..... 28
1.5.2 Theoretical foundations ..... 29
1.5.3 Main results ..... 32
1.6 Applications to non-convex learning and stochastic optimization ..... 37
1.6.1 Non-convex learning under curvature conditions ..... 37
1.6.2 Stochastic optimization ..... 41
1.7 Learning with missing data and Expectation-Maximization algorithms ..... 45
1.7.1 Background ..... 46
1.7.2 Problem-dependent rates for first-order EM ..... 48
1.7.3 Discussion and improvements over previous results ..... 51
1.8 Concluding remarks ..... 54
Chapter 2: Upper Counterfactual Confidence Bounds: a New Optimism Principle for Con- textual Bandits ..... 56
2.1 Introduction ..... 56
2.1.1 Motivation. ..... 56
2.1.2 The contextual MAB problem ..... 57
2.1.3 Introducing UCCB: two equivalent viewpoints ..... 59
2.1.4 Related literature ..... 61
2.1.5 Organization ..... 63
2.2 Upper counterfactual confidence bounds ..... 63
2.2.1 The algorithm ..... 63
2.2.2 Key ideas underlying UCCB ..... 64
2.2.3 Proof sketch of Theorem 8 and Lemma 3 ..... 68
2.3 Generalization to infinite $\mathcal{F}$ ..... 70
2.4 A unified framework for infinite action spaces ..... 72
2.4.1 Illustrative models ..... 73
2.4.2 Counterfactual action divergence ..... 75
2.4.3 The algorithm and regret bound ..... 76
2.4.4 Applications in illustrative examples ..... 79
2.5 Using "optimistic subroutines" to generalize randomized algorithms ..... 82
2.6 Conclusion and future directions ..... 85
Chapter 3: Bayesian Design Principles for Frequentist Sequential Learning ..... 87
3.1 Introduction ..... 87
3.1.1 Background ..... 87
3.1.2 Contributions ..... 89
3.1.3 Related literature ..... 91
3.2 Preliminaries and definition of AIR ..... 91
3.2.1 Problem formulation ..... 91
3.2.2 Algorithmic Information Ratio ..... 93
3.2.3 Relation to IR and DEC ..... 95
3.3 Algorithms ..... 97
3.3.1 A generic regret bound leveraging AIR ..... 97
3.3.2 Adaptive Posterior Sampling (APS) ..... 98
3.3.3 Adaptive Minimax Sampling (AMS) ..... 100
3.3.4 Using approximate maximizers ..... 101
3.4 Application to Bernoulli MAB ..... 101
3.4.1 Simplified APS for Bernoulli MAB ..... 102
3.4.2 Numerical experiments ..... 103
3.5 Key intuitions of the proof ..... 108
3.5.1 Bounding regret by sum of AIR ..... 108
3.5.2 Mimimax theory: from value to construction ..... 109
3.6 Applications to infinite-armed bandits ..... 111
3.6.1 Maximization of AIR for structured bandits ..... 112
3.6.2 Application to Gaussian linear bandits ..... 112
3.6.3 Application to bandit convex optimization ..... 114
3.7 Model-index AIR and application to RL ..... 115
3.7.1 Model-index Algorithmic Information Ratio ..... 116
3.7.2 Near-optimal algorithmic beliefs in closed form ..... 118
3.7.3 Model-index APS ..... 119
3.7.4 Application to reinforcement learning ..... 121
3.8 Conclusion and future directions ..... 124
Conclusion ..... 126
References ..... 127
Appendix A: Appendix for Chapter 1 ..... 138
A. 1 Fast rates in supervised learning with structured convex cost ..... 141
A.1.1 Background ..... 142
A.1.2 Main results and illustrative examples ..... 148
A.1.3 Contributions relative to previous approaches ..... 154
A. 2 Proofs for Section 1.2 and Section 1.3 ..... 158
A.2.1 Proofs for Proposition 1 and its variants ..... 158
A.2.2 Proof of Theorem 1 ..... 161
A.2.3 Estimating loss-dependent rates from data ..... 164
A.2.4 Proof of Theorem 2 ..... 167
A.2.5 Estimating variance-dependent rates from data ..... 179
A.2.6 Auxiliary lemmata ..... 182
A. 3 Proofs for Section 1.5, Section 1.6 and Section 1.7 ..... 182
A.3.1 Proof of Lemma 2 ..... 182
A.3.2 Proof of Proposition 2 ..... 183
A.3.3 Proof of Theorem 3 ..... 187
A.3.4 Proof of Theorem 4 ..... 191
A.3.5 Proof of Corollary 5 ..... 196
A.3.6 Proof of Theorem 6 ..... 197
A.3.7 Proof of Corollary 7 ..... 201
A.3.8 Auxiliary definitions and lemmata ..... 206
A. 4 Proofs for Section A. 1 ..... 208
A.4.1 Proof of Theorem 20 ..... 208
A.4.2 Proof of Corollary 21 ..... 218
Appendix B: Appendix for Chapter 2 ..... 220
B. 1 Proofs for the finite-action setting ..... 220
B.1.1 Proof of Theorem 8. ..... 220
B.1.2 Analysis on the confidence ..... 221
B.1.3 Proof of Lemma 4 ..... 226
B. 2 Proofs for the extensions to infinite function classes ..... 226
B.2.1 Proof of Corollary 9 ..... 226
B.2.2 Proof of Corollary 10 ..... 228
B. 3 Proofs for the infinite-action setting ..... 229
B. 4 Proofs for the "optimistic subroutine" in Section 2.5 ..... 233
Appendix C: Appendix for Chapter 3 ..... 238
C. 1 Extensions and proofs for AIR ..... 238
C.1.1 Extensions of AIR ..... 238
C.1.2 Proof of Theorem 14 ..... 239
C.1.3 Proof of Theorem 16 ..... 241
C.1.4 Proof of Theorem 15 ..... 241
C.1.5 Proof of Theorem 17 and Theorem 25 ..... 243
C. 2 Details and proofs for bandit problems ..... 244
C.2.1 Concave parameterization with Bernoulli reward ..... 244
C.2.2 Simplified APS for Bernoulli MAB ..... 246
C.2.3 Surrogate concave objective with Gaussian reward ..... 248
C.2.4 Simplified APS for Gaussian linear bandits, relationship with IPW ..... 250
C. 3 Proofs for MAIR ..... 251
C.3.1 Proof of Lemma 13 ..... 251
C.3.2 Analysis of sequential estimation ..... 252
C.3.3 Proof for Theorem 18 ..... 254
C.3.4 Proof of Theorem 19 ..... 257
C. 4 Technical backgrounds ..... 258
C.4.1 Conditional entropy ..... 258
C.4.2 Minimax theorem ..... 258
C.4.3 Convex analysis ..... 259
C.4.4 Concentration inequality ..... 260
C.4.5 Information theory ..... 260

## List of Figures

3.1 Sensitivity analysis in a stochastic bandit problem. ..... 104
3.2 Sensitivity analysis in an adversarial bandit problem. ..... 105
3.3 Sensitivity analysis in a "change points" environment. ..... 106
3.4 Comparing APS to "clairvoyant" restarted algorithms in a "change points" envi- ronment. ..... 106
3.5 "Sine curve" reward sequences for 4 arms. ..... 107
3.6 Regret curves in a "sine curve" environment. ..... 107
3.7 Tracking selected arms of APS in a "sine curve" environment. ..... 107
3.8 Tracking selected arms of TS in a "sine curve" environment. ..... 107

## Acknowledgements

Completing this doctoral thesis has been an arduous yet immensely rewarding journey, and I have many people to thank for their support and guidance along the way. First and foremost, I am deeply grateful to my advisor Prof. Assaf Zeevi, who has been an incredible mentor to me. Assaf gave me the opportunity to explore my own interests, and helped me navigate the intricacies of research. His consistent encouragement, generosity, and kindness have made an indelible impact on both my scholarly and personal growth. Assaf, I extend my deepest gratitude for your invaluable advice, inspirational vision, and warm friendship. You have not only imparted profound knowledge but have also shared wisdom that extends far beyond the realm of academia.

I would like to express my sincere gratitude to the members of the DRO division at Columbia University for creating an intellectually stimulating and collaborative environment that fostered learning and research. To my fellow students, I want to extend my heartfelt thanks for the enjoyable moments, lively discussions, constructive feedback, and fruitful collaborations. You have made my graduate school experience unforgettable and enjoyable. I am profoundly appreciative of the faculty members at DRO and the machine learning community at Columbia. Your brilliant ideas and unique perspectives have made a significant contribution to the depth and quality of my research. I would also like to express my special gratitude to my thesis committee members and recommenders: Hongseok Namkoong, Daniel Russo, Jing Dong, Shipra Agrawal, and Wotao Yin. Your expertise and support have been invaluable in shaping my research. To my esteemed collaborators, I offer my sincere appreciation, and I eagerly look forward to our continued collaboration in the future.

Finally, I would like to express my heartfelt gratitude to my brother, Yunzong, and my parents, Xianchun and Hua. Their unwavering love and support have been my guiding light throughout this journey. They have shaped my character, instilled in me the values I hold dear, and provided unwavering encouragement to pursue my dreams. Their presence has been a constant source of strength and inspiration. This work is dedicated to them with profound gratitude and love.

# Chapter 1: Towards Optimal Problem Dependent Generalization Error Bounds in Statistical Learning Theory 

We study problem-dependent rates, i.e., generalization errors that scale near-optimally with the variance, the effective loss, or the gradient norms evaluated at the "best hypothesis." We introduce a principled framework dubbed "uniform localized convergence," and characterize sharp problem-dependent rates for central statistical learning problems. From a methodological viewpoint, our framework resolves several fundamental limitations of existing uniform convergence and localization analysis approaches. It also provides improvements and some level of unification in the study of localized complexities, one-sided uniform inequalities, and sample-based iterative algorithms. In the so-called "slow rate" regime, we provides the first (moment-penalized) estimator that achieves the optimal variance-dependent rate for general "rich" classes; we also establish improved loss-dependent rate for standard empirical risk minimization. In the "fast rate" regime, we establish finite-sample problem-dependent bounds that are comparable to precise asymptotics. In addition, we show that iterative algorithms like gradient descent and first-order ExpectationMaximization can achieve optimal generalization error in several representative problems across the areas of non-convex learning, stochastic optimization, and learning with missing data.

### 1.1 Introduction

### 1.1.1 Background

Problem statement. Consider the following statistical learning setting. Assume that a random sample $z$ follows an unknown distribution $\mathbb{P}$ with support $\mathcal{Z}$. For each realization of $z$, let $\ell(\cdot ; z)$ be a real-valued loss function, defined over the hypothesis class $\mathcal{H}$. Let $h^{*} \in \mathcal{H}$ be the optimal
hypothesis that minimizes the population risk

$$
\mathbb{P} \ell(h ; z):=\mathbb{E}[\ell(h ; z)] .
$$

Given $n$ i.i.d. samples $\left\{z_{i}\right\}_{i=1}^{n}$ drawn from $\mathbb{P}$, our goal, roughly speaking, is to "learn" a hypothesis $\hat{h} \in \mathcal{H}$ that makes the generalization error

$$
\mathcal{E}(\hat{h}):=\mathbb{P} \ell(\hat{h} ; z)-\mathbb{P} \ell\left(h^{*} ; z\right)
$$

as small as possible. This pursuit is ubiquitous in machine learning, statistics and stochastic optimization.

We study problem-dependent rates, i.e., finite-sample generalization errors that scale tightly with problem-dependent parameters, such as the variance, the effective loss, or the gradient norms at the optimal hypothesis $h^{*}$. While the direct dependence of $\mathcal{E}(\hat{h})$ on the sample size $n$ is often well-understood, it typically only reflects an "asymptotic" perspective, placing less emphasis on the scale of problem-dependent parameters. Existing literature leaves several outstanding challenges in deriving problem-dependent rates. These can be broadly categorized into the so-called "slow rate" and "fast rate" regimes, as described below.

Main challenges in the "slow rate" regime. In the absence of strong convexity and margin conditions, the direct dependence of $\mathcal{E}(\hat{h})$ on the sample size $(n)$ is typically no faster than $O\left(n^{-\frac{1}{2}}\right)$. These settings are referred to as the "slow rate" regime. Here, the central challenge is to simultaneously achieve optimal dependence on problem-dependent parameters (the variance or the effective loss at the optimal $h^{*}$ ) and the sample size $(n)$. To the best of our knowledge, this has not been achieved in previous literature for "rich" classes (to be explained shortly).

Perhaps the most widely used framework to study problem-dependent rates is the traditional "local Rademacher complexity" analysis [1, 2, 3], which has become a standard tool in learning theory. However, as we will discuss later, this analysis leads to a dependence on the sample
size ( $n$ ) which is sub-optimal for essentially all "rich" classes (with the exception of parametric classes). The absence of more precise localization analysis also challenges the design of more refined estimation procedures. For example, designing estimators to achieve variance-dependent rates requires penalizing the empirical second moment to achieve the "right" bias-variance tradeoff. Most antecedent work is predicated on either the traditional "local Rademacher complexity" analysis $[4,5]$ or coarser approaches [6, 7]. Thus, to the best of our knowledge, the question of optimal problem-dependent rates for general rich classes is still open.

Main challenges in the "fast rate" regime. When assuming suitable curvature or so-called margin conditions, the direct dependence of $\mathcal{E}(\hat{h})$ on $n$ is typically faster than $O\left(n^{-\frac{1}{2}}\right)$; for that reason we refer to this as the "fast rate" regime. Here, the traditional localization analysis often provides correct dependence on the sample size ( $n$ ), but the complexity parameters (e.g., norms of gradients, Lipchitz constants, etc.) characterizing these generalization errors are not localized and hence are not problem-dependent.

Much progress on problem-dependent rates has been made under particular formulations, such as regression with structured strongly convex cost (e.g., square cost, Huber cost) [8, 9, 10], and binary classification under margin conditions [11, 12]. In particular, [8, 9] point out that the traditional "local Rademacher complexity" analysis fails to provide parameter localization for unbounded regression problems, and propose the so-called "learning without concentration" methodology to obtain problem-dependent rates that do not scale with the worst-case parameters. We are able to recover these results using a more direct concentration-based analysis and remove certain restrictions.

We will also focus on parametric models in the "fast rate" regime, which covers more "modern" non-convex learning problems whose analysis is not aligned with traditional generalization error analysis. For example, in non-convex learning problems one wants generalization error bounds for iterative optimization algorithms; and traditional localization frameworks (which mostly focus on supervised learning) do not cover general stochastic optimization and missing data problems. For
many representative non-convex learning, stochastic optimization and missing data problems, it remains open to provide algorithmic solutions and problem-dependent generalization error bounds that are comparable to optimal asymptotic results.

### 1.1.2 Contributions and organization of the paper

The paper provides contributions both in the framework it develops, as well as its application to improving existing results in several problem areas. In particular, it suggests guidelines for designing estimators and learning algorithms and provides analysis tools to study them. In addition, it provides some level of unification across problem areas. Specifically, the main contributions are as follows.
(1) We introduce a principled framework to study localization in statistical learning, dubbed the "uniform localized convergence" procedure, which simultaneously provides optimal "direct dependence" on the sample size, correct scaling with problem-dependent parameters, and flexibility to unify various problem settings. Section 1.2 provides a description of the framework, and explains how it addresses several fundamental challenges.
(2) In the "slow rate" regime, we employ the above ideas to design the first estimator that achieves optimal variance-dependent rates for general function classes. The derivation is based on a novel procedure that optimally penalizes the empirical (centered) second moment. We also establish improved loss-dependent rates for standard empirical risk minimization, which has computational advantages. Section 1.3 presents these theoretical results (see Section 1.3.2 for the lossdependent rate and Section 1.3.3 for the variance-dependent rate); and Section 1.4 will illustrate our improvements in non-parametric classes and VC classes.
(3) In the "fast rate" regime, we establish a "uniform localized convergence of gradient" framework for parametrized models, and characterize optimal problem-dependent rates for approximate stationary points and iterative optimization algorithms such as gradient descent and first-order Expectation-Maximization. Our results scale tightly with the gradient norms at the optimal parameter, and improve previous guarantees in non-convex learning, stochastic optimization and missing
data problems. See Section 1.5 for the theoretical results and Sections 1.6-1.7 for illustrations of improvements over previous results.
(4) In the "fast rate" regime, we also study supervised learning with structured convex cost, where the hypothesis class can be non-parametric and heavy-tailed. This part of the work has direct relationship to a stream of literature known as "learning without concentration." Our contributions in this setting lie in technical refinements of the generalization error bounds and some unification between one-sided uniform inequalities and concentration of truncated functions; for this reason we defer its treatment to Appendix A.1.

Most formal proofs in the paper are deferred to Appendix A. 2 and Appendix A.3. We provide a guide for reading them in the beginning of the appendix, highlighting the core messages and steps of each technical proof.

### 1.2 The "uniform localized convergence" procedure

### 1.2.1 The current blueprint

Denote the empirical risk

$$
\mathbb{P}_{n} \ell(h ; z):=\frac{1}{n} \sum_{i=1}^{n} \ell\left(h ; z_{i}\right),
$$

and consider the following straightforward decomposition of the generalization error

$$
\begin{equation*}
\mathcal{E}(\hat{h})=\left(\mathbb{P}-\mathbb{P}_{n}\right) \ell(\hat{h} ; z)+\left(\mathbb{P}_{n} \ell(\hat{h} ; z)-\mathbb{P}_{n} \ell\left(h^{*} ; z\right)\right)+\left(\mathbb{P}_{n}-\mathbb{P}\right) \ell\left(h^{*} ; z\right) . \tag{1.2.1}
\end{equation*}
$$

The main difficulty in studying $\mathcal{E}(\hat{h})$ comes from bounding the first term $\left(\mathbb{P}-\mathbb{P}_{n}\right) \ell(\hat{h} ; z)$, since $\hat{h}$ depends on the $n$ samples. The simplest approach, which does not achieve problem-dependent rates, is to bound the uniform error

$$
\sup _{h \in \mathcal{H}}\left(\mathbb{P}-\mathbb{P}_{n}\right) \ell(h ; z)
$$

over the entire hypothesis class $\mathcal{H}$. In order to obtain problem-dependent rates, a natural modification is to consider uniform convergence over localized subsets of $\mathcal{H}$.

We first give an overview of the traditional "local Rademacher complexity" analysis [1, 13, 3]. Consider a generic function class $\mathcal{F}$ that we wish to concentrate, which consists of real-valued functions defined on $\mathcal{Z}$ (e.g., one can set $f(z)=\ell(h ; z)$ ). Denote

$$
\mathbb{P} f:=\mathbb{E}[f(z)], \quad \mathbb{P}_{n} f:=\frac{1}{n} \sum_{i=1}^{n} f\left(z_{i}\right)
$$

The notation $r>0$ will serve as a localization parameter, and $\delta>0$ will serve for high probability arguments. Let $\psi(r ; \delta)$ denote a surrogate function that upper bounds the uniform error within a localized region $\{f \in \mathcal{F}: T(f) \leq r\}$, where we call $T: \mathcal{F} \rightarrow \mathbb{R}_{+}$the "measurement functional." Formally, let $\psi$ be a function that maps $[0, \infty) \times(0,1)$ to $[0, \infty)$, which possibly depends on the observed samples $\left\{z_{i}\right\}_{i=1}^{n}$. Assume $\psi$ satisfies for arbitrary fixed $\delta, r$, with probability at least $1-\delta$,

$$
\begin{equation*}
\sup _{f \in \mathcal{F}: T(f) \leq r}\left(\mathbb{P}-\mathbb{P}_{n}\right) f \leq \psi(r ; \delta) . \tag{1.2.2}
\end{equation*}
$$

By default we ask $\psi(r ; \delta)$ to be a non-decreasing and non-negative function. ${ }^{1}$ The main result of the traditional localization analysis can be summarized as follows. (The statement is obtained by adapting the proof from Section 3.2 in [1]; itself more general than the traditional meta-results [1, 13, 3].)

Statement 1 (current blueprint) Assume that $\psi$ is a sub-root function, i.e., $\psi(r ; \delta) / \sqrt{r}$ is nonincreasing with respect to $r \in \mathbb{R}_{+}$. Assume the Bernstein condition $T(f) \leq B_{e} \mathbb{P} f, B_{e}>0$, $\forall f \in \mathcal{F}$. Then with probability at least $1-\delta$, for all $f \in \mathcal{F}$ and $K>1$,

$$
\begin{equation*}
\left(\mathbb{P}-\mathbb{P}_{n}\right) f \leq \frac{1}{K} \mathbb{P} f+\frac{100(K-1) r^{*}}{B_{e}} \tag{1.2.3}
\end{equation*}
$$

where $r^{*}$ is the "fixed point" solution of the equation $r=B_{e} \psi(r ; \delta)$.

[^0]Since its inception, Statement 1 has become a standard tool in learning theory. However, it requires a rather technical proof, and it appears to be loose when compared with the original assumption (1.2.2)—ideally, we would like to directly extend (1.2.2) to hold uniformly without sacrificing any accuracy. Moreover, some assumptions in the statement are restrictive and might not be necessary.

### 1.2.2 The "uniform localized convergence" principle

We provide a surprisingly simple analysis that greatly improves and simplifies Statement 1. Unlike Statement 1, the following proposition does not require the concentrated functions $g_{f}$ to satisfy the Bernstein condition, and the surrogate function $\psi$ need not to be "sub-root." Despite placing less restrictions, Proposition 1 is able to establish results that are typically "sharper" than the current blueprint. Note that in the proposition, both the measurement functional $T$ as well as the surrogate function $\psi$ are allowed to be sample-dependent.

Proposition 1 (the "uniform localized convergence" argument) For a function class $\mathcal{G}=\left\{g_{f}\right.$ : $f \in \mathcal{F}\}$ and functional $T: \mathcal{F} \rightarrow[0, R]$, assume there is a function $\psi(r ; \delta)$, which is nondecreasing with respect to $r$ and satisfies that $\forall \delta \in(0,1), \forall r \in[0, R]$, with probability at least $1-\delta$,

$$
\begin{equation*}
\sup _{f \in \mathcal{F}: T(f) \leq r}\left(\mathbb{P}-\mathbb{P}_{n}\right) g_{f} \leq \psi(r ; \delta) \tag{1.2.4}
\end{equation*}
$$

Then, given any $\delta \in(0,1)$ and $r_{0} \in(0, R]$, with probability at least $1-\delta$, for all $f \in \mathcal{F}$, either $T(f) \leq r_{0}$ or

$$
\begin{equation*}
\left(\mathbb{P}-\mathbb{P}_{n}\right) g_{f} \leq \psi\left(2 T(f) ; \delta\left(\log _{2} \frac{2 R}{r_{0}}\right)^{-1}\right) \tag{1.2.5}
\end{equation*}
$$

The key intuition behind Proposition 1 is that the uniform restatement of the "localized" argument (1.2.4) is nearly cost-free, because the deviations $\left(\mathbb{P}-\mathbb{P}_{n}\right) g_{f}$ can be controlled solely by the real valued functional $T(f)$. As a result, we essentially only require uniform convergence over an interval $\left[r_{0}, R\right]$. The "cost" of this uniform convergence, namely, the additional $\log _{2}\left(\frac{2 R}{r_{0}}\right)$ term
in (1.2.5), will only appear in the form $\log \left(\delta / \log _{2}\left(\frac{2 R}{r_{0}}\right)\right)$ in high-probability bounds, which is of a negligible $O(\log \log n)$ order in general.

Formally, we apply a "peeling" technique: we take $r_{k}=2^{k} r_{0}$, where $k=1,2, \ldots,\left\lceil\log _{2} \frac{R}{r_{0}}\right\rceil$, and we use the union bound to extend (1.2.4) to hold for all these $r_{k}$. Then for any $f \in \mathcal{F}$ such that $T(f)>r_{0}$ is true, there exists a non-negative integer $k$ such that $2^{k} r_{0}<T(f) \leq 2^{k+1} r_{0}$. By the non-decreasing property of the $\psi$ function, we then have

$$
\left(\mathbb{P}-\mathbb{P}_{n}\right) g_{f} \leq \psi\left(r_{k+1} ; \delta\left(\log _{2} \frac{2 R}{r_{0}}\right)^{-1}\right) \leq \psi\left(2 T(f) ; \delta\left(\log _{2} \frac{2 R}{r_{0}}\right)^{-1}\right)
$$

which is exactly (1.2.5). Interestingly, the proof of the classical result (Statement 1) relies on a relatively heavy machinery that includes more complicated peeling and re-weighting arguments (see Section 3.1.4 in [1]). However, that analysis obscures the key intuition that we elucidate under inequality (1.2.5).

In this paper, we prove localized generalization error bounds through a unified principle, summarized at a high level in the two-step template below.

Principle of uniform localized convergence. First, determine the concentrated functions, the measurement functional and the surrogate $\psi$, and obtain a sharp "uniform localized convergence" argument. Then, perform localization analysis that is customized to the problem setting and the learning algorithm.

Distinct from the blueprint (1.2.3), the right hand side of our "uniform localized convergence" argument (1.2.5) contains a "free" variable $T(f)$ rather than a fixed value $r$. The new principle strictly improves the current blueprint from many aspects, and its merits will be illustrated in the sequel.

### 1.2.3 Merits of "uniform localized convergence"

Our improvements in the "slow rate" regime originate from the noticeable gap between Proposition 1 and Statement 1, illustrated by the following (informal) conclusion.

Statement 2 (improvements over the current blueprint-informal statement) Setting $g_{f}=f$, then under the assumptions of Statement 1, Proposition 1 provides a strict improvement over Statement 1. In particular, the slower $\psi$ grows, the larger the gap between the bounds in the two results, and the bounds are on pair only when $\psi$ is proportional to $\sqrt{r}$, i.e., when the function class $\mathcal{F}$ is parametric and not "rich."

Formalizing as well as providing rigorous justification for this conclusion is relatively straightforward: taking the "optimal choice" of $K$ in Statement 1, we can re-write its conclusion as

$$
\left(\mathbb{P}-\mathbb{P}_{n}\right) f \leq 20 \sqrt{\frac{r^{*} \mathbb{P} f}{B_{e}}}-\frac{r^{*}}{B_{e}} \quad[\text { Statement } 1]
$$

where the right hand side is of order $\sqrt{r^{*} \mathbb{P} f / B_{e}}$ when $\mathbb{P} f<r^{*} / B_{e}$, and order $r^{*} / B_{e}$ when $\mathbb{P} f \leq$ $r^{*} / B_{e}$. Our result (1.2.5) is also of order $r^{*} / B_{e}$ when $\mathbb{P} f \leq r^{*} / B_{e}$. However, for every $f$ such that $\mathbb{P} f>r^{*} / B_{e}$, it is straightforward to verify that under the assumptions in Statetment 1,

$$
\begin{align*}
\psi(2 T(f) ; \delta) & \leq \psi\left(2 B_{e} \mathbb{P} f ; \delta\right) \quad\left[\text { Bernstein condition: } T(f) \leq B_{e} \mathbb{P} f\right] \\
& \leq \frac{\sqrt{2 B_{e} \mathbb{P} f}}{\sqrt{r^{*}}} \psi\left(r^{*} ; \delta\right) \quad[\psi(r ; \delta) \text { is sub-root }] \\
& \leq \sqrt{\frac{2 r^{*} \mathbb{P} f}{B_{e}}} \quad\left[r^{*} \text { is the fixed point of } B \psi(r ; \delta)\right] . \tag{1.2.6}
\end{align*}
$$

Therefore, the argument $\psi(2 T(f) ; \delta) \leq \sqrt{2 r^{*} \mathbb{P} f / B_{e}}$ established by (1.2.6) shows that the "uniform localized convergence" argument (1.2.5) strictly improves over Statement 1 (ignoring negligible $O(\log \log n)$ factors $)$. Statement 2 indicates that the folklore use of fixed values as straightforward complexity control is somewhat questionable. In the "slow rate" regime, the key point to achieve optimal problem-dependent rates is to bound the generalization error using the function $\psi$. Otherwise the bounds will have the "wrong" dependence on the sample size for all "rich" classes.

Interestingly, the merits of our approach in the "fast rate" regime stem from very different perspectives: the removal of the "sub-root" requirement on $\psi$ allows one to achieve parameter localization; and the added flexibility in the choice of $g_{f}$ allows one to prove one-sided uniform
inequalities and uniform convergence of gradient vectors. To better appreciate these, we provide the following informal discussion to help elucidate the key messages. Let $\mu>0$ be the curvature parameter in the "fast rate" regime (a common example is the strong convexity parameter for the loss function). The generalization error is often characterized by the fixed point of $\psi(r ; \delta) / \mu$, where $\psi$ is the surrogate function that governs the uniform error of excess risk. The removal of the "sub-root" restriction is crucial because under curvature and smoothness conditions, the uniform error of excess loss typically grows "faster" than the square root function. Consider surrogate functions of the form

$$
\begin{equation*}
\psi(r ; \delta)=\underbrace{\sqrt{a_{n}^{*} r}}_{\text {problem-dependent }}+\underbrace{c_{n} r}_{\text {linear }} \tag{1.2.7}
\end{equation*}
$$

where $a_{n}^{*}$ is a problem-dependent rate, and $c_{n}$ satisfies $0<c_{n}<1 /(2 \mu)$ when the sample size $n$ satisfies mild conditions. We call $c_{n} r$ with $0<c_{n}<1 /(2 \mu)$ the benign linear component in the sense that when solving the fixed point equation $r=\psi(r ; \delta) / \mu$, that part can be dropped from both side of the equation. As a result, the fixed point solution only depends on order of the problem-dependent component $\sqrt{a_{n}^{*} r}$, and so one obtains problem-dependent rates. In contrast, worst-case parameters will be unavoidable if one wants to use a "sub-root" surrogate function to govern (1.2.7). In traditional analysis, a loose "sub-root" surrogate function is often obtained via two-sided concentration and Lipchitz contraction, making global Lipchitz constants unavoidable. Furthermore, the added flexibility to choose concentrated functions $g_{f}$ is also useful. In particular, we will show that: 1) our framework unifies traditional value-based uniform convergence and uniform convergence of gradient vectors, which is crucial to study non-convex learning problems and sample-based iterative algorithms; and 2) simple "truncated" functions can be used to established one-sided uniform inequalities that are sharper than two-sided ones, which enable recovery of results in unbounded and heavy-tailed regression problems.

### 1.2.4 Unification and improvements over existing localization approaches

Beyond proving new bounds, an important objective of the paper is to provide some level of unification to the methodology of uniform convergence and localization. Here we present a historical review of uniform convergence and localization, and discuss how the "uniform localized convergence" principle unifies and improves existing approaches. We will overview four general settings where localization plays a crucial role in generalization error analysis and algorithm design: 1) the "slow rate" regime; 2) classification under margin conditions; 3) regression under curvature conditions; and 4) non-convex learning and stochastic optimization (the latter three settings belong to the "fast rate" regime).

The "slow rate" regime. The traditional "local Rademacher complexity" analysis is the standard tool to study localized generalization error bounds in the "slow rate" regime; it also influences the design of regularization. Here, our framework resolves the fundamental limitation of the traditional analysis, leading to the first optimal loss-dependent and variance-dependent rates for general "rich" classes (see Section 1.3-1.4).

Classification under margin conditions. One early line of work in the "fast rate" regime focuses on exploiting the margin conditions to establish fast rates in binary classification (e.g., see Section 5.3 in [13]). It can be shown that the Bernstein condition in Statement 1 subsumes standard margin conditions such as Tsybakov's margin condition [16] and Massart's noise condition [17]. Because the loss functions in binary classification are uniformly bounded, the "current blueprint" (Statement 1) already provides a good framework to study these questions. In an orthogonal direction, some recent works study more refined complexity measures as alternatives to the notion "local Rademacher complexity" [12, 18]. While these results are important, they are within the scope of Statement 1 from the perspective of localization machinery. In this setting, the work presented in this paper contributes to unification rather than improvements over specific theoretical results.

Connections to the "learning without concentration" paradigm. In regression problems under curvature conditions, the traditional localization analysis has been widely applied to achieve "fast rates." However, it fails to localize the complexity parameters (e.g., norms of gradients, Lipchitz constants, etc.) in the generalization error bounds. As a consequence, the traditional localization analysis is widely recognized as not being suitable for regression problems with unbounded or heavy-tailed losses; and it may be unfavorable even for uniformly bounded problems because it fails to adapt to the "effective noise level" at the optimal hypothesis. Important progress addressing the aforementioned limitations has been made during the last decade. Focusing on supervised learning problems with structured convex costs (square cost, Huber cost, etc.), the breakthrough works [8, 9] propose the so-called "learning without concentration" framework, where the central notions and proof techniques are quite different from the traditional concentration framework. An interesting direction is to recover these results by directly strengthening the traditional concentration framework, explicitly figuring out which component of the concentrated functions contributes to which part of the surrogate function. By exploiting the intrinsic connections between one-sided uniform inequalities and the "uniform localized convergence" of truncated functions, we are able to establish such a unified analysis and illustrate systematical refinements of the generalization error bounds; this development is deferred to Appendix A. 1 due to space constraints.

Non-convex learning and stochastic optimization. Traditional value-based generalization error analysis relies on properties of global (regularized) empirical risk minimizers. However, nonconvex learning problems and generalization error analysis of iterative optimization algorithms require one to consider uniform convergence of gradient vectors. And it should be noted that existing localization frameworks mostly focus on supervised learning problems and are unable to handle general stochastic optimization or missing data problems. Our framework is able to prove "uniform localized convergence of gradients," which improves existing vector-based uniform convergence frameworks [19, 20, 21] and provides problem-dependent generalization error bounds for iterative algorithms (See Sections 1.5-1.7).

### 1.3 Problem-dependent rates in the "slow rate" regime

### 1.3.1 Preliminaries

Let $\mathcal{V}^{*}$ and $\mathcal{L}^{*}$ be the variance and the "effective loss" at the best hypothesis $h^{*}$ :

$$
\mathcal{V}^{*}:=\operatorname{Var}\left[\ell\left(h^{*} ; z\right)\right], \quad \mathcal{L}^{*}:=\mathbb{P}\left[\ell\left(h^{*} ; z\right)-\inf _{\mathcal{H}} \ell(h ; z)\right] .
$$

In this section we study finite-sample generalization errors that scale tightly with $\mathcal{V}^{*}$ or $\mathcal{L}^{*}$, which we call problem-dependent rates, without invoking strong convexity or margin conditions.

In the slow rate regime, we assume the loss function to be uniformly bounded by $[-B, B]$, i.e., $|\ell(h ; z)| \leq B$ for all $h \in \mathcal{H}$ and $z \in \mathcal{Z}$. This is a standard assumption used in almost all previous works that do not invoke curvature conditions or rely on other problem-specific structure; and our results in the slow rate regime essentially only require this boundedness assumption. Extensions to unbounded targets can be obtained via truncation techniques (see, e.g. [22]), and our problemdependent results allow $B$ to be very large, potentially scaling with $n$.

We represent the complexity through a surrogate function $\psi(r ; \delta)$ that satisfies for all $\delta \in(0,1)$,

$$
\begin{equation*}
\sup _{f \in \mathcal{F}: \mathbb{P}\left[f^{2}\right] \leq r}\left(\mathbb{P}-\mathbb{P}_{n}\right) f \leq \psi(r ; \delta), \tag{1.3.1}
\end{equation*}
$$

with probability at least $1-\delta$, where $\mathcal{F}$ is taken to be the excess loss class

$$
\begin{equation*}
\ell \circ \mathcal{H}-\ell \circ h^{*}:=\left\{z \mapsto\left(\ell(h ; z)-\ell\left(h^{*} ; z\right)\right): h \in \mathcal{H}\right\} . \tag{1.3.2}
\end{equation*}
$$

(From the perspective of Section 1.2.2, we choose the excess losses as the "concentrated functions," and use $T(f)=\mathbb{P}\left[f^{2}\right]$ as the "measurement functional".) To achieve non-trivial complexity control (and ensure existence of the fixed point), we only consider "meaningful" surrogate functions, as stated below.

Definition 1 (meaningful surrogate function) A bivariate function $\psi(r ; \delta)$ defined over $[0, \infty) \times$
$(0,1)$ is called a meaningful surrogate function if it is non-decreasing, non-negative and bounded with respect to $r$ for every fixed $\delta \in(0,1)$.

We note that the above does not place significant restrictions on the choice of the surrogate function. In particular, for the $\psi$ function defined in (1.3.1) and the excess loss class in (1.3.2), the left hand side of (1.3.1) is itself non-decreasing and non-negative; and the boundedness requirement can always be met by setting $\psi(r ; \delta)=\psi\left(4 B^{2} ; \delta\right)$ for all $r \geq 4 B^{2}$. We now give the formal definition of fixed points.

Definition 2 (fixed point) Given a non-decreasing, non-negative and bounded function $\varphi(r) d e$ fined over $[0, \infty)$, we define the fixed point of $\varphi(r)$ to be $\sup \{r>0: \varphi(r) \geq r\}$.

It is well-known that a non-decreasing, non-negative and bounded function only has finite discontinuity points, all of which belong to "discontinuity points of the first kind" (see Definition 4.26 in [23] for more background). Therefore, it is easy to verify that the fixed point of $\varphi(r)$ is the maximal solution to the equation $\varphi(r)=r$.

Given a bounded class $\mathcal{F}$, empirical process theory provides a general way to construct surrogate function by upper bounding the "local Rademacher complexity" $\mathfrak{R}\left\{f \in \mathcal{F}: \mathbb{P}\left[f^{2}\right] \leq r\right\}$ (see Lemma 16 in Appendix A.2.6). We give the definition of Rademacher complexity for completeness.

Definition 3 (Rademacher complexity) For a function class $\mathcal{F}$ that consists of mappings from $\mathcal{Z}$ to $\mathbb{R}$, define

$$
\mathfrak{R F}:=\mathbb{E}_{z, v} \sup _{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} v_{i} f\left(z_{i}\right), \quad \mathfrak{R}_{n} \mathcal{F}:=\mathbb{E}_{v} \sup _{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} v_{i} f\left(z_{i}\right),
$$

as the Rademacher complexiy and the empirical Rademacher complexity of $\mathcal{F}$, respectively, where $\left\{v_{i}\right\}_{i=1}^{n}$ are i.i.d. Rademacher variables for which $\operatorname{Prob}\left(v_{i}=1\right)=\operatorname{Prob}\left(v_{i}=-1\right)=\frac{1}{2}$. In the above we explicitly denote expectation operators with subscripts that describes the distribution relative to which the expectation is computed. $\mathbb{E}_{z}$ means taking expectations over $\left\{z_{i}\right\}_{i=1}^{n}$ and $\mathbb{E}_{v}$ means taking expectations over $\left\{v_{i}\right\}_{i=1}^{n}$.

Furthermore, Dudley's integral bound provides one general solution to construct a computable upper bound on the local Rademacher complexity via the covering number of $\mathcal{F}$. We give the definition of covering number and state Dudley's integral bound for completeness as well.

Definition 4 (covering number) Assume $(\mathcal{M}, \operatorname{met}(\cdot, \cdot))$ is a metric space, and $\mathcal{F} \subseteq \mathcal{M}$. The $\varepsilon$-covering number of the set $\mathcal{F}$ with respect to a metric $\operatorname{met} r(\cdot, \cdot)$ is the size of its smallest $\varepsilon$-net cover:

$$
\mathcal{N}(\varepsilon, \mathcal{F}, \text { met } r)=\min \left\{m: \exists f_{1}, \ldots, f_{m} \in \mathcal{F} \text { such that } \mathcal{F} \subseteq \cup_{j=1}^{m} \mathcal{B}\left(f_{j}, \varepsilon\right)\right\}
$$

where $\mathcal{B}(f, \varepsilon):=\{\tilde{f}: \operatorname{metr}(\tilde{f}, f) \leq \varepsilon\}$.

We call $\log \mathcal{N}(\varepsilon, \mathcal{F}$, metr $)$ the metric entropy of the set $\mathcal{F}$ with respect to a metric metr$(\cdot, \cdot)$. Standard metrics include the $L_{p}\left(\mathbb{P}_{n}\right)$ metric defined by $L_{p}\left(\mathbb{P}_{n}\right)\left(f_{1}, f_{2}\right):=\sqrt[p]{\mathbb{P}_{n}\left(f_{1}(z)-f_{2}(z)\right)^{p}}$ for $p>0$. For function classes characterized by metric entropy conditions, the surrogate function constructed by Dudley's integral bound is often optimal. We use the $L_{2}\left(\mathbb{P}_{n}\right)$ metric for simplicity; the result trivially extends to $L_{p}\left(\mathbb{P}_{n}\right)$ metrics for $p \geq 2$, since $\log \mathcal{N}\left(\varepsilon, \mathcal{F}, L_{2}\left(\mathbb{P}_{n}\right)\right) \leq$ $\log \mathcal{N}\left(\varepsilon, \mathcal{F}, L_{p}\left(\mathbb{P}_{n}\right)\right)$ for any set $\mathcal{F}$ and discretization error $\varepsilon>0$.

Lemma 1 (Dudley's integral bound, [24]) Given $r>0$ and a class $\mathcal{F}$ that consists of functions defined on $\mathcal{Z}$,

$$
\mathfrak{R}_{n}\left\{f \in \mathcal{F}: \mathbb{P}_{n}\left[f^{2}\right] \leq r\right\} \leq \inf _{\varepsilon_{0}>0}\left\{4 \varepsilon_{0}+12 \int_{\varepsilon_{0}}^{\sqrt{r}} \sqrt{\frac{\log \mathcal{N}\left(\varepsilon, \mathcal{F}, L_{2}\left(\mathbb{P}_{n}\right)\right)}{n}} d \varepsilon\right\} .
$$

In what follows, when comparing different complexity parameters, we often use " $\vee$ " to denote the maximum operator, and to interpret correctly, its use should be understood to take precedence over addition but not over multiplication. Throughout we will find it convenient to use "big-oh" notation to simplify various expressions and comparisons that capture order of magnitude effects. For two non-negative sequence $\left\{a_{n}\right\}$ and $\left\{b_{n}\right\}$, we write $a_{n}=O\left(b_{n}\right)$, if $a_{n}$ can be upper bounded by $b_{n}$ up to an absolute constant for sufficiently large $n$. The same expression is often used also in
the context of probabilistic statements in which case it is interpreted as holding on an event which has a specified (typically "high") probability. We write $a_{n}=\Omega\left(b_{n}\right)$ if $b_{n} / a_{n}=O(1)$.

### 1.3.2 Loss-dependent rates via empirical risk minimization

In this section we are interested in loss-dependent rates, which should scale tightly with $\mathcal{L}^{*}$ := $\mathbb{P}\left[\ell\left(h^{*} ; z\right)-\inf _{\mathcal{H}} \ell(h ; z)\right]$; the best achievable "effective loss" on $\mathcal{H}$. The following theorem characterizes the loss-dependent rate of empirical risk minimization (ERM) via a surrogate function $\psi$, its fixed point $r^{*}$, the effective loss $\mathcal{L}^{*}$ and $B$.

Theorem 1 (loss-dependent rate of ERM) For the excess loss class $\mathcal{F}$ in (1.3.2), assume there is a meaningful surrogate function $\psi(r ; \delta)$ that satisfies $\forall \delta \in(0,1)$ and $\forall r>0$, with probability at least $1-\delta$,

$$
\sup _{f \in \mathcal{F}: \mathbb{P}\left[f^{2}\right] \leq r}\left(\mathbb{P}-\mathbb{P}_{n}\right) f \leq \psi(r ; \delta) .
$$

Then the empirical risk minimizer $\hat{h}_{\mathrm{ERM}} \in \arg \min _{\mathcal{H}}\left\{\mathbb{P}_{n} \ell(h ; z)\right\}$ satisfies for any fixed $\delta \in(0,1)$ and $r_{0} \in\left(0,4 B^{2}\right)$, with probability at least $1-\delta$,

$$
\mathcal{E}\left(\hat{h}_{\mathrm{ERM}}\right) \leq \psi\left(24 B \mathcal{L}^{*} ; \frac{\delta}{C_{r_{0}}}\right) \vee \frac{r^{*}}{6 B} \vee \frac{r_{0}}{48 B},
$$

where $C_{r_{0}}=2 \log _{2} \frac{8 B^{2}}{r_{0}}$, and $r^{*}$ is the fixed point of $6 B \psi\left(8 r ; \frac{\delta}{C_{r_{0}}}\right)$.

Remarks. 1) The term $r_{0}$ is negligible since it can be arbitrarily small. One can simply set $r_{0}=B^{2} / n^{4}$, which will much smaller than $r^{*}$ in general ( $r^{*}$ is at least of order $B^{2} \log \frac{1}{\delta} / n$ in the traditional "local Rademacher complexity" analysis, because this term is unavoidable in two-sided concentration inequalities).
2) In high-probability bounds, $C_{r_{0}}$ will only appear in the form $\log \left(C_{r_{0}} / \delta\right)$ ), which is of a negligible $O(\log \log n)$ order, so $C_{r_{0}}$ can be viewed an absolute constant for all practical purposes. As a result,
our generalization error bound can be viewed to be of the order

$$
\begin{equation*}
\mathcal{E}\left(\hat{h}_{\mathrm{ERM}}\right) \leq O\left(\psi\left(B \mathcal{L}^{*} ; \delta\right) \vee \frac{r^{*}}{B}\right) \tag{1.3.3}
\end{equation*}
$$

3) By using the empirical "effective loss," $\mathbb{P}_{n}\left[\ell\left(\hat{h}_{\mathrm{ERM}} ; z\right)-\inf _{\mathcal{H}} \ell(h ; z)\right]$, to estimate $\mathcal{L}^{*}$, the lossdependent rate can be estimated from data without knowledge of $\mathcal{L}^{*}$. We defer the details to Appendix A.2.3.

Comparison with existing results. Under additional restrictions (to be explained later), the traditional analysis (1.2.3) leads to a loss-dependent rate of the order [1]

$$
\begin{equation*}
\mathcal{E}\left(\hat{h}_{\mathrm{ERM}}\right) \leq O\left(\sqrt{\frac{\mathcal{L}^{*} r^{*}}{B}} \vee \frac{r^{*}}{B}\right), \tag{1.3.4}
\end{equation*}
$$

which is strictly worse than our result (1.3.3) due to reasoning following Statement 2. When $B \mathcal{L}^{*} \leq O\left(r^{*}\right)$, both (1.3.3) and (1.3.4) are dominated by the order $r^{*} / B$ so there is no difference between them. However, when $B \mathcal{L}^{*} \geq \Omega\left(r^{*}\right)$, our result (1.3.3) will be of order $\psi\left(B \mathcal{L}^{*} ; \delta\right)$ and the previous result (1.3.4) will be of order $\sqrt{\mathcal{L}^{*} r^{*} / B}$. In this case, the square-root function $\sqrt{\mathcal{L}^{*} r^{*} / B}$ is only a coarse relaxation of $\psi\left(B \mathcal{L}^{*} ; \delta\right)$ : as the traditional analysis requires $\psi$ to be sub-root, we can compare the two orders by

$$
\begin{equation*}
\psi\left(B \mathcal{L}^{*} ; \delta\right) \stackrel{\psi \text { is sub-root }}{\leq} \sqrt{\frac{B \mathcal{L}^{*}}{r^{*}}} \psi\left(r^{*} ; \delta\right) \stackrel{r^{*} \text { is fixed point }}{=} O\left(\sqrt{\frac{\mathcal{L}^{*} r^{*}}{B}}\right) . \tag{1.3.5}
\end{equation*}
$$

The "sub-root" inequality (the first inequality in (1.3.5)) becomes an equality when $\psi(r ; \delta)=$ $O(\sqrt{d r / n})$ in the parametric case, where $d$ is the parametric dimension. However, when $\mathcal{F}$ is "rich," $\psi(r ; \delta) / \sqrt{r}$ will be strictly decreasing so that the "sub-root" inequality can become quite loose. For example, when $\mathcal{F}$ is a non-parametric class we often have $\psi(r ; \delta)=O\left(\sqrt{r^{1-\rho} / n}\right)$ for some $\rho \in(0,1)$. The richer $\mathcal{F}$ is (e.g., the larger $\rho$ is), the more loose the "sub-root" inequality. This intuition will be validated via examples in Section 1.4.

Theorem 1 also applies to broader settings than previous results. For example, in [1] it is assumed that the loss is non-negative, and their original result only adapts to $\mathbb{P} \ell\left(h^{*} ; z\right)$ rather than the "effective loss" $\mathcal{L}^{*}$. Our proof (see Appendix A.2.2) is quite different as we bypass the Bernstein condition (which is traditionally implied by non-negativity, but not satisfied by the class used here), bypass the sub-root assumption on $\psi$, and adapt to the "better" parameter $\mathcal{L}^{*}$.

### 1.3.3 Variance-dependent rates via moment penalization

The loss-dependent rate proved in Theorem 1 contains a complexity parameter $B \mathcal{L}^{*}$ within its $\psi$ function, which may still be much larger than the optimal variance $\mathcal{V}^{*}$. Despite its prevalent use in practice, standard empirical risk minimization is unable to achieve variance-dependent rates in general. An example is given in [4] where $\mathcal{V}^{*}=0$ and the optimal rate is at most $O(\log n / n)$, while $\mathcal{E}\left(\hat{h}_{\text {ERM }}\right)$ is proved to be slower than $n^{-\frac{1}{2}}$.

We follow the path of penalizing empirical second moments (or variance) $[6,7,4,5]$ to design an estimator that achieves the "right" bias-variance trade-off for general, potentially "rich," classes. Our proposed estimator simultaneously achieves correct scaling on $\mathcal{V}^{*}$, along with minimaxoptimal sample dependence ( $n$ ). Besides empirical first and second moments, it only depends on the boundedness parameter $B$, a computable surrogate function $\psi$, and the confidence parameter $\delta$. All of these quantities are essentially assumed known in previous works: e.g., [6, 7] require covering number of the loss class, which implies a computable surrogate $\psi$ via Dudley's integral bound; and estimators in [4, 5] rely on the fixed point $r^{*}$ of a computable surrogate $\psi$.

In order to adapt to $\mathcal{V}^{*}$, we use a sample-splitting two-stage estimation procedure (this idea is inspired by the prior work [5]). Without loss of generality, we assume access to a data set of size $2 n$. We split the data set into the "primary" data set $S$ and the "auxiliary" data set $S^{\prime}$, both of which are of size $n$. We denote $\mathbb{P}_{n}$ the empirical distribution of the "primary" data set, and $\mathbb{P}_{S^{\prime}}$ the empirical distribution of the "auxiliary" data set.

Strategy 1 (the two-stage sample-splitting estimation procedure.) At the first-stage, we derive a preliminary estimate of $\mathcal{L}_{0}^{*}:=\mathbb{P} \ell\left(h^{*} ; z\right)$ via the "auxiliary" data set $S^{\prime}$, which we refer to as $\widehat{\mathcal{L}_{0}^{*}}$.

Then, at the second stage, we perform regularized empirical risk minimization on the "primal" data set $S$, which penalizes the centered second moment $\mathbb{P}_{n}\left[\left(\ell(h ; z)-\widehat{\mathcal{L}_{0}^{*}}\right)^{2}\right]$.

As we will present later, it is rather trivial to obtain a qualified preliminary estimate $\widehat{\mathcal{L}_{0}^{*}}$ via empirical risk minimization. Therefore, we firstly introduce the second-stage moment-penalized estimator, which is more crucial and interesting.

Strategy 2 (the second-stage moment-penalized estimator.) Consider the excess loss class $\mathcal{F}$ in (1.3.2). Let $\psi(r ; \delta)$ be a meaningful surrogate function that satisfies $\forall \delta \in(0,1), \forall r>0$, with probability at least $1-\delta$,

$$
4 \mathfrak{R}_{n}\left\{f \in \mathcal{F}: \mathbb{P}_{n}\left[f^{2}\right] \leq 2 r\right\}+\sqrt{\frac{2 r \log \frac{8}{\delta}}{n}}+\frac{9 B \log \frac{8}{\delta}}{n} \leq \psi(r ; \delta) .
$$

Denote $C_{n}=2 \log _{2} n+5$. Given a fixed $\delta \in(0,1)$, let the estimator $\hat{h}_{\mathrm{MP}}$ be

$$
\hat{h}_{\mathrm{MP}} \in \arg \min _{\mathcal{H}}\left\{\mathbb{P}_{n} \ell(h ; z)+\psi\left(16 \mathbb{P}_{n}\left[\left(\ell(h ; z)-\widehat{\mathcal{L}_{0}^{*}}\right)^{2}\right] ; \frac{\delta}{C_{n}}\right)\right\} .
$$

Given an arbitrary preliminary estimate $\widehat{\mathcal{L}_{0}^{*}} \in[-B, B]$, we can prove that the generalization error of the moment-penalized estimator $\hat{h}_{\mathrm{MP}}$ is at most

$$
\begin{equation*}
\mathcal{E}\left(\hat{h}_{\mathrm{MP}}\right) \leq 2 \psi\left(c_{0}\left[\mathcal{V}^{*} \vee\left(\widehat{\mathcal{L}_{0}^{*}}-\mathcal{L}_{0}^{*}\right)^{2} \vee r^{*}\right] ; \frac{\delta}{C_{n}}\right), \tag{1.3.6}
\end{equation*}
$$

with probability at least $1-\delta$, where $c_{0}$ is an absolute constant, and $r^{*}$ is the fixed point of $16 B \psi\left(r ; \frac{\delta}{C_{n}}\right)$. Moreover, the first-stage estimation error will be negligible if

$$
\begin{equation*}
\left(\widehat{\mathcal{L}_{0}^{*}}-\mathcal{L}_{0}^{*}\right)^{2} \leq O\left(r^{*}\right) \tag{1.3.7}
\end{equation*}
$$

It is rather elementary to show that performing the standard empirical risk minimization on $S^{\prime}$ suffices to satisfy (1.3.7), provided an additional assumption that $\psi$ is a "sub-root" function. We now give our theorem on the generalization error following this two-stage procedure.

Theorem 2 (variance-dependent rate) Let $\widehat{\mathcal{L}_{0}^{*}}=\inf _{\mathcal{H}} \mathbb{P}_{S^{\prime}} \ell(h ; z)$ be attained via empirical risk minimization on the auxiliary data set $S^{\prime}$. Assume that the meaningful surrogate function $\psi(r ; \delta)$ is "sub-root," i.e. $\frac{\psi(r ; \delta)}{\sqrt{r}}$ is non-increasing over $r \in\left[0,4 B^{2}\right]$ for all fixed $\delta$. Then for any $\delta \in\left(0, \frac{1}{2}\right)$, by performing the moment-penalized estimator in Strategy 2, with probability at least $1-2 \delta$,

$$
\mathcal{E}\left(\hat{h}_{\mathrm{MP}}\right) \leq 2 \psi\left(c_{1} \mathcal{V}^{*} ; \frac{\delta}{C_{n}}\right) \vee \frac{c_{1} r^{*}}{8 B}
$$

where $r^{*}$ is the fixed point of $B \psi\left(r ; \frac{\delta}{C_{n}}\right)$ and $c_{1}$ is an absolute constant.

Remarks. 1) In high-probability bounds, $C_{n}$ will only appear in the form $\left.\log \left(C_{n} / \delta\right)\right)$, which is of a negligible $O(\log \log n)$ order, so $C_{n}$ can effectively be viewed as constant for all practical purposes.
2) The "sub-root" assumption in Theorem 2 is only used to to bound the first-stage estimation error (see (1.3.7)). This assumption is not needed for the result (1.3.6) on the second-stage momentpenalized estimator.
3) Replacing $\mathcal{V}^{*}$ by an empirical centered second moment, we can prove a fully data-dependent generalization error bound that is computable from data without the knowledge of $\mathcal{V}^{*}$. We leave the full discussion to Appendix A.2.5.

Comparison with existing results. The best variance-dependent rate attained by existing estimators is of the order [5]

$$
\sqrt{\frac{\mathcal{V}^{*} r^{*}}{B^{2}}} \vee \frac{r^{*}}{B}
$$

which is strictly worse than the rate proved in Theorem 2. The reasoning is similar to Statement 2 and the explanation after Theorem 1: when $\mathcal{V}^{*} \leq O\left(r^{*}\right)$ the two results are essentially identical, but our estimator can perform much better when $\mathcal{V}^{*} \geq \Omega\left(r^{*}\right)$. Because $\psi$ is sub-root and $r^{*}$ is the
fixed point, we can compare the orders of the rates

$$
\begin{aligned}
\psi\left(\mathcal{V}^{*} ; \delta\right) & \leq \sqrt{\frac{\mathcal{V}^{*}}{r^{*}}} \psi\left(r^{*} ; \delta\right) \quad[\psi(r ; \delta) / \sqrt{r} \text { non-increasing for sub-root } \psi] \\
& =O\left(\sqrt{\frac{\mathcal{V}^{*} r^{*}}{B^{2}}}\right) \quad\left[r^{*} \text { is the fixed point of } B \psi(r ; \delta)\right]
\end{aligned}
$$

Since variance-dependent rates are generally used in applications that require robustness or exhibit large worst-case boundedness parameter, $\mathcal{V}^{*} \geq r^{*}$ is the more critical regime where one wants to ensure the estimation performance will not degrade.

Discussion. Per our "uniform localized convergence" principle, the most obvious difficulty in proving Theorem 2 is in establishing (1.3.6): the empirical second moment is sample-dependent, whereas standard tools in empirical process theory like Talagrand's concentration inequality (Lemma 16) require the localized subsets to be independent of the samples. The core techniques in our proof essentially overcome this difficulty by concentrating data-dependent localized subsets to data-independent ones. This idea may be of independent interest; we defer details to Appendix A.2.4.

The tightness of our variance-dependent rates depend on tightness of the computable surrogate function $\psi$. When covering numbers of the excess loss class are given, a direct choice is Dudley's integral bound (see Lemma 1), which is known to be rate-optimal for many important classes.

Previous approaches usually take a simper regularization term [6,5] that is proportional to the square root of the empirical second moment (or empirical variance). That type of penalization is "too aggressive" for rich classes, as Strategy 2 finds the right regularization scheme from the theoretical perspective. One important future direction is to investigate practical choices of $\psi$ (e.g. convex ones) in Strategy 2 that can be optimized efficiently. [4] proposes a regularization approach that preserves convexity of empirical risk. However, based on an equivalence proved in their paper, they have similar limitations to the approaches that penalizes the square root of the empirical variance. Outside the study of variance-dependent rates, integral-based and local-

Rademacher-complexity-based penalization is also used in model selection [25], but the setting and the goal of model selection are very different from the problem we study here.

### 1.4 Illustrative examples in the "slow rate" regime

### 1.4.1 Discussion and illustrative examples

Recall that our loss-dependent rates and variance-dependent (moment-penalized) rates are

$$
\begin{align*}
\mathcal{E}\left(\hat{h}_{\mathrm{ERM}}\right) \leq O\left(\psi\left(B \mathcal{L}^{*} ; \delta\right) \vee \frac{r^{*}}{B}\right) & {[\text { Theorem 1] }}  \tag{1.4.1}\\
\mathcal{E}\left(\hat{h}_{\mathrm{MP}}\right) \leq O\left(\psi\left(\mathcal{V}^{*} ; \delta\right) \vee \frac{r^{*}}{B}\right) & {[\text { Theorem 2] }} \tag{1.4.2}
\end{align*}
$$

respectively. In contrast to our results (1.4.1) (1.4.2), the best known loss-dependent rates [1] and variance-dependent rates [5] are

$$
\begin{array}{cc}
\mathcal{E}\left(\hat{h}_{\mathrm{ERM}}\right) \leq O\left(\sqrt{\frac{\mathcal{L}^{*} r^{*}}{B}} \vee \frac{r^{*}}{B}\right) & \text { [existing result [1]], } \\
\mathcal{E}\left(\hat{h}_{\text {previous }}\right) \leq O\left(\sqrt{\frac{\mathcal{V}^{*} r^{*}}{B^{2}}} \vee \frac{r^{*}}{B}\right) & \text { [existing result [5]], } \tag{1.4.4}
\end{array}
$$

respectively (we use $\hat{h}_{\text {previous }}$ to denote the previous best known moment-penalized estimator proposed in [5]). To illustrate the noticeable gaps between our new results and previous known ones, we compare the two different variance-dependent rates, (1.4.2) and (1.4.4) on two important families of "rich" classes: non-parametric classes of polynomial growth and VC classes. The implications of this comparison will similarly apply to loss-dependent rates.

Before presenting the advantages of the new problem-dependent rates, we would like to discuss how to compute them. In Theorem 1 and Theorem 2, the class of concentrated functions, $\mathcal{F}$, is the excess loss class $\ell \circ \mathcal{H}-\ell \circ h^{*}$ in (1.3.2). As we have mentioned in earlier sections, a general solution for the $\psi$ function is to use Dudley's integral bound (Lemma 1). Knowledge of the metric
entropy of the excess loss class,

$$
\log \mathcal{N}\left(\varepsilon, \ell \circ \mathcal{H}-\ell \circ h^{*}, L_{2}\left(\mathbb{P}_{n}\right)\right)
$$

can be used to calculate Dudley's integral bound and construct the surrogate function $\psi$ needed in our theorems. Note that there is no difference between the metric entropy of the excess loss class and metric entropy of the loss class itself: from the definition of covering number, one has

$$
\mathcal{N}\left(\varepsilon, \ell \circ \mathcal{H}-\ell \circ h^{*}, L_{2}\left(\mathbb{P}_{n}\right)\right)=\mathcal{N}\left(\varepsilon, \ell \circ \mathcal{H}, L_{2}\left(\mathbb{P}_{n}\right)\right)
$$

We comment that almost all existing theoretical literature that discusses general function classes and losses $[1,6,7,5]$ imposes metric entropy conditions on the loss class/excess loss class rather than the hypothesis class $\mathcal{H}$, and we follows that line as well to allow for a seamless comparison of the results. As a complement, we will discuss how to obtain such metric entropy conditions for practical applications in Section 1.4.2.

## Non-parametric classes of polynomial growth

Example 1 (non-parametric classes of polynomial growth) Consider a loss class $\ell \circ \mathcal{H}$ with the metric entropy condition

$$
\begin{equation*}
\log \mathcal{N}\left(\varepsilon, \ell \circ \mathcal{H}, L_{2}\left(\mathbb{P}_{n}\right)\right) \leq O\left(\varepsilon^{-2 \rho}\right) \tag{1.4.5}
\end{equation*}
$$

where $\rho \in(0,1)$ is a constant. Using Dudley's integral bound to find $\psi$ and solving $r \leq O(B \psi(r ; \delta))$, it is not hard to verify that

$$
\psi(r ; \delta) \leq O\left(\sqrt{\frac{r^{1-\rho}}{n}}\right), \quad r^{*} \leq O\left(\frac{B^{\frac{2}{1+\rho}}}{n^{\frac{1}{1+\rho}}}\right) .
$$

As a result, our variance-dependent rate (1.4.2) is of the order

$$
\begin{equation*}
\mathcal{E}\left(\hat{h}_{\mathrm{MP}}\right) \leq O\left(\mathcal{V}^{* \frac{1-\rho}{2}} n^{-\frac{1}{2}} \vee \frac{r^{*}}{B}\right) \tag{1.4.6}
\end{equation*}
$$

which is $O\left(\mathcal{V}^{* \frac{1-\rho}{2}} n^{-\frac{1}{2}}\right)$ when $\mathcal{V}^{*} \geq r^{*}$. In contrast, the previous best-known result (1.4.4) is of the order

$$
\begin{equation*}
\mathcal{E}\left(\hat{h}_{\text {previous }}\right) \leq O\left(\sqrt{\mathcal{V}^{*}} B^{-\frac{\rho}{1+\rho}} n^{-\frac{1}{2+2 \rho}} \vee \frac{r^{*}}{B}\right), \tag{1.4.7}
\end{equation*}
$$

which is $O\left(\sqrt{\mathcal{V}^{*}} B^{-\frac{\rho}{1+\rho}} n^{-\frac{1}{2+2 \rho}}\right)$ when $\mathcal{V}^{*} \geq r^{*}$. Therefore, for arbitrary choice of $n, \mathcal{V}^{*}, B$, the "sub-optimality gap" is

$$
\begin{equation*}
\text { ratio between (1.4.7) and }(1.4 .6):=\frac{\sqrt{\mathcal{V}^{*}} B^{-\frac{\rho}{1+\rho}} n^{-\frac{1}{2+2 \rho}} \vee \frac{r^{*}}{B}}{\mathcal{V}^{*} \frac{1-\rho}{2}} n^{-\frac{1}{2}} \vee \frac{r^{*}}{B} \quad 1 \vee\left(\mathcal{V}^{*}\left(\frac{n}{B^{2}}\right)^{\frac{1}{1+\rho}}\right)^{\frac{\rho}{2}}, \tag{1.4.8}
\end{equation*}
$$

which can be arbitrary large and grows polynomially with $n$.
We consider two stylized regimes where our variance-dependent rate is much better (we use the notation $\approx$ when the left hand side and the right hand side are of the same order).

- The more "traditional" regime: $B \approx 1, \mathcal{V}^{*} \approx n^{-a}$ where $a>0$ is a fixed constant. This regime captures the traditional supervised learning problems where $B$ is not large, but one wants to use the relatively small order of $\mathcal{V}^{*}$ to achieve "faster" rates.
- The "high-risk" regime: $B \approx n^{b}$ where $b>0$ is a fixed constant, and $\mathcal{V}^{*} \ll B^{2}$ (i.e., $\mathcal{V}^{*}$ is much smaller than order $n^{2 b}$ ). This regime captures modern "high-risk" learning problems such as counterfactual risk minimization, policy learning, and supervised learning with limited number of samples. In those settings, the worst-case boundedness parameter is considered to scale with $n$ so that one wants to avoid (or reduce) the dependence on $B$.

In both of the two regimes, generalization errors via naive (non-localized) uniform convergence arguments will be worse than our approach by orders polynomial in $n$, so we only need to compare
with previous variance-dependent rates.

The "traditional" regime. The "sub-optimality gap" $(1.4 .8)$ is $1 \vee\left(\mathcal{V}^{*} n^{\frac{1}{1+\rho}}\right)^{\frac{\rho}{2}}$. It is quite clear that when $\mathcal{V}^{*} \approx n^{-a}$ where $0<a<\frac{1}{1+\rho}$, our variance-dependent rate improves over all previous generalization error rates by orders polynomial in $n$.

The "high-risk" regime. We focus on the simple case $B^{\frac{2}{1+\rho}} \leq \mathcal{V}^{*} \ll 4 B^{2}$ to gain some insight, where our result exhibits an improvement of order $O\left(n^{\frac{\rho}{2}\left(\frac{1}{1+\rho}\right)}\right)$ relative to the previous result. Clearly the larger $\rho$, the more improvement we provide. By letting $\rho \rightarrow 1$ our improvement can be as large as $O\left(n^{\frac{1}{4}}\right)$.

## VC-type classes

Our next example considers VC-type classes. Although this classical example has been extensively studied in learning theory, our results provide strict improvements over antecedents.

Example 2 (VC-type classes) One general definition of VC-type classes (which is not necessarily binary) uses the metric entropy condition. Consider a loss class $\ell \circ \mathcal{H}$ that satisfies

$$
\log \mathcal{N}\left(\varepsilon, \ell \circ \mathcal{H}, L_{2}\left(\mathbb{P}_{n}\right)\right) \leq O\left(d \log \frac{1}{\varepsilon}\right)
$$

where $d$ is th so-called the Vapnik-Chervonenkis (VC) dimension [26]. Using Dudley's integral bound to find the surrogate $\psi$ and solving $r \leq O(B \psi(r ; \delta))$, it can be proven [2] that

$$
\psi(r ; \delta) \leq O\left(\sqrt{\frac{d r}{n} \log \frac{8 B^{2}}{r}} \vee \frac{B d}{n} \log \frac{8 B^{2}}{r}\right), \quad r^{*} \leq O\left(\frac{B^{2} d \log n}{n}\right) .
$$

Recently, [5] proposed a moment-penalized estimator whose generalization error is of the rate

$$
\mathcal{E}\left(\hat{h}_{\text {previous }}\right) \leq O\left(\sqrt{\frac{d \mathcal{V}^{*} \log n}{n}}+\frac{B d \log n}{n}\right),
$$

in the worst case without invoking other assumptions. This result has a $O(\log n)$ gap compared with the $\Omega\left(\sqrt{\frac{d \mathcal{V}^{*}}{n}}\right)$ lower bound [27], which holds for arbitrary sample size. There is much recent interest focused on the question of when the sub-optimal $\log n$ factor can be removed $[28,5]$.

By applying Theorem 2, our refined moment-penalized estimator gives a generalization error bound of tighter rate

$$
\begin{equation*}
\mathcal{E}\left(\hat{h}_{\mathrm{MP}}\right) \leq O\left(\sqrt{\frac{d \mathcal{V}^{*} \log \frac{8 B^{2}}{\mathcal{V}^{*}}}{n}} \vee \frac{B d \log n}{n}\right) \tag{1.4.9}
\end{equation*}
$$

This closes the $O(\log n)$ gap in the regime $\mathcal{V}^{*} \geq \Omega\left(\frac{B^{2}}{(\log n)^{\alpha}}\right)$, where $\alpha>0$ is an arbitrary positive constant. Though this is not the central regime, it is the first positive result that closes the notorious $O(\log n)$ gap without invoking any additional assumptions on the loss/hypothesis class (e.g., the rather complex "capacity function" assumption introduced in [5]). We anticipate additional improvements are possible under further assumptions on the hypothesis class and the loss function.

### 1.4.2 Problem areas to which "localization" theory is applicable

In practical applications it is more standard to consider metric entropy conditions of the hypothesis class $\mathcal{H}$. In view of this, we introduce two important settings where metric entropy on the loss/excess loss class can be obtained from metric entropy conditions on $\mathcal{H}$. Thus, the improvements illustrated in Section 1.4.1 can be directly transferred to these application areas.

Supervised learning with Lipchitz continuous cost. In supervised learning, the data $z$ is a feature-label pair $(x, y)$, and the $\operatorname{loss} \ell(h ; z)$ is of the form

$$
\ell(h ; z)=\ell_{\mathrm{SV}}(h(x), y)
$$

where $\ell_{\mathrm{sv}}: \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ is a fixed cost function that is $L_{\mathrm{SV}}$-Lipchitz continuous with respect to its first argument,

$$
\ell_{\mathrm{sv}}\left(u_{1}, y\right)-\ell_{\mathrm{sv}}\left(u_{2}, y\right) \leq L_{\mathrm{sv}}\left|u_{1}-u_{2}\right|, \quad \forall u_{1}, u_{2}, y \in \mathbb{R} .
$$

For hypothesis classes characterized by metric entropy conditions, properties are preserved because

$$
\log \mathcal{N}\left(\varepsilon, \ell \circ \mathcal{H}, L_{2}\left(\mathbb{P}_{n}\right)\right) \leq \log \mathcal{N}\left(\frac{\varepsilon}{L_{\mathrm{sv}}}, \mathcal{H}, L_{2}\left(\mathbb{P}_{n}\right)\right)
$$

Note that $L_{\mathrm{SV}}$ only depends on the cost function and is usually of constant order. Our theory naturally applies to supervised learning problems where the cost function is Lipchitz continuous (for example, the $\ell_{1}$ cost, the hinge cost, the ramp cost, etc.). Moreover, it is simple to show that any surrogate function of the localized empirical process $\sup _{\mathbb{P}\left[\left(h(x)-h^{*}(x)\right)^{2}\right] \leq r}\left(\mathbb{P}-\mathbb{P}_{n}\right)\left(\ell_{\mathrm{SV}}(h(x), y)\right)$ will lead to a construction of the surrogate function $\psi$ required in Theorem 1 and Theorem 2.

Counterfactual risk minimization. Denote $x \in \mathcal{X}$ the feature and $t \in \mathcal{T}$ the treatment (e.g. $\mathcal{T}=\{0,1\}$ in binary treatment experimental design), and $c(x, t)$ the unknown cost function. A hypothesis (policy) $h$ is a map from $\mathcal{X} \times \mathcal{T}$ to [0,1] such that $\sum_{t \in \mathcal{T}} h(x, t)=1$. Thus, a hypothesis (policy) essentially maps features to a distribution over treatments. We consider the standard formulation of "learning with logged bandit feedback," dubbed "counterfactual risk minimization" [7]: a batch of samples $\left\{\left(x_{i}, t_{i}, c_{i}\right)\right\}_{i=1}^{n}$ are obtained by applying a known policy $h_{0}$, so that $t_{i}$ is sampled from $h_{0}\left(x_{i}, \cdot\right)$ and one can only observe the cost $c_{i}$ associated with $t_{i}$. We write $z=(x, t, c)$ and let

$$
\begin{equation*}
\ell\left(h ; z_{i}\right)=\frac{c_{i}}{h_{0}\left(x_{i}, t_{i}\right)} h\left(x_{i}, t_{i}\right) \tag{1.4.10}
\end{equation*}
$$

be the "constructed loss" using importance sampling. It is straightforward to show that the population risk $\mathbb{P} \ell(h ; z)$ is equal to the expected cost of policy $h$, so determining good policies requires one to minimize the generalization error $\mathcal{E}(\hat{h})$. It is usually convenient to obtain metric entropy condition of the loss/excess loss class by using the linearity structure of (1.4.10). In particular, from the Cauchy-Schwartz inequality we can prove that

$$
\begin{equation*}
\log \mathcal{N}\left(\varepsilon, \ell \circ \mathcal{H}, L_{2}\left(\mathbb{P}_{n}\right)\right) \leq \log \mathcal{N}\left(\frac{\varepsilon}{\gamma_{n}}, \mathcal{H}, L_{4}\left(\mathbb{P}_{n}\right)\right) \tag{1.4.11}
\end{equation*}
$$

where $\gamma_{n}:=\sqrt[4]{\mathbb{P}_{n}\left[\left(\frac{c(x, t)}{h_{0}(x, t)}\right)^{4}\right]}$ only depends on the functions $c, h_{0}$ in the given problem, and the samples rather than the worst-case parameters. A systematical challenge in counterfactual risk minimization is that the worst-case boundedness parameter, $\sup _{h, z}|\ell(h ; z)|$, is typically very large, since the inverse probability term $\frac{1}{h_{0}\left(x_{i}, t_{i}\right)}$ in (1.4.10) is typically large in the worst case.

### 1.5 Problem-dependent rates in the parametric "fast rate" regime

### 1.5.1 Background

When assuming suitable curvature or margin conditions, the direct dependence of the generalization error on $n$ is typically faster than $O\left(n^{-\frac{1}{2}}\right)$. We call this the "fast rate" regime. A well-known example is, when the hypothesis class is parametrized and the loss is strongly convex with respect to the parameter, in which case the direct dependence of the generalization error on $n$ is typically the "parametric rate" $O\left(n^{-1}\right)$. In the "fast rate" regime, existing problem-dependent rates are mostly studied in supervised learning with structured convex cost; see Section 1.2.4 for a historical review of existing localization approaches and problem-dependent rates. Through our proposed "uniform localized convergence" procedure, we can recover results in $[8,9]$ for supervised learning problems with structured convex cost, and our approach is able to systematically weaken some restrictions. (We defer the discussion to Appendix A. 1 as the contributions there mostly lie in unification and some technical improvements.)

In this and the next two sections we study more modern applications in the fast rate regime,
focusing on computationally efficient estimators and algorithms, where the derivation of sharp problem-dependent rates remains an open question. A secondary objective is to provide unification to vector-based uniform convergence analysis. The main focal points are the following.

Non-convexity, stationary points, and iterative algorithms. Classical generalization error analysis relies on the property of global empirical risk minimizers. However, many important machine learning problems are non-convex. For those problems, guarantees on global empirical risk minimizes are not sufficient, and therefore one typically targets guarantees on stationary points and the iterates produced by optimization algorithms. This motivates us to study uniform convergence of gradients and sample-based iterative algorithms. Existing generalization error bounds in this area are typically not localized, and connections to traditional localization frameworks is not fully understood.

General formulation of stochastic optimization. Existing problem-dependent rates mostly focus on supervised learning settings, with specialized assumptions on the problem structure. Hence, it is important to characterize problem-dependent generalization error bounds for more general stochastic optimization problems, in which the classical asymptotic results do not depend on the parametric dimension $d$ and global parameters. Existing methods, however, typically give rise to dimension-dependent factors and "large" global parameters.

Organization of this section is as follows. In Section 1.5.2, we will strengthen the "uniform convergence of gradients" idea by developing a theory for "uniform localized convergence of gradients." In Section 1.5.3 we will provide problem-dependent rates for approximate stationary points of empirical risk and iterates of the gradient descent algorithm.

### 1.5.2 Theoretical foundations

Recently, the idea of "uniform convergence of gradients" [19, 20, 29] has been applied successfully to many non-convex learning and stochastic optimization problems. These works do not
consider problem-dependent rates, and their results typically rely on various global parameters, like global Lipchitz constants and the radius of the parametric set. In this subsection we strengthen these ideas by developing a theory of "uniform localized convergence of gradients." This theory will be proven to be more powerful in deriving problem-dependent rates. Before moving to state a key assumption, we introduce some additional notation.

Notation for the parametric "fast rate" regime. We write the loss function as $\ell(\theta ; z)$ where $\theta \in \mathbb{R}^{d}$ is the parameter representation of the hypothesis $h$. Consider a compact set $\Theta \subseteq \mathbb{R}^{d}$ and let $\theta^{*}$ be the best parameter within $\Theta$, which satisfies $\theta^{*} \in \arg \min _{\Theta} \mathbb{P} \ell(\theta ; z)$. Denote $\|\cdot\|$ to be the $L_{2}$ norm in $\mathbb{R}^{d}$, noting that most of our results can be generalized to matrix learning problems by considering the Frobenius norm. We let $\mathcal{B}^{d}\left(\theta_{0}, \rho\right):=\left\{\theta \in \mathbb{R}^{d}:\left\|\theta-\theta_{0}\right\| \leq \rho\right\}$ denote a ball with center $\theta_{0} \in \mathbb{R}^{d}$ and radius $\rho$. We assume that there are two radii $\Delta_{m}, \Delta_{M}$ such that $\mathcal{B}^{d}\left(\theta^{*}, \Delta_{m}\right) \subseteq \Theta \subseteq \mathcal{B}^{d}\left(\theta^{*}, \Delta_{M}\right)$. We would like to provide guarantee for the generalization error

$$
\mathcal{E}(\hat{\theta}):=\mathbb{P} \ell(\hat{\theta} ; z)-\mathbb{P} \ell\left(\theta^{*} ; z\right) .
$$

We state a key assumption of our framework.

Assumption 1 (statistical noise of smooth population risk) For all $\theta_{1}, \theta_{2} \in \Theta, \frac{\nabla \ell\left(\theta_{1} ; z\right)-\nabla \ell\left(\theta_{2} ; z\right)}{\left\|\theta_{1}-\theta_{2}\right\|}$ is a $\beta$-sub-exponential random vector. Formally there exist $\beta>0$ such that for any unit vector $u \in \mathcal{B}(\mathbf{0}, 1)$ and $\theta_{1}, \theta_{2} \in \Theta$,

$$
\mathbb{E}\left\{\exp \left(\frac{\left|u^{T}\left(\nabla \ell\left(\theta_{1} ; z\right)-\nabla \ell\left(\theta_{2} ; z\right)\right)\right|}{\beta\left\|\theta_{1}-\theta_{2}\right\|}\right)\right\} \leq 2 .
$$

From Jensen's inequality and convexity of the exponential function, a simple consequence of Assumption 1 is that the population risk is $\beta$-smooth: for any $\theta_{1}, \theta_{2} \in \Theta$,

$$
\begin{equation*}
\left\|\mathbb{P} \nabla \ell\left(\theta_{1} ; z\right)-\mathbb{P} \nabla \ell\left(\theta_{2} ; z\right)\right\| \leq \beta\left\|\theta_{1}-\theta_{2}\right\| . \tag{1.5.1}
\end{equation*}
$$

Smoothness of the population risk is a standard assumption in the optimization literature. Compared with the smoothness condition (1.5.1), Assumption 1 is a stronger distributional assumption: it requires the random vectors $\left\{\frac{\nabla \ell\left(\theta_{1} ; z\right)-\nabla \ell\left(\theta_{2} ; z\right)}{\left\|\theta_{1}-\theta_{2}\right\|}\right\}$ to be a $\beta$-sub-exponential for all $\theta_{1}, \theta_{2} \in \Theta$, while the smoothness condition (1.5.1) only concerns the expectation of these random vectors. Certain distributional assumptions are necessary to analyze the generalization performance of unbounded losses (e.g., see the "Hessian statistical noise" assumption in [19], and the "moment-equivalence" assumptions in $[30,31]$ ). Assumption 1 imposed here is applicable to many smooth machine learning models studied in previous literature, and its verification is often no much harder than verification of smoothness conditions. This can be explained through the following lemma and discussion.

Lemma 2 (Hessian statistical noise implies Assumption 1) Assumption 1 is satisfied if for all $\theta \in \Theta$, the random Hessian matrix $\nabla^{2} \ell(\theta ; z)$ is a $\beta$-sub-exponential matrix. Formally, Assumption 1 is satisfied when there exist $\beta>0$ such that for any unit vectors $u_{1}, u_{2} \in \mathcal{B}^{d}(\mathbf{0}, 1)$ and any $\theta \in \Theta$,

$$
\begin{equation*}
\mathbb{E}\left\{\exp \left(\frac{1}{\beta}\left|u_{1}^{T} \nabla^{2} \ell(\theta ; z) u_{2}\right|\right)\right\} \leq 2 \tag{1.5.2}
\end{equation*}
$$

By Lemma 2, one only needs to compute the Hessian matrices and verify they are sub-exponential over $\Theta$. For instance, many statistical estimation problems have $\nabla^{2} \ell(\theta ; z)$ proportional to $z z^{T}$ (or $x x^{T}$ when the problem is a supervised learning problem and $z=(x, y)$ is the feature-label pair). By assuming the data $z$ (or $x$ ) is a $\tau$-sub-Gaussian vector, $z z^{T}$ (or $x x^{T}$ ) becomes a $\tau^{2}$-sub-exponential matrix. If the remaining quantities in $\nabla^{2} \ell(\theta ; z)$ can be uniformly bounded by some constant $C_{0}$, then Assumption 1 holds with $C_{0} \tau^{2}$. Note that similar steps are needed to verify the smoothness conditions.

We carefully choose a function class $\mathcal{G}=\left\{g_{(\theta, v)}: \theta \in \Theta, v \in \mathcal{B}^{d}\left(\mathbf{0}, \max \left\{\Delta_{M}, \frac{1}{n}\right\}\right)\right\}$ to apply concentration, where each element is a function $g_{(\theta, v)}: \mathcal{Z} \rightarrow \mathbb{R}$ defined by

$$
\begin{equation*}
g_{(\theta, v)}(z)=\left(\nabla \ell(\theta ; z)-\nabla \ell\left(\theta^{*} ; z\right)\right)^{T} v . \tag{1.5.3}
\end{equation*}
$$

By applying Proposition 1 with the "concentrated functions" $g_{(\theta, v)}$ and the "measurement functional" $T$ defined by $T((\theta, v))=\left\|\theta-\theta^{*}\right\|^{2}+\|v\|^{2}$, we obtain the key argument of "uniform localized convergence" for gradients.

Proposition 2 (uniform localized convergence of gradients) Under Assumption $1, \forall \delta \in(0,1)$, with probability at least $1-\delta$, for all $\theta \in \Theta$,

$$
\begin{align*}
& \left\|\left(\mathbb{P}-\mathbb{P}_{n}\right)\left(\nabla \ell(\theta ; z)-\nabla \ell\left(\theta^{*} ; z\right)\right)\right\| \\
& \leq c \beta \max \left\{\left\|\theta-\theta^{*}\right\|, \frac{1}{n}\right\}\left(\sqrt{\frac{d+\log \frac{4 \log _{2}\left(2 n \Delta_{M}+2\right)}{\delta}}{n}}+\frac{d+\log \frac{4 \log _{2}\left(2 n \Delta_{M}+2\right)}{\delta}}{n}\right), \tag{1.5.4}
\end{align*}
$$

where $c$ is an absolute constant.

Distinct from previous results on "uniform convergence of gradients," which give the same upper bound on $\left\|\left(\mathbb{P}-\mathbb{P}_{n}\right) \nabla \ell(\theta ; z)\right\|$ for all $\theta \in \Theta$, the right hand side of (1.5.4) in Proposition 2 scales linearly with $\left\|\theta-\theta^{*}\right\|$ for all $\theta$ such that $\left\|\theta-\theta^{*}\right\| \geq \frac{1}{n}$. Therefore, Proposition 2 provides a refined, "localized" upper bound on the concentration of gradients. This property makes Proposition 2 the key in deriving problem-dependent rates.

### 1.5.3 Main results

In order to obtain tight problem-dependent rates, we require a very mild assumption on the noise at the optimal point $\theta^{*}$. We also ask $\theta^{*}$ to be the unique population risk minimizer in $\Theta$ in this subsection.

Assumption 2 (noise at the optimal point) There is a unique population risk minimizer $\theta^{*}=$ $\arg \min _{\Theta} \mathbb{P} \ell(\theta ; z)$, and the gradient at $\theta^{*}$ satisfies the Bernstein condition: there exists $G_{*}>0$ such that for all $2 \leq k \leq n$,

$$
\begin{equation*}
\mathbb{E}\left[\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{k}\right] \leq \frac{1}{2} k!\mathbb{E}\left[\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{2}\right] G_{*}^{k-2} . \tag{1.5.5}
\end{equation*}
$$

We note that this assumption is very mild because $G_{*}$ only depends on gradients at $\theta^{*}$, and $G_{*}$ will only appear in the $O\left(n^{-2}\right)$ terms in our theorems. Our approach also allows many other noise conditions at $\theta^{*}$ (e.g., those for heavy-tailed noise). At a high level, the order of our generalization error bounds only depends on the concentration of $\mathbb{P}_{n} \nabla \ell\left(\theta^{*} ; z\right)$ relative to $\mathbb{P} \nabla \ell\left(\theta^{*} ; z\right)$, which barely depends on noise at the single point $\theta^{*}$ and can be analyzed under various types of noise conditions. We introduce Assumption 2 here because it leads to the asymptotically optimal problem-dependent parameter $\mathbb{P}\left[\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{2}\right]$ and simplifies comparison with previous literature.

Now we turn to establish problem-dependent rates under a curvature condition. While our methodology is widely-applicable without restriction to particular curvature conditions, we will focus on the Polyak-Łojasiewicz (PL) condition, which is known to be one of the weakest conditions that guarantee linear convergence of optimization algorithms [32] as well as fast-rate generalization error [20].

Assumption 3 (Polyak-Łojasiewicz condition) There exist $\mu>0$ such that for all $\theta \in \boldsymbol{\Theta}$,

$$
\mathbb{P} \ell(\theta ; z)-\mathbb{P} \ell\left(\theta^{*} ; z\right) \leq \frac{\|\mathbb{P} \nabla \ell(\theta ; z)\|^{2}}{2 \mu}
$$

The PL condition is weaker than many others recently introduced in the areas of matrix recovery, deep learning, and learning dynamical systems, such as "one-point convexity" [33, 34], "star convexity" [35], and "quasar convexity" [36, 37], not to mention the classical "strong convexity." It is also referred to as "gradient dominance condition" in previous literature [20]. Under suitable assumptions on their inputs, many popular non-convex models have been shown to satisfy the PL condition (sometimes locally rather than globally). An incomplete list of these models includes: matrix factorization [38], neural networks with one hidden layer [33], ResNets with linear activations [39], binary linear classification [19], robust regression [19], phase retrieval [40], blind deconvolution [41], linear dynamical systems [37], mixture of two Gaussians [21], to name a few.

While the PL condition is known to be one of the weakest conditions that can be used to establish linear convergence to the global minimum (see [32] for its relationship with other common
curvature conditions), the generalization aspects of such structural non-convex learning problems have not been fully understood. In particular, existing generalization error bounds often contain global Lipchitz parameters that can be large for unbounded smooth losses.

Our next theorem provides problem-dependent bounds for approximate stationary points of the empirical risk, under the PL condition of the population risk. The theorem implies that optimization procedures that find stationary points of the empirical risk are also learning algorithms for the population risk.

Theorem 3 (generalization error of the approximate stationary point) Under Assumptions 1, 2 and $3, \forall \delta \in(0,1)$, with probability at least $1-\delta$, we have the following results:
(a) there exist approximate stationary points of the empirical risk, $\widehat{\theta} \in \Theta$ such that

$$
\begin{equation*}
\left\|\mathbb{P}_{n} \nabla \ell(\theta ; z)\right\| \leq \sqrt{\frac{2 \mathbb{P}\left[\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{2}\right] \log \frac{4}{\delta}}{n}}+\frac{G_{*} \log \frac{4}{\delta}}{n} . \tag{1.5.6}
\end{equation*}
$$

(b) for all $\hat{\theta}$ that satisfy (1.5.6), when $n \geq \frac{c \beta^{2}}{\mu^{2}}\left(d+\log \frac{4 \log \left(2 n \Delta_{M}+1\right)}{\delta}\right)$, where $c$ is an absolute constant, we have

$$
\mathcal{E}(\hat{\theta}) \leq \frac{64 \mathbb{P}\left[\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{2}\right] \log \frac{4}{\delta}}{\mu n}+\frac{32 G_{*}^{2} \log ^{2} \frac{4}{\delta}+4 \mu^{2}}{\mu n^{2}} .
$$

Ignoring higher order terms and absolute constants, Theorem 3 implies a problem-dependent bound

$$
\begin{equation*}
\mathcal{E}(\hat{\theta}) \leq O\left(\frac{\mathbb{P}\left[\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{2}\right] \log \frac{1}{\delta}}{\mu n}\right), \tag{1.5.7}
\end{equation*}
$$

which scales tightly with the problem-dependent parameter $\mathbb{P}\left[\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{2}\right]$. The proof of Theorem 4 can be found in Appendix A.3.4. Optimality and implications of this problem-dependent rate will be discussed shortly after we present an additional theorem.

Many optimization algorithms, including gradient descent [32], stochastic gradient descent [42], and non-convex SVRG [43] can efficiently find (approximate) stationary points of the em-
pirical risk. However, convergence of these optimization algorithms is mostly studied under assumptions on the empirical risk. The next theorem demonstrates that under assumptions on the population risk, gradient descent provably achieve "small" generalization error. These type of results are challenging to prove because properties of the population risk may not transfer to the empirical risk.

Consider the gradient descent algorithm with fixed step size $\alpha$ and initialization $\theta^{0}$, generating iterates according to

$$
\begin{equation*}
\theta^{t+1}=\theta^{t}-\alpha \mathbb{P}_{n} \nabla \ell\left(\theta^{t} ; z\right), \quad t=0,1, \ldots \tag{1.5.8}
\end{equation*}
$$

Theorem 4 (generalization error of gradient descent) Assume Assumptions 1, 2, 3. Then for an initialization $\theta^{0} \in \mathcal{B}^{d}\left(\theta^{*}, \sqrt{\frac{\mu}{\beta}} \Delta_{m}\right)$ and step size $\frac{1}{\beta}$, the iterates of the gradient descent algorithm (1.5.8) satisfy for any fixed $\delta \in(0,1)$, with probability at least $1-\delta$, for all $t=0,1, \ldots$,

$$
\begin{equation*}
\mathcal{E}\left(\theta^{t}\right) \leq \underbrace{\frac{16 \mathbb{P}\left[\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{2}\right] \log \frac{4}{\delta}}{\mu n}+\frac{8 G_{*}^{2} \log ^{2} \frac{4}{\delta}+\mu^{2}}{\mu n^{2}}}_{\text {statistical error }}+\left(1-\frac{\mu}{2 \beta}\right)^{t} \mathcal{E}\left(\theta^{0}\right) \tag{1.5.9}
\end{equation*}
$$

provided that the sample size $n$ is large enough such that that the "statistical error" term in (1.5.9) is smaller than $\frac{\mu}{2} \Delta_{m}^{2}$ and $n \geq \frac{c \beta^{2}}{\mu^{2}}\left(d+\log \frac{8 \log _{2}\left(2 n \Delta_{M}+2\right)}{\delta}\right)$, where $c$ is an absolute constant.

Theorem 4 is the first broad scope result on the generalization error of gradient descent under the PL condition. It implies that after a logarithmic number of iterations, gradient descent achieves the problem-dependent rate (1.5.7). Note that the algorithm only requires the initialization condition in the theorem, rather than any knowledge of $\Theta$ and the problem-dependent parameters. The proof of Theorem 4 can be found in Appendix A.3.4, and the main idea is applicable to many other optimization algorithms as well. For example, in Section 1.7 we provide a similar analysis to the first-order Expectation-Maximization algorithm.

Optimality of the problem-dependent rates in Theorem 3 and Theorem 4. It is well-known in the asymptotic statistics literature [26] that when $n$ tends to infinity, under mild local conditions,

$$
\begin{equation*}
\sqrt{n}\left(\hat{\theta}_{\mathrm{ERM}}-\theta^{*}\right) \xrightarrow{d} N\left(\mathbf{0}, H^{-1} Q H^{-1}\right), \tag{1.5.10}
\end{equation*}
$$

where $H=\mathbb{P} \nabla^{2} \ell\left(\theta^{*} ; z\right), Q=\mathbb{P}\left[\nabla \ell\left(\theta^{*} ; z\right) \nabla \ell\left(\theta^{*} ; z\right)^{T}\right]$, and $\xrightarrow{d}$ means convergence in distribution. The asymptotic rate (1.5.10) is often information theoretically optimal under suitable conditions [44] (e.g., it matches the Hájek-Le Cam asymptotic minimax lower bound [45, 46] when the loss $\ell(\theta ; z)$ is a $\log$ likelihood function). The generalization error bounds in Theorem 3 and Theorem 4 , which are of the order

$$
\begin{equation*}
\mathcal{E}(\hat{\theta}) \leq O\left(\frac{\mathbb{P}\left[\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{2}\right] \log \frac{1}{\delta}}{\mu n}\right), \tag{1.5.11}
\end{equation*}
$$

are natural finite-sample versions of the "ideal" asymptotic benchmark (1.5.10). This is because the generalization error $\mathcal{E}(\hat{\theta})$ can be approximated by the quadratic form $\left(\hat{\theta}-\theta^{*}\right)^{T} H\left(\hat{\theta}-\theta^{*}\right), 1 / \mu$ is a natural proxy for the inverse Hessian $H^{-1}$, and $\mathbb{P}\left[\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{2}\right.$ is a natural proxy for $Q$.

In both Theorem 3 and Theorem 4, the sample complexity required to make the generalization error smaller than a fixed $\varepsilon>0$ is

$$
\begin{equation*}
n \geq \Omega\left(\frac{\beta^{2} d}{\mu^{2}}+\frac{\mathbb{P}\left[\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{2}\right] \log \frac{1}{\delta}}{\mu \varepsilon}\right) \tag{1.5.12}
\end{equation*}
$$

Here we only consider the "interesting" case of $\varepsilon \in\left(0, \frac{\mu \Delta_{m}^{2}}{2}\right)$ in Theorem 4; otherwise the initialization point $\theta^{0}$ will satisfy $\mathcal{E}\left(\theta^{0}\right) \leq \varepsilon$. Clearly, the sample complexity threshold $n \geq \Omega\left(\beta^{2} d / \mu^{2}\right)$ scales with the dimension $d$. This threshold is sharp up to absolute constants-there exist simple linear regression constructions where we require $\Omega\left(\beta^{2} d / \mu^{2}\right)$ samples [47] to make the empirical Hessian positive definite. As a result, our overall sample complexity (1.5.12) is essentially the sharpest result one can expect under the aforementioned assumptions.

### 1.6 Applications to non-convex learning and stochastic optimization

In this section we will compare our problem-dependent rates with previous results from two areas: non-convex learning and stochastic optimization. Another topic that nicely illustrates the advantages of our approach, Expectation-Maximization algorithms for missing data problems, is deferred to Section 1.7.

### 1.6.1 Non-convex learning under curvature conditions

In this subsection we discuss generalization error bounds for non-convex losses that satisfy the Polyak-Łojasiewicz (PL) condition. The PL condition is one of the weakest curvature conditions that have been rigorously and extensively studied in the areas of matrix recovery, deep learning, learning dynamical systems and learning mixture models. See our prior discussion under Assumption 3 and the reference thereof for representative models that satisfy this condition, and its relationship with other curvature conditions. The topic has attracted much recent attention because there is some empirical evidence suggesting that modern deep neural networks might satisfy this condition in large neighborhoods of global minimizers [34, 35].

For structured non-convex learning problems, a benchmark approach to prove generalization error bounds is "uniform convergence of gradients." [19] presents the "uniform convergence of gradients" principle and proves dimension-dependent generalization error bounds to several representative problems; [20] extends this idea to obtain norm-based generalization error bounds. We will compare our problem-dependent rates with these results.

Comparison with the results in Mei et al. [19]. The main regularity assumptions imposed in [19] include: 1) an assumption on the statistical noise of the Hessian matrices, whose content is similar to our Assumption $1 ;$ and 2$)$ an assumption that the random gradients $\nabla \ell(\theta ; z)$ are $G$-subGaussian for all $\theta \in \Theta$, which is not used in our framework (in contrast, we only impose a mild assumption on the gradient noise at $\theta^{*}$ ). They also assume the Hessian is Lipchitz continuous which we view as redundant.

The theoretical foundation in [19] is the following result on the (global) uniform convergence of gradients: with probability at least $1-\delta$,

$$
\begin{equation*}
\sup _{\theta \in \Theta}\left\|\left(\mathbb{P}-\mathbb{P}_{n}\right) \nabla \ell(\theta ; z)\right\| \leq O\left(G \sqrt{\frac{d \log n \log \frac{1}{\delta}}{n}}\right) . \tag{1.6.1}
\end{equation*}
$$

The sub-Gaussian parameter $G$ is larger than the global Lipchitz constant, and can be quite large in practical applications. From (1.6.1), when the population risk satisfies the PL condition, the generalization error for a stationary point $\hat{\theta}$ of the empirical risk can be bounded as follows:

$$
\begin{equation*}
\mathcal{E}(\hat{\theta}) \leq O\left(\frac{G^{2} d \log n \log \frac{1}{\delta}}{\mu n}\right) . \tag{1.6.2}
\end{equation*}
$$

[19] also provides guarantees for iterates of the gradient descent algorithm, but the analysis is specialized to the three applications considered in the paper. It is worth mentioning that [19] also studies the high-dimensional setting and provides a series of important results; we will not compare with those.

Our approach improves both the result (1.6.2) as well as the methodology (1.6.1) as follows.

- Our Theorem 3 and Theorem 4 provide generalization error bounds for approximate stationary points and iterates of the gradient descent algorithm, which are of the order

$$
\mathcal{E}(\hat{\theta}) \leq O\left(\frac{\mathbb{P}\left[\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{2}\right] \log \frac{1}{\delta}}{\mu n}\right)
$$

Focusing on the the dominating $O\left(n^{-1}\right)$ term, our new result replaces $O\left(G^{2} d \log n\right)$ in the numerator with the typically much smaller localized quantity $\mathbb{P}\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{2}$. In fact, from the definition of sub-Gaussian vectors, one can prove (see, e.g. [48]) that

$$
\mathbb{P}\left[\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{2}\right] \ll \sup _{\Theta} \mathbb{P}\|\nabla \ell(\theta ; z)\|^{2} \leq 8 G^{2} d,
$$

so our bounds are always more favorable than (1.6.2). In passing, we also remove a super-
fluous $\log n$ factor by using generic chaining rather than simple discretization.

- Our Proposition 2 on the localized uniform convergence of gradients,

$$
\left\|\left(\mathbb{P}-\mathbb{P}_{n}\right)\left(\nabla \ell(\theta ; z)-\nabla \ell\left(\theta^{*} ; z\right)\right)\right\| \leq O\left(\beta\left\|\theta-\theta^{*}\right\| \sqrt{\frac{d+\log \frac{2 \log \left(2 n \Delta_{M}\right)}{\delta}}{n}}\right),
$$

strengthens (1.6.1) to a localized result, under fewer assumptions.

We illustrate our improvements on a particular non-convex learning example.
Example 3 (non-convex regression with non-linear activation) Consider the model

$$
\begin{equation*}
\ell(\theta ; z)=\left(\eta\left(\theta^{T} x\right)-y\right)^{2} \tag{1.6.3}
\end{equation*}
$$

where $\eta(\cdot)$ is a non-linear activation function, and there exists $\theta^{*} \in \Theta$ such that $\mathbb{E}[y]=\eta\left(x^{T} \theta^{*}\right)$. This model has been empirically shown to be superior relative to convex formulations in several applications [49, 50, 51], and is representative of the class of one-hidden-layer neural network models. Popular choices of $\eta$ include sigmoid link $\eta(t)=\left(1+e^{-t}\right)^{-1}$ and probit link $\eta(t)=\Phi(t)$ where $\Phi$ is the Gaussian cumulative distribution function. Under mild regularity conditions, the population risk $\mathbb{P} \ell(\theta ; z)$ satisfies the $P L$ condition and the statistical noise conditions.

Assumption 4 (regularity conditions for Example 3) (a) $\|x\|_{\infty}$ is uniformly bounded by $\tau$, the feasible parameter set $\Theta$ is given by $\left\{\theta \in \mathbb{R}^{d}:\|\theta\| \leq \frac{\Delta_{M}}{2}\right\}$, and $B$ is the worst-case boundedness parameter of $\left(\eta\left(\theta^{T} x\right)-y\right)^{2}$ (which can scale with $n$ ). (b) there exist $C_{\eta}>0, c_{\eta}>0$ such that

$$
\sup _{|t| \leq \Delta_{M} \tau \sqrt{d}} \max \left\{\eta^{\prime}(t), \eta^{\prime \prime}(t)\right\} \leq C_{\eta}, \quad \inf _{|t| \leq \Delta_{M} \tau \sqrt{d}} \eta^{\prime}(t) \geq c_{\eta} .
$$

(c) The feature vector $x$ spans all directions in $\mathbb{R}^{d}$, that is, $\mathbb{E}\left[x x^{T}\right] \geq \gamma \tau^{2} I_{d \times d}$ for some $0<\gamma<1$.

Under Assumption 4, all of our proposed assumptions in Theorem 3 and Theorem 4 are satisfied. In particular: Assumption 1 holds with $\beta=2 C_{\eta}\left(C_{\eta}+\sqrt{B}\right) \tau^{2}$; Assumption 2 holds with
$G_{*}=2 C_{\eta} \tau \sqrt{B d}$; and Assumption 3 holds with $\mu=\frac{2 c_{\eta}^{3} \tau^{2} \gamma}{C_{\eta}}$ (see Appendix A.3.5 for details). Let $\hat{\theta}$ be the approximate stationary point in Theorem 3, or the output of the gradient descent algorithm in Theorem 4 (after running sufficiently many iterations), we have the following corollary.

Corollary 5 (generalization error bound for Example 3) Under Assumption 4,

$$
\begin{equation*}
\mathcal{E}(\hat{\theta}) \leq O\left(\frac{d \mathcal{L}^{*}\left(\tau C_{\eta}\right)^{2} \log \frac{1}{\delta}}{\mu n}\right) \tag{1.6.4}
\end{equation*}
$$

where $\mathcal{L}^{*}:=\mathbb{P}\left[\left(y-\eta\left(\theta^{* T} x\right)\right)^{2}\right]$ is the population risk at the optimal parameter. $\theta^{*}$.
Since the sub-Gaussian parameter of the random gradient satisfies $G \leq O\left(\tau C_{\eta} \sqrt{B}\right)$ under Assumption 4, the result (1.6.2) in Mei et al. [19] implies a generalization error bound of the order

$$
\begin{equation*}
\mathcal{E}(\hat{\theta}) \leq O\left(\frac{d B\left(\tau C_{\eta}\right)^{2} \log n \log \frac{1}{\delta}}{\mu n}\right) \quad \text { [existing result [19]], } \tag{1.6.5}
\end{equation*}
$$

where $B=\sup _{\theta, x, y}\left(\eta\left(\theta^{T} x\right)-y\right)^{2}$ is the worst-case boundedness parameter. Let us now compare our result (1.6.4) with the the existing result (1.6.5): 1) our result (1.6.4) improves the worst-case boundedness parameter $B$, replacing it with the much smaller optimal risk $\mathcal{L}^{*}$; and 2 ) it removes the superfluous logarithmic factor $\log n$.

Comparison with the norm-based generalization error bound in Foster et al. [20]. Let us now compare our problem-dependent rates with the norm-based bounds in [20] (it is worth mentioning that they also provide novel results in the infinite-dimensional and high-dimensional settings). Under the formulation in Example 3 and Assumption 4, the generalization error bound proved in [20] is of the order

$$
\begin{equation*}
\mathcal{E}(\hat{\theta}) \leq O\left(\frac{d^{2} B\left(\tau C_{\eta}\right)^{4}+d B\left(\tau C_{\eta}\right)^{2} \log \frac{1}{\delta}}{\mu n}\right) \quad[\text { converted from [20]], } \tag{1.6.6}
\end{equation*}
$$

for an approximate stationary point $\hat{\theta}$. To be specific, their original result assumes $\|x\|$ to be uniformly bounded by $D$ and the generalization error scales with $D^{4}$. Under the standard assumption
$\|x\|_{\infty} \leq \tau$ (or $x$ being a $\tau$-sub-Gaussian random vector), $D^{4}$ is of order $\tau^{4} d^{2}$ so their result does not achieve optimal dependence on $d$ (in the original statements in [20] there is a potential misunderstanding of the dependence on $d$ ). Besides improving the worst-case boundedness parameter $B$ to the optimal risk $\mathcal{L}^{*}$, our result (1.6.4) further improves (1.6.6) by order $d\left(\tau C_{\eta}\right)^{2}$.

Lastly, we comment that there is no formal guarantee on how to find $\hat{\theta}$ by an optimization algorithm in [20]. They merely establish the generalization error bound for approximate stationary points, but analysis of an optimization algorithm is more challenging because properties of the population risk may not carry over to the empirical risk.

### 1.6.2 Stochastic optimization

The parametric learning setting we discussed is sometimes referred to as "stochastic optimization" [52, 53]. Beyond supervised learning, stochastic optimization also covers operations research and system control problems, where the dimension $d$ may no longer be pertinent in the generalization error bound for sufficiently large $n$ (i.e., the bound should be dimension-independent in the asymptotic regime). Therefore, it is preferable to prove norm-based generalization error bounds, which do not explicitly scale with $d$ after a certain sample complexity threshold.

We compare our results with previous ones in the area of stochastic optimization. Those results typically assume the population risk to be strongly convex, i.e., there exists $\mu>0$ such that $\forall \theta_{1}, \theta_{2} \in \Theta$,

$$
\mathbb{P} \ell\left(\theta_{1} ; z\right)-\mathbb{P} \ell\left(\theta_{2} ; z\right)-\left(\mathbb{P} \nabla \ell\left(\theta_{2} ; z\right)\right)^{T}\left(\theta_{1}-\theta_{2}\right) \geq \frac{\mu}{2}\left\|\theta_{1}-\theta_{2}\right\|^{2}
$$

While this assumption is much more restrictive than our Assumption 3, we note that our problemdependent rate and sample complexity results are novel even in this well-studied strongly convex setting.

Recall that our problem-dependent generalization error bounds in Theorem 3 and Theorem 4
are of the order

$$
\begin{equation*}
\mathcal{E}(\hat{\theta}) \leq O\left(\frac{\mathbb{P}\left[\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{2}\right] \log \frac{1}{\delta}}{\mu n}\right), \tag{1.6.7}
\end{equation*}
$$

provided $n \geq \Omega\left(\beta^{2} d / \mu^{2}\right)$; and the sample complexity (to achieve $\varepsilon$ generalization error) is

$$
\begin{equation*}
n \geq \Omega\left(\frac{\beta^{2} d}{\mu^{2}}+\frac{\mathbb{P}\left[\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{2}\right] \log \frac{1}{\delta}}{\mu \varepsilon}\right) \tag{1.6.8}
\end{equation*}
$$

Our results are natural finite-sample extensions of the classical "asymptotic normality" result (1.5.10), and hence are the sharpest results one can expect under aforementioned assumptions (see the discussion at the end of Section 1.5).

Comparison with the classical result from Shapiro et al. [54]. Perhaps the most representative result on the generalization error of empirical risk minimization (also referred to as "sample average approximation") in the stochastic optimization literature, is Corollary 5.20 from the monograph [54]. When the random gradient $\nabla \ell(\theta ; z)$ is $G$-sub-Gaussian for all $\theta \in \Theta$, that result implies

$$
\begin{equation*}
\mathcal{E}\left(\hat{\theta}_{\mathrm{ERM}}\right) \leq O\left(\frac{G^{2} d \log n \log \frac{1}{\delta}}{\mu n}\right) . \tag{1.6.9}
\end{equation*}
$$

One advantage of (1.6.9) is that it does not require the population risk to be smooth. However, the explicit dependence on $d$ and the global sub-Gaussian parameter $G$ in (1.6.9) make it less favorable for some operations research applications and M-estimation problems, where the asymptotic complexity (1.5.10) does not depend on $d$. It is easy to show that our problem-dependent generalization error bound (1.6.7) strictly improves on this classical result. Specifically, under the sub-Gaussian distributional assumptions on gradients, one can prove that

$$
\mathbb{P}\left[\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{2}\right] \ll \sup _{\theta \in \Theta} \mathbb{P}\left[\|\nabla \ell(\theta ; z)\|^{2}\right] \leq 8 d G^{2}
$$

Plugging this into (1.6.7) and (1.6.9), we observe that our bound improves on (1.6.9) by removing dependence on the worst-case $L_{2}$ norm of the gradient over the entire parameter space $\Theta$.

Comparison with results obtained from the "online to batch conversion" [55]. By assuming the population risk is strongly convex and satisfies the following "uniform Lipchitz continuous" condition,

$$
\left|\ell\left(\theta_{1} ; z\right)-\ell\left(\theta_{2} ; z\right)\right| \leq L_{\text {unif }}\left\|\theta_{1}-\theta_{2}\right\|, \quad \forall z \in \mathcal{Z}, \quad \forall \theta_{1}, \theta_{2} \in \Theta,
$$

[55] proves an "online to batch conversion" that relates the regret of an algorithm (on past data) to the generalization performance (on future data). As a result, the output $\hat{\theta}_{S G D}$ of certain stochastic gradient methods (also referred to as "stochastic approximation" in the stochastic optimization literature) can be proved to satisfy

$$
\begin{equation*}
\mathcal{E}\left(\hat{\theta}_{\mathrm{SGD}}\right) \leq O\left(\frac{L_{\mathrm{unif}}^{2} \log \frac{1}{\delta}}{n}\right) \tag{1.6.10}
\end{equation*}
$$

Note that (1.6.10) does not require any sample size threshold. In contrast, our problem-dependent generalization error bound (1.6.7) provides an improved rate, but only as long as the sample size condition $n \geq \Omega\left(\beta^{2} d / \mu^{2}\right)$ is satisfied, because in this case

$$
\mathbb{P}\left[\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{2}\right] \ll \sup _{\theta \in \Theta, z \in \mathcal{Z}}\|\nabla \ell(\theta ; z)\|^{2}=L_{\text {unif }}^{2} .
$$

Plugging this into (1.6.7) and (1.6.10), this claimed improvement can be immediately verified.

Comparison with loss-dependent bounds for ERM in [56]. By imposing both strong convexity and a uniform smoothness condition, [56] systematically provides dimension-independent generalization error bounds for empirical risk minimization. However, there are several limitations in their approach: 1) their sample complexity threshold to achieve dimension-independent generalization error is sub-optimal for many popular problems; and 2) many of their assumptions are restrictive
and not required (as our analysis shows).
The main source of the limitations in [56] is the assumption that $\ell(\theta ; z)$ admits a uniform smooth parameter $\beta_{\text {unif }}$, i.e.,

$$
\begin{equation*}
\left\|\nabla \ell\left(\theta_{1}, z\right)-\nabla \ell\left(\theta_{2}, z\right)\right\| \leq \beta_{\text {unif }}\left\|\theta_{1}-\theta_{2}\right\|, \quad \forall z \in \mathcal{Z}, \quad \forall \theta_{1}, \theta_{2} \in \Theta . \tag{1.6.11}
\end{equation*}
$$

This quantity serves as the main complexity proxy. With additional assumptions that $\ell(\theta, z)$ is non-negative and convex for all $z$, Theorem 3 in [56] proves that when

$$
n \geq \Omega\left(\frac{\beta_{\text {unif }} d \log n}{\mu}\right),
$$

empirical risk minimization achieves the problem-dependent bound

$$
\begin{equation*}
\mathcal{E}\left(\hat{\theta}_{\mathrm{ERM}}\right) \leq O\left(\frac{\beta_{\mathrm{unif}} \mathcal{L}^{*} \log \frac{1}{\delta}}{\mu n}\right) \tag{1.6.12}
\end{equation*}
$$

However, as $\beta_{\text {unif }}$ is effectively the largest value of the operator norm of the Hessian— $\sup _{\theta ; z}\left\|\nabla^{2} \ell(\theta ; z)\right\|_{\text {op }}$, it scales with dimension $d$ in most statistical estimation problems. As a result, the sample complexity threshold $\Omega\left(\beta_{\text {unif }} d \log n / \mu\right)$ becomes sub-optimal for most statistical regression problems. For example, consider a simple linear regression set up:

$$
\begin{equation*}
\ell(\theta ;(x, y))=\left(y-\theta^{T} x\right)^{2}, \quad y=x^{T} \theta^{*}+v, \quad v \sim N(0,1), \quad x \sim N\left(0, \tau^{2} I_{d \times d}\right) . \tag{1.6.13}
\end{equation*}
$$

In this example, we have $\mu=1$ and $\beta_{\text {unif }}=\Omega\left(\tau^{2} d\right)$, so the sample complexity needed to achieve $\varepsilon$ accuracy in [56] is

$$
\begin{equation*}
n \geq \Omega\left(\tau^{2} d^{2} \log n+\frac{d \mathcal{L}^{*} \tau^{2}}{\varepsilon}\right) \quad[\text { sample complexity [56]] } \tag{1.6.14}
\end{equation*}
$$

In contrast, our sample complexity (1.6.8) is

$$
n \geq \Omega\left(\tau^{2} d+\frac{d \mathcal{L}^{*} \tau^{2}}{\varepsilon}\right) \quad[\text { sample complexity (1.6.8)] }
$$

in this example. Therefore, the $\Omega\left(\tau^{2} d^{2} \log n\right)$ term in (1.6.14) is sub-optimal primarily because the analysis in [56] relies on the uniform smoothness parameter $\beta_{\text {unif }}$. Moreover, their assumptions that $\ell(\theta ; z)$ is non-negative and convex for all $z$ may rule out interesting stochastic optimization applications. These are not required by our framework and results.

Comparison to loss-dependent bounds for stochastic gradient methods in [57, 58, 59] The approach in [56] has been extended to analyze a variant of the stochastic gradient descent algorithm [57], but the results hold only in expectation rather than with high probability, and they have similar limitations to the results in [56]. This approach has also been extended to non-convex stochastic optimization problems [58], where the generalization error bounds are of the from (1.6.10)—they contain the uniform Lipchitz parameter $L_{\text {unif }}$ and are not problem-dependent.

Recently, [59] provided generalization error bounds for various optimization algorithms including SGD and non-convex SVRG, under the assumptions that the loss functions satisfy the PL condition, admit a uniform smooth parameter $\beta_{\text {unif }}$ as in (1.6.11), and are non-negative. The generalization error bounds in [59] (see its Theorem 1) include a statistical error term of the order (1.6.12) and an optimization error term. As we have argued under (1.6.12) through the linear regression example, the statistical error term in their bounds may be sub-optimal for many problems including linear regression, because its dependence on the uniform smooth parameter $\beta_{\text {unif }}$.

### 1.7 Learning with missing data and Expectation-Maximization algorithms

In this section we introduce a broad class of applicable non-convex and semi-supervised learning problems, in the area of "learning with missing data." We again apply our proposed "uniform localized convergence" framework and prove a variant of Theorem 4, which gives the sharpest local convergence rate for first-order Expectation-Maximization (EM) algorithms in many widely
studied problems. Our analysis improves the framework introduced recently in Balakrishnan et al. [21].

### 1.7.1 Background

Convex maximum likelihood estimation problems will generally become non-convex when there is missing or unobserved data. Assume the data $(z, w)$ is generated from an unknown distribution specified by the true parameter $\theta^{*} \in \mathbb{R}^{d}$, where $z \in \mathcal{Z}$ corresponds to the observable data, and $w \in \mathcal{W}$ corresponds to the unobservable data (also referred to as the "latent variable"). For every $\theta \in \mathbb{R}^{d}$, let $f_{\theta}(z, w)$ be the likelihood of observing $z$ conditioned on $w$, if the underlying distribution is specified by $\theta$. (Throughout this section we will assume the existence of density functions for simplicity.) Consider the loss function

$$
\begin{equation*}
\ell(\theta ; z)=-\log \left[\int_{\mathcal{W}} f_{\theta}(z, w) d w\right] . \tag{1.7.1}
\end{equation*}
$$

Our goal is to estimate the true parameter $\theta^{*}$, which minimizes the population risk

$$
\mathbb{P} \ell(\theta ; z)=\int_{\mathcal{Z}} \ell(\theta ; z) d z
$$

over all $\theta \in \mathbb{R}^{d}$. (Equivalently, $\theta^{*}$ maximizes the population log-likelihood function.) The main challenge is that $\mathbb{P} \ell(\theta ; z)$ is typically non-convex, despite the fact that the conditional loglikelihood function $\log f_{\theta}(z, w)$ would usually be convex with respect to $\theta$, if both $z$ and $w$ were observable.

The following $\ell_{\theta^{\prime}}(\theta ; z)$ function provides a convex upper bound on $\ell(\theta ; z)$, and can be interpreted as the conditional expectation of the loss, as if $\theta^{\prime}$ is the true parameter $\theta^{*}$ :

$$
\ell_{\theta^{\prime}}(\theta ; z)=-\int_{\mathcal{W}} k_{\theta^{\prime}}(w \mid z) \log f_{\theta}(z, w) d w,
$$

where $k_{\theta^{\prime}}(w \mid z)$ is the conditional density of $w$ given $z$. Denote $\nabla_{\theta} \ell_{\theta^{\prime}}(\theta ; z)$ the gradient of $\ell_{\theta^{\prime}}(\theta ; z)$
when fixing $\theta^{\prime}$. From the definition of $k_{\theta^{\prime}}(w \mid z)$, it is easy to verify that the vector-value of $\nabla_{\theta} \ell_{\theta^{\prime}}(\theta ; z)$ at $\theta^{\prime}$ is exactly the gradient of $\ell\left(\theta^{\prime} ; z\right)$ : that is, for all $\theta, \theta^{\prime} \in \Theta$,

$$
\begin{equation*}
\left.\nabla_{\theta} \ell_{\theta^{\prime}}(\theta ; z)\right|_{\theta=\theta^{\prime}}=\nabla \ell\left(\theta^{\prime} ; z\right) \tag{1.7.2}
\end{equation*}
$$

In view of the identity (1.7.2), it is known [21] that gradient descent on the empirical risk $\mathbb{P}_{n} \ell(\theta ; z)$ is equivalent to the first-order Expectation-Maximization algorithm: at the $t$-th iteration, the "expectation" step calculates the sample average $\mathbb{P}_{n} \ell_{\theta^{t}}(\theta ; z)$, and the "maximization" step executes the first-order update

$$
\begin{equation*}
\theta^{t+1}=\theta^{t}-\alpha \nabla \mathbb{P}_{n} \nabla \ell_{\theta^{t}}\left(\theta^{t} ; z\right)=\theta^{t}-\alpha \nabla \mathbb{P}_{n} \nabla \ell\left(\theta^{t} ; z\right), \tag{1.7.3}
\end{equation*}
$$

where $\alpha>0$ is the step size. First-order EM is known to be more computationally efficient than standard EM, and more amendable for analysis [21].

Examples of learning with missing data problems for which the above observations apply include the followings.

Example 4 (Mixture of two Gaussians) In this problem, the missing variable $w \in\{-1,1\}$ is an indicator of the underlying mixture component, which has $\frac{1}{2}$ probability to be 1 and the other $\frac{1}{2}$ probability to be -1 . Conditioned on $w$, the observable variable $z$ is generated as follows.

$$
(z \mid w=1) \sim N\left(\theta^{*}, \sigma^{2} I_{d \times d}\right), \quad(z \mid w=-1) \sim N\left(-\theta^{*}, \sigma^{2} I_{d \times d}\right) .
$$

For this problem, we have

$$
\ell_{\theta^{\prime}}(\theta ; z)=\frac{w_{\theta^{\prime}}(z)}{2}\|z-\theta\|^{2}+\frac{\left(1-w_{\theta^{\prime}}(z)\right)}{2}\|z+\theta\|^{2}
$$

where $w_{\theta^{\prime}}(z)=e^{-\frac{\left\|\theta^{\prime}-z\right\|^{2}}{2 \sigma^{2}}}\left[e^{-\frac{\left\|\theta^{\prime}-z\right\|^{2}}{2 \sigma^{2}}}+e^{-\frac{\left\|\theta^{\prime}+z\right\|^{2}}{2 \sigma^{2}}}\right]^{-1}$.

Example 5 (Mixture of two component linear regression) In this problem, $x \sim N\left(0, I_{d \times d}\right)$ is a
random feature vector, and $w \in\{-1,1\}$ is a missing indicator variable that has $\frac{1}{2}$ probability to be 1 and $\frac{1}{2}$ probability to be -1 . Conditioned on $w$ and $x$, the label variable $y$ is generated as follows.

$$
(y \mid w=1, x) \sim N\left(x^{T} \theta^{*}, \sigma^{2}\right), \quad(y \mid w=-1, x) \sim N\left(-x^{T} \theta^{*}, \sigma^{2}\right) .
$$

In this problem, the observable variable $z$ is the feature-label pair $(x, y)$, and we have

$$
\ell_{\theta^{\prime}}(\theta ; z)=\frac{w_{\theta^{\prime}}(x, y)}{2}\left(y-x^{T} \theta\right)^{2}+\frac{1-w_{\theta^{\prime}}(x, y)}{2}\left(y+x^{T} \theta\right)^{2},
$$

where $w_{\theta^{\prime}}(x, y)=e^{-\frac{\left(x^{T} \theta^{\prime}-y\right)^{2}}{2 \sigma^{2}}}\left[e^{-\frac{\left(x^{T} \theta^{\prime}-y\right)^{2}}{2 \sigma^{2}}}+e^{-\frac{\left(x^{T} \theta^{\prime}+y\right)^{2}}{2 \sigma^{2}}}\right]^{-1}$.

### 1.7.2 Problem-dependent rates for first-order EM

Motivated by the breakthrough work Balakrishnan et al. [21], we assume that the feasible parameter space $\Theta$ contains the true parameter $\theta^{*}$, and satisfies the two assumptions.

Assumption 5 (strong convexity of $\mathbb{P} \boldsymbol{C}_{\boldsymbol{\theta}^{*}}(\boldsymbol{\theta} ; \boldsymbol{z})$ ) There exists $\mu_{1}>0$ such that $\forall \theta_{1}, \theta_{2} \in \Theta$

$$
\mathbb{P} \ell_{\theta^{*}}(\theta ; z)-\mathbb{P} \ell_{\theta^{*}}\left(\theta^{*} ; z\right) \leq \frac{\left\|\mathbb{P} \nabla \ell_{\theta^{*}}(\theta ; z)\right\|^{2}}{2 \mu_{1}}
$$

Recall that $\mathbb{P} \ell_{\theta^{*}}(\theta ; z)$ is the underlying "true" log likelihood with respect to parameter $\theta$, which is unknown due to lack of information on $\theta^{*}$. It is standard to assume that $\mathbb{P} \ell_{\theta^{*}}(\theta ; z)$ is a strongly convex when there is no missing data [60,21].

Assumption 6 (gradient smoothness) There exists $0<\mu_{2}<\mu_{1}$ such that $\forall \theta \in \Theta$

$$
\left\|\mathbb{P} \nabla \ell_{\theta}(\theta ; z)-\mathbb{P} \nabla \ell_{\theta^{*}}(\theta ; z)\right\| \leq \mu_{2}\left\|\theta-\theta^{*}\right\|
$$

Assumption (6) is also assumed in [21]. While this assumption does not typically hold over all $\theta \in \mathbb{R}^{d}$, it is often satisfied with small enough $\mu_{2}$ over a local region around the true parameter $\theta^{*}$ [21, 61]. Under the above two assumptions, according to the identity (1.7.2), the gradient of the
population risk,

$$
\mathbb{P} \nabla \ell(\theta ; z)=\mathbb{P} \nabla \ell_{\theta}(\theta ; z),
$$

can be viewed as a perturbation of $\nabla \ell_{\theta^{*}}(\theta ; z)$ —the gradient of the strongly convex function $\mathbb{P} \ell_{\theta^{*}}(\theta ; z)$. Therefore, Assumption 5 and Assumption 6 play a similar role to that of the PL condition that we have analyzed in Section 1.5.3. The following theorem can be viewed as a modification of our previous Theorem 4, where the proof is tailored to these new assumptions but the key ideas remain mostly similar.

Theorem 6 (generalization error of first-order EM) Assume Assumptions 1, 2, 5, 6, and assume access to an initialization $\theta^{0} \in \mathcal{B}^{d}\left(\theta^{*}, \Delta_{m}\right)$. For any fixed $\delta \in(0,1)$, iterates of the first-order EM algorithm $\left\{\theta^{t}\right\}$ generated by (1.7.3) with the fixed step size $\frac{2}{\beta+\mu_{1}}$ satisfy with probability at least $1-\delta$ and all $t=0,1, \ldots$,

$$
\begin{align*}
\mathcal{E}\left(\theta^{t}\right) & \leq \frac{16 \beta}{\mu_{1}^{2}}\left(\sqrt{\frac{2 \mathbb{P}\left[\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{2}\right] \log \frac{4}{\delta}}{n}}+\frac{G_{*} \log \frac{4}{\delta}+\mu_{1}}{n}\right)^{2}+\left(1-\frac{2 \mu_{1}-\mu_{2}}{2\left(\beta+\mu_{1}\right)}\right)^{2 t} \beta\left\|\theta^{0}-\theta^{*}\right\|^{2}, \quad \text { and } \\
\left\|\theta^{t}-\theta^{*}\right\| & \leq \frac{4}{\mu_{1}}\left(\sqrt{\frac{2 \mathbb{P}\left[\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{2}\right] \log \frac{4}{\delta}}{n}}+\frac{G_{*} \log \frac{4}{\delta}+\mu_{1}}{n}\right)+\left(1-\frac{2 \mu_{1}-\mu_{2}}{2\left(\beta+\mu_{1}\right)}\right)^{t}\left\|\theta^{0}-\theta^{*}\right\|, \tag{1.7.4}
\end{align*}
$$

provided the sample size condition $n \geq \max \left\{\frac{c \beta^{2}}{\mu_{1}^{2}}\left(d+\log \frac{8 \log _{2}\left(2 n \Delta_{M}+2\right)}{\delta}\right), \frac{128 \mathrm{P}\left[\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{2}\right] \log \frac{4}{\delta}}{\mu_{1} \Delta_{M}}, \frac{8 G_{*} \log \frac{4}{\delta}+8 \mu_{1}}{\mu_{1} \Delta_{M}}\right\}$ holds, where c is an absolute constant.

We comment that it is usually straightforward to verify Assumption 1 for a specific missing data applications (no harder than verifying it on the completely observable case). From Lemma 2, we only need to show that the hessian matrices $\left\{\frac{\partial^{2}}{\partial \theta^{2}}\left[-\log f_{\theta}(z, w)\right]\right\}_{\theta \in \Theta}$ are sub-exponential for
all fixed $w$ and $\theta$. That is,

$$
\begin{equation*}
\mathbb{E}\left\{\exp \left(\frac{1}{\beta}\left|u_{1}^{T}\left(\frac{\partial^{2}\left[\log f_{\theta}(z, w)\right]}{\partial \theta^{2}}\right) u_{2}\right|\right)\right\} \leq 2, \tag{1.7.5}
\end{equation*}
$$

for any unit vectors $u_{1}, u_{2}$, any $z$ and any $\theta \in \Theta$. Typically, condition (1.7.5) simply requires that the observable data $z$ be a sub-Gaussian vector, regardless of the (fixed) values the unobservable data $w$ take.

As a result, Theorem 6 applies to a broad class of "learning with missing data" problems, including Example 4 and Example 5. In order to validate Assumption 6 on these two examples, a common strategy [21] is to assume the signal-to-noise ratio (SNR) to be lower bounded as

$$
\begin{equation*}
\frac{\left\|\theta^{*}\right\|}{\sigma} \geq \eta, \tag{1.7.6}
\end{equation*}
$$

for some absolute constant $\eta>0$. The following corollary holds under identical assumptions on $\eta$ as in [21].

Corollary 7 (Theorem 6 applied to Example 4 and Example 5) In both Example 4 and Example 5, after sufficiently many iterations, the first-order EM algorithm with step size 1 satisfies the generalization error bound

$$
\mathcal{E}\left(\theta^{t}\right) \leq O\left(\frac{\sigma^{2} d \log \frac{1}{\delta}}{n}\right)
$$

Specifically,

- For the Gaussian mixture model (Example 4), assuming the signal-to-noise condition (1.7.6) holds, and the initialization point satisfies $\theta^{0} \in \mathcal{B}^{d}\left(\theta^{*}, \frac{\left\|\theta^{*}\right\|}{4}\right)$, then the result of Theorem 6 holds with $\beta=1, G_{*}=\sigma \sqrt{d}, \mu_{1}=1$ and $\mu_{2}=c_{1}\left(1+\frac{1}{\eta^{2}}+\eta^{2}\right) e^{-c_{2} \eta^{2}}$, where $c_{1}, c_{2}$ are absolute constants.
- For the mixture of linear regression model (Example 5), assuming the signal-to-noise condi-
tion (1.7.6), and the initialization point satisfies $\theta^{0} \in \mathcal{B}^{d}\left(\theta^{*} ; \frac{\left\|\theta^{*}\right\|}{32}\right)$, then the result of Theorem 6 holds with $\beta=1, G_{*}=\sigma \sqrt{d}, \mu_{1}=1$ and $\mu_{2}=\frac{1}{4}$.

Proofs can be found in Appendix A.3.7. Notably, our results do not depends on $\left\|\theta^{*}\right\|$, and hence refine those in [21]. Our generalization error bounds for EM in Example 4 and Example 5 are information-theoretically optimal, because $O\left(\sigma^{2} d / n\right)$ is also the optimal rate for learning from a single Gaussian without mixture and linear regression without mixture!

### 1.7.3 Discussion and improvements over previous results

In this subsection, we will first compare our general theory with the methodology in Balakrishnan et al. [21]. Then, we will discuss the improvement over [21] as illustrated in Example 4 and Example 5. Lastly, we will provide some intuition pertaining to this improvement.

Improvements in the methodology. We now restate the theoretical result from Balakrishnan et al. [21] on the estimation error of first-order EM. Assume with probability at least $1-\delta$,

$$
\begin{equation*}
\sup _{\Omega}\left\|\left(\mathbb{P}-\mathbb{P}_{n}\right) \nabla \ell_{\theta}(\theta ; z)\right\| \leq \varepsilon_{\text {unif }}(n, \delta) . \tag{1.7.7}
\end{equation*}
$$

When the sample size $n$ is large enough, [21] proves that the first-order EM iterates $\left\{\theta^{t}\right\}_{t=0}^{\infty}$ satisfy

$$
\begin{equation*}
\left\|\theta^{t}-\theta^{*}\right\| \leq O\left(\frac{\varepsilon_{\text {unif }}(n, \delta)}{\mu_{1}-\mu_{2}}\right)+\left(1-\frac{2 \mu_{1}-\mu_{2}}{\beta+\mu_{1}}\right)^{t}\left\|\theta^{0}-\theta^{*}\right\| . \tag{1.7.8}
\end{equation*}
$$

Compared with (1.7.4) in Theorem 6, the approach in [21] has two main limitations.

- The result (1.7.8) contains a loose, global uniform convergence terms $\varepsilon_{\text {unif }}(n, \delta)$ defined via (1.7.7). In contrast, our Theorem 4 suggests that the statistical error only depends on the concentration of $\mathbb{P}_{n} \nabla \ell\left(\theta^{*} ; z\right)$ relative to $\mathbb{P} \nabla \ell\left(\theta^{*} ; z\right)$ at the single point $\theta^{*}$. The precise improvement will be illustrated on Example 4 and Example 5 shortly.
- [21] does not discuss how to calculates the complex uniform convergence term $\varepsilon^{\text {unif }}(n, \delta)$
for general models. In fact, [21] only calculate this term for Example 4. For the rest of the applications they consider, they turn to analyze sample-splitting heuristics. Although these heuristics are easier to analyze, they are less common in practice. In contrast, our Theorem 6 applies to general models without leaving the uniform convergence term unspecified.

Improvements on the examples. For the mixture of two Gaussians (Example 4), [21] proves that after sufficiently many iterations, the first-order EM algorithm satisfies the generalization error bound

$$
\begin{equation*}
\mathcal{E}\left(\theta^{t}\right) \leq O\left(\frac{\left\|\theta^{*}\right\|^{2}\left(1+\frac{\left\|\theta^{*}\right\|^{2}}{\sigma^{2}}\right) d \log \frac{1}{\delta}}{n}\right) \quad[\text { GMM result [21]] } \tag{1.7.9}
\end{equation*}
$$

and for the mixture of regressions (Example 5), [21] proves that after sufficient iterations, the first-order EM algorithm satisfies the generalization error bound

$$
\begin{equation*}
\mathcal{E}\left(\theta^{t}\right) \leq O\left(\frac{\left(\sigma^{2}+\left\|\theta^{*}\right\|^{2}\right) d \log \frac{1}{\delta}}{n}\right) \quad \text { [regression result [21]]. } \tag{1.7.10}
\end{equation*}
$$

In contrast, our problem-dependent generalization error bounds given by Corollary 7 are of the order

$$
\mathcal{E}\left(\theta^{t}\right) \leq O\left(\frac{\sigma^{2} d \log \frac{1}{\delta}}{n}\right) \quad[\text { Corollary } 7]
$$

which exhibits an improvement over the previous results (1.7.9) (1.7.10) from [21], under identical assumptions on the signal-noise ratio $\left(\left\|\theta^{*}\right\| / \sigma \geq \eta\right.$, where $\eta$ is a sufficiently large absolute constant specified in [21]). In particular, in the high signal-to-noise ratio regime, $\left\|\theta^{*}\right\|^{2} \gg \sigma^{2}$ so our improvements are significant.

Tight characterization of the statistical error is traditionally considered challenging in the area of mixture models. Only recently, [62] provided a refined analysis of the mixture of regression problem (Example 5), and proved that the achievable generalization error is indeed of the order
$\left(\sigma^{2} d \log \frac{1}{\delta} / n\right)$. However, the analysis in [62] is fairly involved and customized to the specifics of the mixture of regression setting, and it is not clear how to extend the analysis to more general problems. Our theory can be applied to quite general settings, and moreover simplifies existing approaches.

From our theory, the optimal $O\left(\sigma^{2} d \log \frac{1}{\delta} / n\right)$ characterization is very natural. Theorem 6 indicates the crucial fact that statistical error of the first-order EM algorithm only relies on $\mathbb{P}\left[\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{2}\right]$, a quantity that depends only on the optimal parameter $\theta^{*}$.

We now use Example 4 to illustrate the simplicity of our analysis. Define the function $g: \mathbb{R} \rightarrow$ $\mathbb{R}^{+}$as

$$
\begin{equation*}
g(u)=\frac{2 e^{-\frac{\left\|2 \theta^{*}-u\right\|^{2}}{2 \sigma^{2}}}}{e^{-\frac{\|u\|^{2}}{2 \sigma^{2}}}+e^{-\frac{\left\|2 \theta^{*}-u\right\|^{2}}{2 \sigma^{2}}}}, \tag{1.7.11}
\end{equation*}
$$

where $u$ is a random vector drawn from $N\left(0, \sigma^{2} I_{d \times d}\right)$. In the high SNR regime, $g(u)$ is anticipated to be very close to zero with high probability, due to the fact that

$$
\frac{\|u\|^{2}}{2 \sigma^{2}} \gg \frac{\left\|2 \theta^{*}-u\right\|^{2}}{2 \sigma^{2}} .
$$

In the Gaussian mixture model, whether $w=1$ or $w=-1$, it is straightforward to show that when conditioned on $w$, the random vector $\left(\nabla \ell\left(\theta^{*} ; z\right) \mid w\right)$ has the same distribution as $u(1-g(u))+$ $\theta^{*} g(u)$. As a result, we have

$$
\begin{equation*}
\mathbb{P}\left[\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{2}\right]=\mathbb{E}_{u}\left[\left\|u \cdot(1-g(u))+\theta^{*} \cdot g(u)\right\|^{2}\right] . \tag{1.7.12}
\end{equation*}
$$

As $g(u)$ is very close to 0 with high probability, $\mathbb{P}\left[\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{2}\right]$ should only scale with $\mathbb{E}_{u}\left[\|u\|^{2}\right]=$ $\sigma^{2} d$ rather than $\left\|\theta^{*}\right\|$. This intuition also applies to other examples like the mixture of linear regression model.

### 1.8 Concluding remarks

This paper provides contributions both in the "uniform localized convergence" approach it develops, as well as the applications thereof to various problems areas. Below we highlight some key implications.

From a methodological viewpoint, our approach resolves some fundamental limitations of the existing uniform convergence and localization analysis methods, such as the traditional "local Rademacher complexity" analysis and the "uniform convergence of gradients." At a high-level, it provides some general guidelines to derive generalization error bounds that are sharper than the worst-case uniform error. In particular, the following observations are of particular interest: 1) problem-dependent rates can often be explained by uniform inequalities whose right hand side is a function of the "free" variable $T(f) ; 2$ ) the choice of surrogate function and concentrated function are flexible, and our proposed framework brings some level of unification to localized complexities, vector-based uniform convergence results and one-sided uniform inequalities; and 3) "uniform localized convergence" arguments are also suitable to study regularization and iterative algorithms. These observations lead to a unified perspective on problem-dependent rates in various problem settings studied in the paper.

Many problem-dependent generalization error bounds proved in the paper may be of independent interest. Their study also informs the design of optimal procedures. For example: in the "slow rate" regime, we propose the first (moment-penalized) estimator that achieves optimal variance-dependent rates for general "rich" classes; and in the parametric "fast rate" regime, we show that efficient algorithms like gradient descent and the first-order Expectation-Maximization algorithm can achieve optimal problem-dependent rates in several representative problems from non-convex learning, stochastic optimization, and learning with missing data.

There are several future directions for this line of research. Applications to machine learning problems with unobservable components or causal effects may be very promising, as the focus there is typically on avoiding worst-case parameter-dependence. It is particularly intriguing to
extend our approach to study singular statistical models such as mixture models in the low SNR regime and over-parameterized low-rank matrix factorization [63]. Another interesting topic is applying the "uniform localized convergence" principle to study distributional robustness, since there are profound connections between the latter and variation-based regularization [4, 64]. Lastly, extension of our framework to overparameterized models is interesting from both the theoretical and practical viewpoints. Our results in the slow rate regime may be directly applicable to the study of overparameterized neural networks, in particular, if one has sharp upper bounds on the local Rademacher complexity. While there has been much recent interest in proving norm-based upper bounds on the global Rademacher complexity for neural network models [65, 66], proving meaningful upper bounds on the local Rademahcer complexity remains largely open. It is worthy mentioning that the recent negative results in $[67,68]$ neither apply to our general framework nor the notion of problem-dependency we suggest (distribution-dependent quantities that depend on the best hypothesis). It is possible that combining our framework with more suitable concentrated functions and localized subsets (e.g., generalizing the data-dependent subset considered in [69]) may shed light also on the study of some overparameterized models.

# Chapter 2: Upper Counterfactual Confidence Bounds: a New Optimism Principle for Contextual Bandits 

The principle of optimism in the face of uncertainty is one of the most widely used and successful ideas in multi-armed bandits and reinforcement learning. However, existing optimistic algorithms (primarily UCB and its variants) are often unable to deal with large context spaces. Essentially all existing well performing algorithms for general contextual bandit problems rely on weighted action allocation schemes; and theoretical guarantees for optimism-based algorithms are only known for restricted formulations. In this paper we study general contextual bandits under the realizability condition, and propose a simple generic principle to design optimistic algorithms, dubbed "Upper Counterfactual Confidence Bounds" (UCCB). We show that these algorithms are provably optimal and efficient in the presence of large context spaces. Key components of UCCB include: 1) a systematic analysis of confidence bounds in policy space rather than in action space; and 2) the potential function perspective that is used to express the power of optimism in the contextual setting. We further show how the UCCB principle can be extended to infinite action spaces, by constructing confidence bounds via the newly introduced notion of "counterfactual action divergence."

### 2.1 Introduction

### 2.1.1 Motivation.

Algorithms that rely on the "optimism principle" have been a major cornerstone in the study of multi-armed bandit (MAB) and reinforcement learning problems. Roughly speaking, optimistic algorithms are those that choose a deterministic action at each round, based on some optimistic estimate of future rewards. Perhaps the most representative example is the celebrated Upper Con-
fidence Bounds (UCB) algorithm and its many variants. Popularity of optimistic algorithms stems from their simplicity and effectiveness: the analysis of UCB-type algorithms are usually more straightforward than alternative approaches, so they have become the "meta-algorithms" for more complex settings. They are also often preferable to weighted allocations among actions because of the ability to discard sub-optimal actions and achieve superior instance-dependent empirical performances.

Despite their prevalent use in traditional bandit problems, existing UCB-type algorithms have a glaring drawback in contextual MAB settings: their regret often scales with the cardinality of the context space. (Notable exceptions are the special "linear payoff" formulation [70] and its generalized-linear variant [71].) In particular, despite encouraging empirical observations [72], optimism-based algorithms provably achieve sub-linear regret only under restrictive distributional assumptions [73]. This motivates the main problem studied in the paper:

## Is there a generic principle that ensures that optimistic algorithms are optimal and efficient for

 general contextual bandit problems?Interestingly, whether computationally efficient or not, almost all existing solutions to general contextual bandits [74, 75, 76, 77, 78, 79] rely on weighted, randomized allocations among actions at each round-we refer to these as "randomized algorithms" in the paper. Moreover, there is little focus on contextual MAB with infinite actions, which we believe to be a natural setting to illustrate simplicity and universality of optimism-based algorithms. These observations motivate us to search for a new optimism principle in the presence of large context spaces.e is little focus on contextual MAB with infinite actions, which we believe to be a natural setting to illustrate simplicity and universality of optimism-based algorithms. These observations motivate us to search for a new optimism principle in the presence of large context spaces.

### 2.1.2 The contextual MAB problem

The canonical stochastic contextual bandit problem can be described as follows. Let $\mathcal{A}$ be the action set (in the initial parts of the paper, one can think of $\mathcal{A}$ as the integer set $\{1, \ldots, K\}$, which
we generalize later on), and $\mathcal{X}$ be the space of contexts that supports the distribution $\mathcal{D}_{\mathcal{X}}$ (e.g., $\mathcal{X}$ can be a subset of Euclidean space). For all $x \in \mathcal{X}, a \in \mathcal{A}$, denote $\mathcal{D}_{x, a}$ a reward distribution determined by context $x$ and action $a$. At each round $t=1, \ldots, T$, the agent first observes a context $x_{t}$ drawn i.i.d. according to $\mathcal{D}_{X}$. She then chooses an action $a_{t} \in \mathcal{A}$ based on $x_{t}$ and the history $H_{t-1}$ generated by $\left\{x_{i}, a_{i}, r_{i}\left(x_{i}, a_{i}\right)\right\}_{i=1}^{t-1}$, and finally observes the reward $r_{t}\left(x_{t}, a_{t}\right)$, which is conditionally independent and distributed according to the distribution $\mathcal{D}_{x_{t}, a_{t}}$. We assume the rewards take values in the interval $[0,1]$. An admissible contextual bandit algorithm Alg is a (possibly randomized) procedure that associates each realization of $\left\{H_{t-1}, x_{t}\right\}$ with an action $a_{t}$ to employ at round $t$.

Previous literature on contextual MAB problems can be sorted into two categories: the realizable setting and the agnostic setting. In the realizable setting, the agent has access to a function class $\mathcal{F}$, with its members $f \in \mathcal{F}$ being mappings from $\mathcal{X} \times \mathcal{A}$ to $[0,1]$. The following is referred to as the realizability condition $[70,76,73,78,79]$ :

Assumption 7 (realizability) There exists $f^{*} \in \mathcal{F}$ such that for all $t \geq 1, x \in \mathcal{X}, a \in \mathcal{A}$, the conditional mean reward, $\mathbb{E}\left[r_{t}\left(x_{t}, a_{t}\right) \mid x_{t}=x, a_{t}=a\right]$, is equal to $f^{*}(x, a)$.

We call a mapping $\pi: \mathcal{X} \rightarrow \mathcal{A}$ from the context space $\mathcal{X}$ to the action set $\mathcal{A}$ a "policy." (Those mappings may be referred to more precisely as "deterministic stationary policies;" in this paper we often just refer to them as "policies" with slight abuse of terminology.) Let $\pi_{f^{*}}$, defined by $\pi_{f^{*}}(x)=\arg \max f^{*}(x, a)$, be the "ground truth" optimal policy. The cumulative (pathwise) regret of a contextual bandit algorithm Alg compared with the optimal policy $\pi_{f^{*}}$ after $T$ rounds is

$$
\operatorname{Regret}(T, \mathrm{Alg}):=\sum_{t=1}^{T}\left(r_{t}\left(x_{t}, \pi_{f^{*}}\left(x_{t}\right)\right)-r_{t}\left(x_{t}, a_{t}\right)\right)
$$

and the agent aims to minimize this cumulative regret. The agnostic setting [74, 80, 75, 77], on the other hand, does not make such realizability assumption; instead, algorithms are compared with the best policy within a given policy class. In this paper we focus on the realizable setting which lends itself more naturally to the design of optimism-based algorithms.

We present some examples of the realizable setting. The most well-studied contextual MAB problems are simple variants of the "linear payoff" model [70, 71]

$$
\begin{equation*}
\mathcal{F}=\left\{f: f(x, a)=\theta^{T} x_{a}, \theta \in \Theta\right\}, \quad \Theta, \mathcal{X} \subseteq \mathbb{R}^{d}, x=\left(x_{a}\right)_{a \in\{1, \ldots, K\}} . \tag{2.1.1}
\end{equation*}
$$

One motivation towards general function classes is to encompass models of the form

$$
\begin{equation*}
\mathcal{F}=\left\{f: f(x, a)=g_{a}(x), \quad g_{a} \in \mathcal{G}, a \in\{1, \ldots, K\}\right\}, \tag{2.1.2}
\end{equation*}
$$

where parameters of $g_{a}: \mathcal{X} \rightarrow \mathbb{R}$ can be distinct for different actions [81, 82]; it is also desirable to handle complex nonlinear models (such as neural networks) which are much more expressive than their linear counterparts.

On the computation side, we make the rather benign assumption the the agent has access to a pre-specified least square oracle over $\mathcal{F}$. Formally, after the agent inputs the historical data $\left\{x_{i}, a_{i}, r_{i}\left(x_{i}, a_{i}\right)\right\}_{i=1}^{t-1}$, the least square oracle outputs a solution $\widehat{f_{t}} \in \mathcal{F}$ that provides the best fit, namely,

$$
\begin{equation*}
\widehat{f_{t}} \in \underset{f \in \mathcal{F}}{\arg \min } \sum_{i=1}^{t-1}\left(f\left(x_{i}, a_{i}\right)-r_{i}\left(x_{i}, a_{i}\right)\right)^{2} \tag{2.1.3}
\end{equation*}
$$

This is the simplest optimization oracle assumed in the contextual bandit literature. We assume the least square oracle to be deterministic, for simplicity, as there may be multiple solutions to (2.1.3).

### 2.1.3 Introducing UCCB: two equivalent viewpoints

This subsection will describe the UCCB principle introduced in this paper from two equivalent viewpoints: 1) implicitly, it is an upper confidence bound rule in policy space; and 2) explicitly, it calculates the upper confidence bound via simulating counterfactual action trajectories rather than using the original action trajectory. For illustration purpose we focus on the finite-action setting where $\mathcal{A}=\{1, \ldots, K\}$; extension to infinite action spaces will be discussed later in Section 2.4.

Implicit strategy: maximizing upper confidence bounds in policy space. Let $\Pi$ be the policy space that contains all deterministic stationary policies $\pi: \mathcal{X} \rightarrow\{1, \ldots, K\}$. The core idea of UCCB is to choose policies that maximize certain upper confidence bounds in the policy space $\Pi$. After initialization, for each round $t$, data $\left\{\left(x_{i}, a_{i}\right), r_{i}\right\}_{i=1}^{t-1}$ is sent to an offline least square oracle to compute the estimator $\widehat{f_{t}} \in \mathcal{F}$. Without the need to "see" $x_{t}$, the agent selects the optimistic policy $\pi_{t} \in \Pi$ (which is a mapping from $\mathcal{X}$ to the action set $\{1, \ldots, K\}$ ) such that

$$
\begin{equation*}
\pi_{t} \in \underset{\pi \in \Pi}{\arg \max }\left\{\mathbb{E}_{x}\left[\widehat{f}_{t}(x, \pi(x)]+\mathbb{E}_{x}\left[\frac{\beta_{t}}{\sum_{i=1}^{t-1} \mathbb{1}\left\{\pi(x)=\pi_{i}(x)\right\}}\right]+\frac{K \beta_{t}}{t}\right\}\right. \tag{2.1.4}
\end{equation*}
$$

where the expectation $\mathbb{E}_{x}[\cdot]$ is history-independent and taken with respect to the distribution $\mathcal{D}_{X}$ (over the random context $x$ ), and $\beta_{t}$ is a parameter related to the complexity of the function class. (When there are multiple solutions to (2.1.4), we take $\pi_{t}$ to be the unique solution such that for all other solutions $\pi^{\prime}$ to (2.1.4) and all $x \in \mathcal{X}$, the index of the action $\pi_{t}(x)$ is smaller than the index of the action $\pi^{\prime}(x)$.) Then the agent observes $x_{t}$ and selects the action $a_{t}=\pi_{t}\left(x_{t}\right)$. The right hand side of (2.1.4) is an upper confidence bound on the true expected reward of $\pi$, because we can prove that with high probability, for all $\pi \in \Pi$,

$$
\left\lvert\, \mathbb{E}_{x}\left[f^{*}(x, \pi(x))\right]-\mathbb{E}_{x}\left[\widehat{f}_{t}(x, \pi(x)] \left\lvert\, \leq \mathbb{E}_{x}\left[\frac{\beta_{t}}{\sum_{i=1}^{t-1} \mathbb{1}\left\{\pi(x)=\pi_{i}(x)\right\}}\right]+\frac{K \beta_{t}}{t} .\right.\right.\right.
$$

Explicit strategy: constructing confidence bounds via counterfactual actions. The distribution $\mathcal{D}_{X}$ is unknown so there are both statistical and computational challenges in the optimization over policies. However, since our proposed policy optimization problem (2.1.4) is decomposable across contexts, there is an equivalent strategy where no explicit policy optimization is required: at round $t$, after observing $x_{t}$, the agent selects the optimistic action

$$
a_{t} \in \underset{a \in\{1, \ldots, K\}}{\arg \max }\left\{\widehat{f}_{t}\left(x_{t}, a\right)+\frac{\beta_{t}}{\sum_{i=1}^{t-1} \mathbb{1}\left\{a=\widetilde{a}_{t, i}\right\}}\right\}
$$

(ties broken by choosing the action with the smallest index), where $\left\{\widetilde{a}_{t, i}\right\}_{i=1}^{t-1}$ is the counterfactual action trajectory for context $x_{t}$, defined as realizations of all past chosen policies $\left\{\pi_{i}\right\}_{i=1}^{t-1}$ on the context $x_{t}$. To recover the counterfactual actions, at round $t$, the agent runs an inner loop to sequentially generate $\widetilde{a}_{t, 1}, \ldots, \widetilde{a}_{t, t-1}$ : for $i=1, \ldots, t-1$,

$$
\widetilde{a}_{t, i} \in \underset{a \in\{1, \ldots, K\}}{\arg \max }\left\{\widehat{f}_{i}\left(x_{t}, a\right)+\frac{\beta_{i}}{\sum_{j=1}^{i-1} \mathbb{1}\left\{a=\widetilde{a}_{t, j}\right\}}\right\},
$$

(ties are broken by choosing the action with the smallest index). Our approach is clearly quite distinct from previous variants of UCB, as we construct confidence bounds by using simulated counterfactual actions rather than using the actual selected actions.

The UCCB principle leads to provably efficient optimism-based algorithms for general function classes: their regret bounds do not scale with the cardinality of the context spaces, and the required offline least square oracle is feasible for most natural function classes.

### 2.1.4 Related literature

We review previous works in the following three areas.

Randomized solutions to general contextual bandits (with finite actions). In the agnostic contextual bandits setting, the minimax regret is $O(\sqrt{K T \log |\bar{\Pi}|})^{1}$ given a finite policy class $\bar{\Pi} \subset \Pi$. The earliest optimal solution to agnostic contextual bandits is the EXP 4 algorithm [74] whose computation is linear in $|\bar{\Pi}|$. There are two optimal oracle-efficient randomized algorithms using the cost-sensitive classification (CSC) oracle: Randomized UCB [75] and ILOVETOCONBANDIT [77].

In this paper, we focus on the realizable contextual bandits setting. Here, the minimax regret of stochastic contextual bandits is $O(\sqrt{K T \log |\mathcal{F}|})$ for a general finite function class $\mathcal{F}$. In [76] the non-efficient algorithm Regressor Elimination was proposed to achieve optimal regret.

[^1][78] proposed the use of an online regression oracle and gave an optimal and oracle-efficient algorithm called SquareCB, however the online regression oracle is only computationally efficient for specific function classes.

The open problem of optimal realizable contextual MAB with an offline least square oracle was first solved by [79], with a randomized algorithm called FALCON. One very inspiring aspect of FALCON is that weighted allocation in policy space can be implicitly achieved by weighted allocation over actions under the realizability assumption-this implication was referred to as "bypassing the monster" in [79]. This motivates the investigation in the present paper that considers implicit optimization over policies when designing optimistic algorithms. Unlike the FALCON algorithm, our approach is predicated on computing counterfactual action trajectories.

Variants of UCB for particular contextual bandit problems. Variants of LinUCB are wellknown to be regret-optimal and efficient for simple variants of (2.1.1). However, for general function classes, existing variants of UCB typically have their regret scaling with $|\mathcal{X}|$ [83], except under strong assumptions on the data distribution [73]. UCB has also been used as a subroutines in contextual bandits when the functions in $\mathcal{F}$ admit smoothness or Lipchitz continuity over $\mathcal{X}$ [84, 85]. These works are usually based on discretization of $\mathcal{X}$.

Contextual bandits with infinite actions. There is far less discussion of the infinite-action contextual bandit problem with general function classes. [86] studies how to reduce realizable contextual MAB with infinite actions to an online learning oracle called knows-what-it-knows (KWIK), but this oracle is only known to exist for restricted function classes. [78] studies how to combine general function classes with a linear action model (our illustrative example (2.4.1) in Section 2.4). However, their results crucially rely on the restrictive assumption that the action set $\mathcal{A}$ is the unit ball, and they assume access to the online regression oracle which is not computationally efficient in general. Lastly, [87] studies infinite-action contextual bandits in a quite general agnostic setting. Their formulation and results are quite different from ours, and they do not provide a computationally efficient algorithm.

### 2.1.5 Organization

In Section 2.2 we introduce an optimal and efficient optimistic algorithm in the finite-action setting, and explain the key ideas underlying its principles. For illustrative purpose we assume the function class $\mathcal{F}$ to be finite in Section 2.2, and present extensions to infinite function classes in Section 2.3. In Section 2.4 we introduce a unified framework for contextual bandits with infinite action spaces, and present several interesting examples for which our work gives rise to the first efficient solutions. In Section 2.5 we propose an optimistic subroutine to generalize randomized algorithms to the infinite-action setting.

### 2.2 Upper counterfactual confidence bounds

Following previous works $[76,73,78,79]$, we start by assuming $\mathcal{A}=\{1, \ldots, K\},|\mathcal{F}|<\infty$, and target the "gold standard" in this area-Regret $(T, \mathrm{Alg}) \leq \tilde{O}(\sqrt{K T \log |\mathcal{F}|})$, which emphasises the logarithmic scaling in the cardinality $|\mathcal{F}|$. This is mainly for illustrative purposes, and we discuss extensions to infinite function classes in Section 2.3. A relatively new setting which has essentially not been explored is the infinite-action setting, which we will discuss in Section 2.4.

### 2.2.1 The algorithm

We present the algorithm that formalizes the high-level descriptions presented in Section 2.1.3, where $\left\{\beta_{t}\right\}_{t=1}^{\infty}$ are tuning parameters that depends on the statistical complexity of $\mathcal{F}$. With the choice $\beta_{t}=\sqrt{17 t \log \left(2|\mathcal{F}| t^{3} / \delta\right) / K}$ for finite $\mathcal{F}$, the algorithm is simple and achieves $\tilde{O}(\sqrt{K T \log |\mathcal{F}|})$ regret, which is optimal up to $\log T$ factors. On the computation side, the algorithm executes no more than $T^{2}$ maximizations over actions and no more than $T$ calls to the regression oracle.

Theorem 8 (Regret for Algorithm 1) Under Assumption 7 and fixing $\delta \in(0,1)$, set the parameter $\beta_{t}$ in Algorithm 1 to be

$$
\beta_{t}=\sqrt{17 t \log \left(2|\mathcal{F}| t^{3} / \delta\right) / K}
$$

```
Algorithm 1 Upper Counterfactual Confidence Bounds (UCCB)
Input tuning parameters \(\left\{\beta_{t}\right\}_{t=1}^{\infty}\).
    for round \(t=1,2, \ldots, K\) do
        Choose action \(t\).
    for round \(t=K+1, K+2, \ldots\) do
        Compute \(\widehat{f_{t}} \in \arg \min _{f \in \mathcal{F}} \sum_{i=1}^{t-1}\left(f\left(x_{i}, a_{i}\right)-r_{i}\left(x_{i}, a_{i}\right)\right)^{2}\) via the least square oracle.
        Observe \(x_{t}\).
        for \(i=K+1, K+2, \ldots, t\) do
            Calculate the counterfactual action \(\widetilde{a}_{t, i}\) by
\[
\widetilde{a}_{t, i} \in \underset{a \in \mathcal{A}}{\arg \max }\left\{\widehat{f}_{i}\left(x_{t}, a\right)+\frac{\beta_{i}}{\sum_{j=K+1}^{i-1} \mathbb{1}\left\{a=\widetilde{a}_{t, j}\right\}+1}\right\} .
\]
(ties broken by taking the action with the smallest index)
8: \(\quad\) Take \(a_{t}=\widetilde{a}_{t, t}\) and observe reward \(r_{t}\left(x_{t}, a_{t}\right)\).
```

Then with probability at least $1-\delta$, for all $T \geq 1$, the regret of Algorithm 1 after $T$ rounds is upper bounded by
$\operatorname{Regret}(T, \operatorname{Algorithm} 1) \leq 2 \sqrt{17 K T \log \left(2|\mathcal{F}| T^{3} / \delta\right)}(\log (T / K)+1)+\sqrt{2 T \log (2 / \delta)}+K$.

Remark: Recall that our offline regression step can be solved by first-order algorithms and does not require any computation related to the confidence interval (i.e., maintaining a subset of $\mathcal{F}$ or inverting the Hessian). Therefore, despite having much broader applicability, Algorithm 1 is also simpler than many variants of UCB [70, 83, 73] from a computational perspective. The only comparable algorithm to Algorithm 1 is a randomized algorithm—FALCON in [79], which requires $O(T)$ maximizations over actions and $O(\log T)$ calls to the offline least square oracle. However, we believe our optimistic solution should be preferable in many practical settings as we do not require randomization and our regret bound exhibits much smaller constants.

### 2.2.2 Key ideas underlying UCCB

We now explain three key ideas underlying UCCB.

Key idea 1: building confidence bounds for policies. Previous literature typically refers to the optimism principle as choosing the optimistic action that has the largest estimate on the current context [83, 88, 73]-optimism is analyzed in the action space. In contrast, we view policies as decisions and build confidence bounds in policy space. The key step in our approach is to characterize the confidence bounds of the function estimate $\hat{f}_{t}$, which is the output of the least square oracle given the history $H_{t-1}$.

For an admissible non-randomized contextual bandit algorithm, at each round $t$ there exists a deterministic stationary policy $\pi_{t}$ such that the chosen action $a_{t}$ is equal to $\pi_{t}\left(x_{t}\right)$ for any realization of $x_{t}$. Equivalently, the algorithm selects $\pi_{t}$ based on $H_{t-1}$ and chooses the action $a_{t}=\pi_{t}\left(x_{t}\right)$ at round $t$. Through this viewpoint, the following lemma is applicable to all admissible nonrandomized contextual bandit algorithms:

Lemma 3 (confidence of policies) Consider an admissible non-randomized contextual bandit algorithm that selects $\pi_{t}$ based on $H_{t-1}$ (and chooses the action $a_{t}=\pi_{t}\left(x_{t}\right)$ ) at each round $t$. Then $\forall \delta \in(0,1)$, with probability at least $1-\delta / 2$, for all $t>K$ and all $\pi \in \Pi$, the estimation error on the expected reward of $\pi$ is bounded by

$$
\begin{equation*}
\left|\mathbb{E}_{x}\left[\widehat{f_{t}}(x, \pi(x))\right]-\mathbb{E}_{x}\left[f^{*}(x, \pi(x))\right]\right| \leq \sqrt{\mathbb{E}_{x}\left[\frac{1}{\sum_{i=1}^{t-1} \mathbb{1}\left\{\pi(x)=\pi_{i}(x)\right\}}\right]} \sqrt{68 \log \left(2|\mathcal{F}| t^{3} / \delta\right)} \tag{2.2.1}
\end{equation*}
$$

The proof of Lemma 3 may be interesting in its own right; a proof ketch will be presented in Section 2.2.3, and full details are deferred to Appendix B.1.2.

Key idea 2: the potential function perspective. The idea to establish confidence bounds in policy space is natural when one takes a potential function perspective. From the potential function perspective, the cumulative regret of an optimistic algorithm can be approximately bounded by the sum of confidence bounds at all rounds. Therefore, we would like to establish a uniform upper bound whatever the trajectory of policies is, which usually depends on the "entropy" of the policies. Although the number of policies is "large," the "entropy" of the policies is essentially bounded by
$\tilde{O}(K)$ in the following manner.

Lemma 4 (contextual potential lemma) Let $\pi_{t}$ be the policy that chooses action tregardless of $x$ for $t=1, \ldots, K$, and from round $K+1$ up to $T$, its actions are given by any deterministic stationary policy. Then for all $T>K$,

$$
\sum_{t=K+1}^{T} \mathbb{E}_{x}\left[\frac{1}{\sum_{j=1}^{t-1} \mathbb{1}\left\{\pi_{t}(x)=\pi_{j}(x)\right\}}\right] \leq K+K \log (T / K) .
$$

The above lemma applies to all admissible non-randomized contextual bandit algorithms that choose each action once at the first $K$ rounds, regardless of the order by which they are chosen. Proof of this lemma follows from the observation that for every $x \in \mathcal{X}$, the historical sum of $\mathbb{1}\left\{\pi_{t}(x)=\pi_{j}(x)\right\}$ will never exceeds a "per-context entropy" $O(K \log T)$. In short, analyzing confidence bounds in policy space helps us take expectation over the "per-context entropy," and successfully avoid the dependence on $|\mathcal{X}|$.

Key idea 3: the relaxation tricks and efficient computation. Following Lemma 3 and Lemma 4, a natural "upper confidence bound" strategy is to choose the policy that maximizes the following (unrelaxed) upper confidence bound:

$$
\pi_{t} \in \underset{\pi \in \Pi_{\mathcal{F}}}{\arg \max }\left\{\mathbb{E}_{x}\left[\widehat{f}_{t}(x, \pi(x)]+\sqrt{\mathbb{E}_{x}\left[\frac{1}{\sum_{i=1}^{t-1} \mathbb{1}\left\{\pi(x)=\pi_{i}(x)\right\}}\right]} \sqrt{68 \log \left(2|\mathcal{F}| t^{3} / \delta\right)}\right\},\right.
$$

where $\Pi_{\mathcal{F}}$ is the policy class defined by $\Pi_{\mathcal{F}}=\left\{\pi_{f}: \pi_{f}(x) \in \arg \max _{a \in \mathcal{A}} f(x, a), \forall x \in \mathcal{X}\right\}$, which contains $\pi_{f^{*}}$. While we can prove this strategy leads to optimal regret bounds, it is not directly feasible: 1) the distribution $\mathcal{D}_{X}$ is unknown; and 2) the optimization over policies is computationally intractable. To solve this issue, we introduce two relaxations: we "agnostically" optimize over the full policy space $\Pi$ rather than $\Pi_{\mathcal{F}}$; and we use a simple inequality to relax the confidence bound proved in Lemma 3, which we call the "square trick".

Lemma 5 (the "square trick" relaxation) The inequality (2.2.1) can be further relaxed to

$$
\begin{equation*}
\left|\mathbb{E}_{x}\left[\widehat{f}_{t}(x, \pi(x))\right]-\mathbb{E}_{x}\left[f^{*}(x, \pi(x))\right]\right| \leq \mathbb{E}_{x}\left[\frac{\beta_{t}}{\sum_{i=1}^{t-1} \mathbb{1}\left\{\pi(x)=\pi_{i}(x)\right\}}\right]+\frac{K \beta_{t}}{t} . \tag{2.2.2}
\end{equation*}
$$

Proof. Simply relax (2.2.1) by the Arithmetic Mean-Geometric Mean inequality.
By performing the two relaxations stated above, we only need to consider the optimization problem

$$
\begin{equation*}
\pi_{t} \in \underset{\pi \in \Pi}{\arg \max }\left\{\mathbb{E}_{x}\left[\widehat{f}(x, \pi(x)]+\mathbb{E}_{x}\left[\frac{\beta_{t}}{\sum_{i=1}^{t-1} \mathbb{1}\left\{\pi(x)=\pi_{i}(x)\right\}}\right]\right\} .\right. \tag{2.2.3}
\end{equation*}
$$

This is a "per context" optimization problem, where optimality at every context implies optimality of $\pi_{t}$ over the full policy space $\Pi$. The algorithm does not need to calculate $\pi_{t}$ explicitly in every step. Instead, the algorithm observes $x_{t}$, and calculates all the counterfactual actions $\pi_{1}\left(x_{t}\right), \pi_{2}\left(x_{t}\right), \ldots, \pi_{t-1}\left(x_{t}\right)$ as if the past policies were applied at $x_{t}$. Using these counterfactual actions, the algorithm calculates a counterfactual confidence, and chooses an optimistic action $a_{t}$ that maximize the upper confidence bound stated in (2.2.3).

The formula to calculate the counterfactual action $\pi_{i}\left(x_{t}\right)$,

$$
\pi_{i}\left(x_{t}\right) \in \underset{a \in \mathcal{A}}{\arg \max }\left\{\widehat{f_{i}}\left(x_{t}, a\right)+\frac{\beta_{i}}{\sum_{j=1}^{i-1} \mathbb{1}\left\{a=\pi_{j}\left(x_{t}\right)\right\}}\right\},
$$

requires us to the compute the sequence $\left\{\widetilde{a}_{t, i}\right\}_{i=1}^{t}$ in a recursive manner: for $i=1, \ldots, t$, compute

$$
\widetilde{a}_{t, i} \in \underset{a \in \mathcal{A}}{\arg \max }\left\{\widehat{f}_{i}\left(x_{t}, a\right)+\frac{\beta_{i}}{\sum_{j=K+1}^{i-1} \mathbb{1}\left\{a=\widetilde{a}_{t, j}\right\}+1}\right\} .
$$

And finally we take $a_{t}=\pi_{t}\left(x_{t}\right)=\widetilde{a}_{t, t}$. Therefore, we can explain the explicit steps in Algorithm 1 via the following (obvious) equivalence:

Lemma 6 (equivalence between Algorithm 1 and implicit strategy (2.2.3)) After the first $K$ initialization rounds, Algorithm 1 produce the same pathwise actions as those produced by the poli-
cies $\left\{\pi_{t}\right\}_{t>K}$ chosen by the upper-confidence-bound rule (2.2.3) and a specific tie-breaking rule (i.e., when there are multiple solutions to (2.2.3), taking $\pi_{t}$ to be the unique solution such that for all other solutions $\pi^{\prime}$ to (2.2.3) and all $x \in \mathcal{X}$, the index of the action $\pi_{t}(x)$ is smaller than the index of the action $\left.\pi^{\prime}(x)\right)$.

Based on all the lemmas that we introduce in this subsection, one can prove the $\widetilde{O}(\sqrt{K T \log |\mathcal{F}|})$ regret bound for Algorithm 1 through relatively standard techniques. The full proof is deferred to Appendix B.1, and a sketch is provided below.

### 2.2.3 Proof sketch of Theorem 8 and Lemma 3

In this subsection we present a proof sketch of Theorem 8 (the cumulative regret of Algorithm 1) and Lemma 3 (confidence bounds in policy space, whose relaxation leads to Lemma 5).

Proof sketch of Theorem 8. From Lemma 6, we know Algorithm 1 implicitly chooses the optimistic policy $\pi_{t}$ (i.e., solution of (2.2.3)) at each round $t$. We prove the regret bound on the event where the inequality (2.2.2) holds true for all $\pi \in \Pi$. From Lemma 5, the measure of this event is at least $1-\frac{\delta}{2}$.

Optimism of Algorithm 1 in policy space suggests that for all $t>K$,

$$
\begin{array}{r}
\mathbb{E}_{x}\left[f^{*}\left(x, \pi_{f^{*}}(x)\right] \leq \mathbb{E}_{x}\left[\widehat{f}_{t}\left(x, \pi_{f^{*}}(x)\right)\right]+\mathbb{E}_{x}\left[\frac{\beta_{t}}{\sum_{i=1}^{t} \mathbb{1}\left\{\pi_{f^{*}}(x)=\pi_{i}(x)\right\}}\right]+\frac{K \beta_{t}}{t}\right. \\
\leq \underset{\pi \in \Pi}{\arg \max }\left\{\mathbb{E}_{x}\left[\widehat{f_{t}}(x, \pi(x))\right]+\mathbb{E}_{x}\left[\frac{\beta_{t}}{\sum_{i=1}^{t} \mathbb{1}\left\{\pi(x)=\pi_{i}(x)\right\}}\right]\right\}+\frac{K \beta_{t}}{t} \\
\\
=\mathbb{E}_{x}\left[\widehat{f_{t}}\left(x, \pi_{t}(x)\right)\right]+\mathbb{E}_{x}\left[\frac{\beta_{t}}{\sum_{i=1}^{t} \mathbb{1}\left\{\pi_{t}(x)=\pi_{i}(x)\right\}}\right]+\frac{K \beta_{t}}{t} \\
\leq \mathbb{E}_{x}\left[f^{*}\left(x, \pi_{t}(x)\right)\right]+\mathbb{E}_{x}\left[\frac{2 \beta_{t}}{\sum_{i=1}^{t} \mathbb{1}\left\{\pi_{t}(x)=\pi_{i}(x)\right\}}\right]+\frac{2 K \beta_{t}}{t}
\end{array}
$$

where the first and the last inequality are due to Lemma 5; and the second inequality due to maxi-
mization over policies. Therefore, the expected regret incurred at round $t$ is bounded by

$$
\begin{equation*}
\mathbb{E}_{x}\left[f^{*}\left(x, \pi_{f^{*}}(x)\right]-\mathbb{E}_{x}\left[f^{*}\left(x, \pi_{t}(x)\right] \leq \mathbb{E}_{x}\left[\frac{2 \beta_{t}}{\sum_{i=1}^{t-1} \mathbb{1}\left\{\pi_{t}(x)=\pi_{i}(x)\right\}}\right]+\frac{2 K \beta_{t}}{t} .\right.\right. \tag{2.2.4}
\end{equation*}
$$

Taking the telescoping sum of (2.2.4) and applying the contextual potential lemma (Lemma 4), we can prove

$$
\begin{equation*}
\sum_{t=1}^{T} \mathbb{E}_{x}\left[f^{*}\left(x, \pi_{f^{*}}(x)-f^{*}\left(x, \pi_{t}(x)\right)\right] \leq 2 \sqrt{17 K T \log \left(2|\mathcal{F}| T^{3} / \delta\right)}(\log (T / K)+1)+K\right. \tag{2.2.5}
\end{equation*}
$$

By Azuma's inequality and Lemma 6, with probability at least $1-\delta / 2$, we can bound the regret by

$$
\begin{equation*}
\operatorname{Regret}(T, \operatorname{Algorithm} 1) \leq \mathbb{E}_{x}\left[f^{*}\left(x, \pi_{f^{*}}(x)-f^{*}\left(x, \pi_{t}(x)\right)\right]+\sqrt{2 T \log (2 / \delta)}\right. \tag{2.2.6}
\end{equation*}
$$

Finally we combine (2.2.5) and (2.2.6) by a union bound to finish the proof.

Proof sketch of Lemma 3. The proof of Lemma 3 includes three key steps: characterization of the estimation error (inequality (2.2.7)); a counting argument (inequality (2.2.8)); and applying Cauchy-Schwartz inequality to (2.2.8). Now we describe these key steps.

The following lemma, which holds for arbitrary algorithms, characterizes the estimation errors of an arbitrary sequence of estimators.

Lemma 7 (uniform convergence over all sequences of estimators) For an arbitrary contextual bandit algorithm, $\forall \delta \in(0,1)$, with probability at least $1-\delta / 2$,

$$
\begin{align*}
& \sum_{i=1}^{t-1} \mathbb{E}_{x_{i}, a_{i}}\left[\left(f_{t}\left(x_{i}, a_{i}\right)-f^{*}\left(x_{i}, a_{i}\right)\right)^{2} \mid H_{i-1}\right] \leq 68 \log \left(2|\mathcal{F}| t^{3} / \delta\right) \\
& \quad+2 \sum_{i=1}^{t-1}\left(f_{t}\left(x_{i}, a_{i}\right)-r_{i}\left(x_{i}, a_{i}\right)\right)^{2}-\left(f^{*}\left(x_{i}, a_{i}\right)-r_{i}\left(x_{i}, a_{i}\right)\right)^{2} \tag{2.2.7}
\end{align*}
$$

uniformly over all $t \geq 2$ and all fixed sequence $f_{2}, f_{3}, \cdots \in \mathcal{F}$.

Proof of Lemma (2.2.7) can be found in Appendix B.1.2.

Consider the contextual bandit algorithm that choose $\pi_{t}$ based on $H_{t-1}$ at each round $t$, the left hand side of $(2.2 .7)$ is equal to $\sum_{i=1}^{t-1} \mathbb{E}_{x}\left[\left(f\left(x, \pi_{i}(x)\right)-f^{*}\left(x, \pi_{i}(x)\right)\right)^{2}\right]$. Then by using the fact that $\forall \pi \in \Pi$, for all $x \in \mathcal{X}$,

$$
\mathbb{1}\left\{\pi(x)=\pi_{i}(x)\right\}\left(f_{t}(x, \pi(x))-f^{*}\left(x_{i}, \pi(x)\right)\right)^{2} \leq\left(f_{t}\left(x, \pi_{i}(x)\right)-f^{*}\left(x, \pi_{i}(x)\right)\right)^{2},
$$

we obtain the key inequality

$$
\begin{align*}
\mathbb{E}_{x}\left[\sum_{i=1}^{t-1} \mathbb{1}\{\pi(x)=\right. & \left.\left.\pi_{i}(x)\right\}\left(f_{t}(x, \pi(x))-f^{*}\left(x_{i}, \pi(x)\right)\right)^{2}\right] \leq 68 \log \left(2|\mathcal{F}| t^{3} / \delta\right) \\
& +2 \sum_{i=1}^{t-1}\left(f_{t}\left(x_{i}, a_{i}\right)-r_{i}\left(x_{i}, a_{i}\right)\right)^{2}-\left(f^{*}\left(x_{i}, a_{i}\right)-r_{i}\left(x_{i}, a_{i}\right)\right)^{2} . \tag{2.2.8}
\end{align*}
$$

We then apply Cauchy-Schwartz inequality to lower bound the left hand side of (2.2.8), and take $f_{t}=\widehat{f}_{t}$ be the least square solutions to upper bound the right hand side of (2.2.8).

### 2.3 Generalization to infinite $\mathcal{F}$

Extensions of our theory to "infinite" $\mathcal{F}$ with statistical complexity notions of covering number and parametric dimension are straightforward. Technically speaking, we only require some standard uniform convergence arguments to modify Lemma 7. We will first show that our results trivially generalizes to parametric $\mathcal{F}$ with suitable continuity, and then extend our results to general function classes following some more careful covering arguments.

Parametric dimension. Assume $\mathcal{F}$ is parametrized by a compact set $\Theta \subset \mathbb{R}^{d}$ whose diameter is bounded by $\Delta$, and satisfies

$$
\begin{equation*}
\left|f_{\theta_{1}}(x, a)-f_{\theta_{2}}(x, a)\right| \leq L\left\|\theta_{1}-\theta_{2}\right\|, \tag{2.3.1}
\end{equation*}
$$

uniformly over $x \in \mathcal{X}$ and $a \in \mathcal{A}$. This case clearly covers many previous structured models (variants of the "linear payoff" formulation (2.1.1)).

Corollary 9 (extension to infinite $\mathcal{F}$ via parametric dimension) Under Assumption 7 and the assumption (2.3.1) and fixing $\delta \in(0,1)$, set the parameter $\beta_{t}$ in Algorithm 1 to be

$$
\beta_{t}=\sqrt{34 t / K} \sqrt{d \log (2+\Delta L t)+\log \left(2 t^{3} / \delta\right)+1} .
$$

Then Algorithm 1 satisfies that with probability at least $1-\delta$, for all $T \geq 1$,

$$
\operatorname{Regret}(T, \operatorname{Algorithm} 1) \leq 2 K \beta_{T}(\log (T / K)+1)+\sqrt{2 T \log (2 / \delta)}+K=\tilde{O}(\sqrt{K T d})
$$

Remark: While this regret bound has a worse dependence on $K$ in the "linear payoff" formulation (2.1.1) compared with SupLinUCB in [70] (whose regret is logarithmic in $K$ ), Algorithm 1 can be applied in more general parametric settings and enjoys much lower computational demands (there is no need to invert any Hessian). While the square-root dependence on $K$ can not be improved for general $\mathcal{F}$ (see the lower bound in [76]), we can improve this dependence for structured models by applying our results in Section 2.4.

Covering number formulation. Our results can be extended to general (possibly non-parametric) function classes via covering numbers and standard uniform convergence techniques. We consider formulation (2.3.2)—a major target of previous works on general contextual bandits [81, 73, 78]. We assume access to a general function class $\mathcal{G}$ that contains mappings from $\mathcal{X}$ to $[0,1]$, and assume

$$
\begin{equation*}
\mathcal{F}=\left\{f: f(x, a)=g_{a}(x), \quad g_{a} \in \mathcal{G}\right\} . \tag{2.3.2}
\end{equation*}
$$

Definition 5 (covering number) For a function class $\mathcal{G}$ that contains mappings from $\mathcal{X}$ to $[0,1]$ and fixed $n \in \mathbb{Z}_{+}$, an empirical $L_{1}$ cover on a sequence $x_{1}, \ldots, x_{n}$ at scale $\varepsilon$ is a set $U \subseteq \mathbb{R}^{n}$ such that $\forall g \in \mathcal{G}, \exists u \in U, \frac{1}{n} \sum_{i=1}^{n}\left|g\left(x_{n}\right)-u_{n}\right| \leq \varepsilon$. We define the covering number $\mathcal{N}_{1}\left(\mathcal{G}, \varepsilon,\left\{x_{i}\right\}_{i=1}^{n}\right)$ to
be the size of the smallest such cover.

Given careful covering arguments proved in [81, 73], the following extension is straightforward:

Corollary 10 (extension to infinite $\mathcal{F}$ via covering number) Under Assumption 7 and the assumption (2.3.2), given $T \geq 1$ and $\delta \in(0,1)$, by setting all the parameters $\beta_{t}$ in Algorithm 1 to be a fixed value

$$
\beta=\sqrt{T K} \cdot \inf _{\varepsilon>0}\left\{25 \varepsilon T+80 \log \left(\frac{8 K T^{3} \mathbb{E}_{\left\{x_{i}\right\}_{i=1}^{T}} \mathcal{N}_{1}\left(\mathcal{G}, \varepsilon,\left\{x_{i}\right\}_{i=1}^{T}\right)}{\delta}\right)\right\} .
$$

Then, Algorithm 1 satisfies that with probability at least $1-\delta$,

$$
\operatorname{Regret}(T, \operatorname{Algorithm} 1) \leq 2 K \beta(\log (T / K)+1)+\sqrt{2 T \log (2 / \delta)}+K
$$

### 2.4 A unified framework for infinite action spaces

In this section we study infinite-action contextual bandits to illustrate the simplicity and applicability of the UCCB principle. In context-free settings, discussion on infinite actions can be sorted into two streams. The first stream studies variants of the linear action model. Prominent examples include linearly parametrized bandit [89, 88], and parametrized bandit with generalized linear model [90]. The second stream is based on discretization over actions and reduction to the finite-action setting (e.g., Lipchitz bandit ([91]). We focus on the first stream here, as it exhibits additional challenges of efficient exploration beyond the finite-action setting.

To focus on the core messages, we assume $\mathcal{F}$ to be finite and function in $\mathcal{F}$ take values in $[0,1]$. We propose a generic algorithm (Algorithm 2) that achieves

$$
\operatorname{Regret}(T, \operatorname{Algorithm} 2) \leq \tilde{O}(\sqrt{\mathcal{E} \log |\mathcal{F}| T})
$$

for many models of interest. Here we call $\mathcal{E}:=\mathbb{E}_{x}\left[\mathcal{E}_{x}\right]$ the "average decision entropy," where $\mathcal{E}_{x}$ is (informally) the complexity of the "fixed- $x$-model" where the context is fixed to be $x$. Note
that unlike previous complexity measures such as "Eluder dimension" [83], the "average decision entropy" $\mathcal{E}$ does not scale with $|\mathcal{X}|$ so that this complexity measure is much more useful in the contextual settings. We will present several interesting illustrative examples, and present key ideas of our algorithm using these examples.

### 2.4.1 Illustrative models

In the context-free infinite-action bandits literature, it is well-known that $\tilde{O}(\sqrt{T})$-type regret is only possible for structured models, among which variants of linear bandits are the preponderant models. As a result, our framework mainly targets settings where all "fixed- $x$-model" are variants of linear bandits.

Example 6 (contextual bandit with linear action model) Given a general vector-valued function class $\mathcal{G}$ that contains mappings from $\mathcal{X}$ to $\mathbb{R}^{d}$, let

$$
\begin{equation*}
\mathcal{F}=\left\{f: \exists g \in \mathcal{G} \text { s.t. } f(x, a)=g(x)^{\top} a, \forall x \in \mathcal{X}, \forall a \in \mathcal{A}\right\} . \tag{2.4.1}
\end{equation*}
$$

We assume $\mathcal{A} \subset \mathbb{R}^{d}$ is an arbitrary compact set, and is available for the agent at all rounds. This formulation is a strict generalization of the finite-action realizable contextual bandit problem we studied in previous sections (it reduces to the $K$-armed setting when $\mathcal{A}$ is the set of $K$ element vectors in $\mathbb{R}^{K}$ ). Another special case where $\mathcal{A}$ is restricted to be the unit ball is studied in Foster and Rakhlin [78], but a general solution to arbitrary compact action set is still open. Moreover, Foster and Rakhlin [78] requires online regression oracles which are not computationally efficient in general. Formulation (2.4.1) was also studied in [92] but the goal there was off-policy evaluation rather than regret minimization.

With knowledge on linear bandits we can prove $\mathcal{E}_{x}=d$ for all $x \in \mathcal{X}$. (detailed explanation is deferred to Section 2.4.4). Therefore $\mathcal{E}=d$, which is independent of the number of actions, and the order of regret is expected to be $\tilde{O}(\sqrt{d \log |\mathcal{F}| T})$.

Example 7 (contextual bandit with generalized linear action model.) Consider a broader choice of models, which contains generalized linear action models and allows a mapping $\varphi$ :

$$
\begin{equation*}
\mathcal{F}=\left\{f: \exists g \in \mathcal{G} \text { s.t. } f(x, a)=\sigma_{x}\left(g(x)^{\top} \varphi(x, a)\right), \forall x \in \mathcal{X}, \forall a \in \mathcal{A}\right\}, \tag{2.4.2}
\end{equation*}
$$

where for every $x \in \mathcal{X}, \sigma_{x}: \mathbb{R} \rightarrow[0,1]$ is a known link function that satisfies

$$
\frac{\sup _{a} \sigma_{x}^{\prime}\left(\left\langle g^{*}(x), \varphi(x, a)\right\rangle\right)}{\inf _{a} \sigma_{x}^{\prime}\left(\left\langle g^{*}(x), \varphi(x, a)\right\rangle\right)} \leq \kappa_{x}
$$

and $\varphi: \mathcal{X} \times \mathcal{A} \rightarrow \mathbb{R}^{d}$ is a known compactness-preserving mapping (e.g., continuous mappings). This model generalizes (2.4.1) and allows more flexibility. When we set $\varphi(x, a)=x_{a}$, we see that this model is significantly broader in scope than the simple "linear payoff" formulation (2.1.1), as $g(x)$ is a general function that depends on $x$ rather than a fixed parameter $\theta$.

Our analysis will show that $\mathcal{E}_{x}=\kappa_{x}^{2}$ d for all $x \in \mathcal{X}$ (detailed explanation is deferred to Section 2.4.4), so that $\mathcal{E}=\mathbb{E}_{x}\left[\kappa_{x}^{2}\right] d$, and the order of regret is expected to be $\tilde{O}\left(\sqrt{E_{x}\left[\kappa_{x}^{2}\right] d \log |\mathcal{F}| T}\right)$.

Example 8 (heterogeneous action set) Many real-world, customized pricing and personalized healthcare applications have a high dimensional action set $\mathcal{A}$, but the "effective dimension" of available actions after observing $x$ is usually much smaller. To model these applications, consider the reward model

$$
\begin{equation*}
\mathcal{F}=\left\{f: \exists g \in \mathcal{G} \text { s.t. } f(x, a)=\sigma_{x}\left(g(x)^{\top} a\right), \forall x \in \mathcal{X}, \forall a \in \mathcal{A}(x)\right\}, \tag{2.4.3}
\end{equation*}
$$

where for all $x \in \mathcal{X}$ we assume a compact action set $\mathcal{A}(x) \subset \mathcal{A}$, and assume $\mathcal{A}(x)$ is contained in a $d_{x}$-dimensional subspace. When the agent observes context $x$, she can only choose her action from $\mathcal{A}(x)$.

For this model we have $\mathcal{E}_{x}=\kappa_{x}^{2} d_{x}$ (detailed explanation is deferred to Section 2.4.4) so that $\mathcal{E}=\mathbb{E}_{x}\left[\kappa_{x}^{2} d_{x}\right]$. The salient point here is the we avoid dependence on the full dimension d. Regret therefore scales as $\tilde{O}\left(\sqrt{\mathbb{E}_{x}\left[\kappa_{x}^{2} d_{x}\right] \log |\mathcal{F}| T}\right)$.

### 2.4.2 Counterfactual action divergence

The main modification required for infinite-action settings is predicated on a central concept called "counterfactual action divergence," which generalizes the term $\left(\sum_{i=1}^{n} \mathbb{1}\left\{a=a_{i}\right\}\right)^{-1}$ that was used in Algorithm 1. This new concept characterizes "how much information" is learned from action $a$ given a sequence $\left\{a_{i}\right\}_{i=1}^{n}$, on the "fixed- $x$-model."

Definition 6 (counterfactual action divergence) For fixed integer $n$, a context $x$, an action $a$ and a sequence of actions $\left\{a_{i}\right\}_{i=1}^{n}$, we say $V_{x}\left(a \|\left\{a_{i}\right\}_{i=1}^{n}\right)$ is a proper choice of the counterfactual action divergence between $a$ and $\left\{a_{i}\right\}_{i=1}^{n}$ evaluated at $x$, if

$$
V_{x}\left(a| |\left\{a_{i}\right\}_{i=1}^{n}\right) \geq \sup _{f \in \mathcal{F}}\left\{\frac{\left|f(x, a)-f^{*}(x, a)\right|^{2}}{\sum_{i=1}^{n}\left(f\left(x, a_{i}\right)-f^{*}\left(x, a_{i}\right)\right)^{2}}\right\} .
$$

We define $V_{x}(a \| \emptyset)=\infty$ in the case $n=1$.

Using the definition of counterfactual action divergence, the expectation

$$
\begin{equation*}
\mathbb{E}_{x}\left[V_{x}\left(\pi(x) \|\left\{\pi_{i}(x)\right\}_{i=1}^{t-1}\right)\right], \tag{2.4.4}
\end{equation*}
$$

can be used to construct an upper confidence bound on the expected reward of policy $\pi$ given the past chosen policies $\{\pi\}_{i=1}^{t-1}$. Similar to the finite-action setting, the agent chooses the optimistic policy $\pi_{t}$ that maximizes this confidence bound, and chooses $a_{t}=\pi_{t}\left(x_{t}\right)$ without explicitly computing $\pi_{t}$-this is achieved by sequentially recovering counterfactual actions, as will be illustrated in our proposed Algorithm.

Convenient choices of $\left.V_{x}\left(a \|\left\{a_{i}\right\}_{i=1}^{n}\right\}_{i=1}^{n}\right)$ should be taken case by case for different problems. In the following lemma, we present closed-form choices of $\left.V_{x}\left(a \|\left\{a_{i}\right\}_{i=1}^{n}\right\}_{i=1}^{n}\right)$ in all our illustrative examples.

Statement 3 (illustration of counterfactual action divergences) In the illustrative examples, the counterfactual action divergences are given as follows (and taken as $\infty$ when inverse of matrices is not well-defined):

- finite-action contextual bandit:

$$
V_{x}\left(a| |\left\{a_{i}(x)\right\}_{i=1}^{n}\right)=\frac{1}{\sum_{i=1}^{n} \mathbb{1}\left\{a=a_{i}\right\}} .
$$

- linear action model (2.4.1):

$$
\begin{equation*}
\left.V_{x}\left(a \mid\left\{a_{i}\right\}_{i=1}^{n}\right\}_{i=1}^{n}\right)=a^{\top}\left(\sum_{i=1}^{n}\left[a_{i} a_{i}^{\top}\right]\right)^{-1} a \tag{2.4.5}
\end{equation*}
$$

- generalized linear action model (2.4.2):

$$
V_{x}\left(a \|\left\{a_{i}\right\}_{i=1}^{n}\right)=\kappa_{x}^{2} \varphi(x, a)^{\top}\left(\sum_{i=1}^{n}\left[\varphi\left(x, a_{i}\right) \varphi\left(x, a_{i}\right)^{\top}\right]\right)^{-1} \varphi(x, a) .
$$

- generalized linear action model with heterogeneous action sets (2.4.3):

$$
V_{x}\left(a \|\left\{a_{i}(x)\right\}_{i=1}^{n}\right)=\kappa_{x}^{2} b_{x, a} \top\left(\sum_{i=1}^{n}\left[b_{x, a_{i}} b_{x, a_{i}}^{\top}\right]\right)^{-1} b_{x, a},
$$

where $b_{x, a}$ is the coefficient vector of a with a basis $\left\{A_{x, 1}, \ldots, A_{x, d_{x}}\right\}$ of $\mathcal{A}(x)$, i.e.,

$$
a=\left[A_{x, 1}, \ldots, A_{x, d_{x}}\right] b_{x, a}
$$

### 2.4.3 The algorithm and regret bound

Algorithm 2 is a generalization of Algorithm 1 to the infinite-action setting. It can be applied to most parametric action models that have been studied in the context-free setting, and handles heterogeneous action sets. Recall that $\mathcal{E}:=\mathbb{E}\left[\mathcal{E}_{x}\right]$ is the average decision entropy of the problem, for which we will give the formal definition later. The "initialization oracle" and the "action maximization oracle" will also be explained shortly.

Algorithm 2 essentically provide a reduction from contextual models to the "fixed- $x$-models." The regret of an optimistic algorithm is usually upper bounded by the sum of confidence bounds.

```
Algorithm 2 Upper Counterfactual Confidence Bound-Infinite Action (UCCB-IA)
Input tuning parameters \(\left\{\beta_{t}\right\}_{t=1}^{\infty}\).
    for round \(t=1,2, \ldots\) do
        Compute \(\widehat{f_{t}}=\arg \min _{f \in \mathcal{F}} \sum_{t=1}^{t-1}\left(f\left(x_{t}, a_{t}\right)-r_{t}\left(x_{i}, a_{t}\right)\right)^{2}\) via the least square oracle.
        Observe \(x_{t}\), use the initialization oracle to obtain initializations \(\left\{A_{x_{t}, i}\right\}_{i=1}^{d_{x}}\).
        for \(i=1,2, \ldots, t \vee d_{x}\) do
            Take \(\widetilde{a}_{t, i}=A_{x_{t}, i}\).
        for \(i=t \wedge\left(d_{x}+1\right), \ldots, t\) do
            Use the action maximization oracle to compute counterfactual actions:
\[
\widetilde{a}_{t, i} \in \underset{a \in \mathcal{A}\left(x_{t}\right)}{\arg \max }\left\{\widehat{f}_{i}\left(x_{t}, a\right)+\beta_{i} V_{x_{t}}\left(a \|\left\{\widetilde{a}_{t, j}\right\}_{j=1}^{i-1}\right)\right\} .
\]
8: \(\quad\) Take \(a_{t}=\widetilde{a}_{t, t}\) and observe reward \(r_{t}\left(x_{t}, a_{t}\right)\).
```

In our case, the sum of expectations (2.4.4) is decomposable over contexts, so tractability of the "fixed- $x$-models" suffices to make Algorithm 2 provably efficient. Formally, we require regularity conditions so that the "fixed- $x$-models" are solvable by the optimism principle. Motivated by the standard potential arguments used in the linear bandit literature, we make Assumption 8 below. Verification of this assumption on Examples 6-8 will be presented in the next section.

Assumption 8 (per-context models are solvable by optimism) There exists counterfactual action divergences such that the following are satisfied:
i)for all $x \in \mathcal{X}$, there exists $d_{x}$ actions $A_{x, 1}, \ldots, A_{x, d_{x}} \in \mathcal{A}(x)$ such that $V_{x}\left(a \|\left\{A_{x, i}\right\}_{i=1}^{d_{x}}\right)<\infty$ for all $a \in \mathcal{A}(x)$.
ii) For all $x \in \mathcal{X}$, there exists $\mathcal{E}_{x}>0$ such that for all $T \geq 1$ and all sequences $\left\{a_{t}\right\}_{t=1}^{T}$ that satisfy $\left\{a_{t}\right\}_{t=1}^{d_{x} \wedge T}=\left\{A_{x, t}\right\}_{t=1}^{d_{x} \wedge T}$, we have

$$
\sum_{t=1}^{T}\left[1 \wedge V_{x}\left(a_{t} \|\left\{a_{j}\right\}_{j=1}^{t-1}\right)\right] \leq \mathcal{E}_{x} \operatorname{poly}(\log T)
$$

for all $x \in \mathcal{X}$, where $\operatorname{poly}(\cdot)$ is a fixed polynomial-scale function.

Given positive values $\mathcal{E}_{x}$ that satisfies condition ii) in Assumption 8 , we define $\mathcal{E}_{x}:=\mathbb{E}_{x}\left[\mathcal{E}_{x}\right]$ to be (a proper choice of) the "average decision entropy" of the problem. The "average decision
entropy" of a problem is not unique, and any "proper" choice of $\mathcal{E}$ leads to a rigorous regret bound of Algorithm 2.

Besides the least-square oracle, Algorithm 2 uses two other optimization oracles that are necessary in the infinite-action setting: 1) a deterministic initialization oracle which returns $\left\{A_{x, i}\right\}_{i=1}^{d_{x}}$ satisfying Assumption 8 after inputting $\mathcal{A}(x)$ (this is standard for Examples 6-8 using the theory of barycentric spanners, see the next subsection); and 2) a deterministic action maximization oracle whose output is a maximizer of a function over the feasible region $\mathcal{A}(x)$.

After imposing the regularity conditions proposed in Assumption 8, the regret of Algorithm 2 can be bounded as the follows.

Theorem 11 (Regret of Algorithm 2) Under Assumptions 7 and 8 and fixing $\delta \in(0,1)$, let

$$
\beta_{t}=\sqrt{17 t \log \left(2|\mathcal{F}| t^{3} / \delta\right) / \mathcal{E}}
$$

Then with probability at least $1-\delta$, for all $T \geq 1$ the regret of Algorithm 2 after $T$ rounds is upper bounded by
$\operatorname{Regret}(T, \operatorname{Algorithm} 2) \leq 2 \sqrt{17 \mathcal{E} T \log \left(2|\mathcal{F}| T^{3} / \delta\right)}(\operatorname{poly}(\log T)+1)+\sqrt{2 T \log (2 / \delta)}+\mathcal{E}$.

This theorem immediately provides regret bounds for all our illustrative examples, which we will discuss in the next subsection.

Finally, we give a high-level interpretation of the average decision entropy $\mathcal{E}$ : if the expectation (2.4.4) is the "discrete" partial gradient of a potential function, then the historical sum has the path independence property-that is, the historical sum of (2.4.4) can be bounded by the maximum value of a potential function, which is characterized by the average decision entropy $\mathcal{E}$. Since $\mathcal{E}$ is the average rather than the sum of the effective complexities of all "fixed- $x$-models," UCCB provides a generic solution to achieve optimal regret bounds that do not scale with $|\mathcal{X}|$.

### 2.4.4 Applications in illustrative examples

In this subsection we will carefully go through the three illustrative examples. We summarize the conclusions in the following corollary:

Corollary 12 (Theorem 11 applied to illustrative examples) Examples 6-8 satisfy Assumptions 8 with the average decision entropy given by

- linear action model (2.4.1): $\mathcal{E}=d$.
- generalized linear action model (2.4.2): $\mathcal{E}=\mathbb{E}_{x}\left[\kappa_{x}^{2}\right] d$.
- generalized linear action model with heterogeneous action sets (2.4.3): $\mathcal{E}=\mathbb{E}_{x}\left[\kappa_{x}^{2} d_{x}\right]$.

Now we give a verification in the remaining parts of this subsection.

## Contextual bandits with linear action model (Example 6).

We begin with contextual bandits with linear action model (2.4.1), with the homogeneous action set $\mathcal{A}$. For this problem, Algorithm 1 only needs to compute the initialization actions $A_{1}, \ldots, A_{d}$ once, and use them during the first $d$ rounds. This suffices to complete the required initialization for all contexts.

Based on well-known results in the linear bandit literature, it is straightforward to show that $\mathcal{E}=d$, because we can take $\mathcal{E}_{x}=d$ for every per-context model. The details are as follows.

As shown in Statement 3, for all $x \in \mathcal{X}$, we choose the counterfactual action divergence between any $a_{t}$ and any sequence $\left\{a_{i}\right\}_{i=1}^{t-1}$ evaluated at $x$ to be

$$
V_{x}\left(a_{t} \|\left\{a_{i}\right\}_{i=1}^{t-1}\right)=a_{t}^{\top}\left(\sum_{i=1}^{t-1}\left[a_{i} a_{i}^{\top}\right]\right)^{-1} a_{t} .
$$

Following the standard approach in the linear bandit literature (e.g., see [89]), we choose the $d$ initialization actions $\left\{A_{i}\right\}_{i=1}^{d}$ to be the barycentric spanner of $\mathcal{A}$. A barycentric spanner is a set of $d$ vectors, all contained in $\mathcal{A}$, such that every vector in $\mathcal{A}$ can be expressed as a linear combination
of the spanner with coefficients in $[-1,1]$. An efficient algorithm to find the barycentric spanner for an arbitrary compact set is given in [93].

The following result follows Lemma 9 in [89] ${ }^{2}$, which is often referred to as the "elliptical potential lemma": let $a_{i}=A_{i}$ for $i=1, \ldots, d$, then for all $T>d$ and all trajectory $\left\{a_{t}\right\}_{t=d+1}^{T}$,

$$
\sum_{t=d+1}^{T}\left[1 \wedge a_{t}^{\top}\left(\sum_{i=1}^{t-1}\left[a_{i} a_{i}^{\top}\right]\right)^{-1} a_{t}\right] \leq 2 d \log T
$$

Therefore, we obtain for all $T \geq 1$ and all $x \in \mathcal{X}$,

$$
\begin{equation*}
\sum_{t=1}^{T}\left[1 \wedge V_{x}\left(a_{t} \|\left\{a_{i}\right\}_{i=1}^{t-1}\right)\right] \leq 2 d \log T+d \leq 3 d \log T \tag{2.4.6}
\end{equation*}
$$

By taking $\left\{A_{i}\right\}_{i=1}^{d}$ to be a barycentric spanner of $\mathcal{A}$, setting $\mathcal{E}=d$, and taking poly $(\log T)=$ $3 \log T$, Assumption 8 holds for problem (2.4.1). Despite the illustration here, we also note that our Assumption 8 is not restricted to any particular choice of initialization actions and $\mathcal{E}$ : there are other ways to choose linearly independent initialization actions, giving rise to a slightly different $\operatorname{poly}(\log T)$ term in Assumption 8 (see, e.g., Lemma 11 in [88]).

## Contextual bandits with generalized linear action model (Example 7).

For the problem formulation (2.4.2), we can take $\mathbb{E}\left[\mathcal{E}_{x}\right]=\mathbb{E}\left[\kappa_{x}^{2}\right] d$. The details are as follows.
As shown in Statement 3, given $x \in \mathcal{X}$, we choose the counterfactual action divergence between any $a_{t}$ and any sequence $\left\{a_{i}\right\}_{i=1}^{t-1}$ evaluated at $x$ to be

$$
V_{x}\left(a_{t} \|\left\{a_{i}\right\}_{i=1}^{t-1}\right)=\kappa_{x}^{2} \varphi\left(x, a_{t}\right)^{\top}\left(\sum_{i=1}^{t-1}\left[\varphi\left(x, a_{i}\right) \varphi\left(x, a_{i}\right)^{\top}\right]\right)^{-1} \varphi\left(x, a_{t}\right) .
$$

Given $x \in \mathcal{X}$, we take $\left\{A_{x, i}\right\}_{i=1}^{d}$ such that $\left\{\varphi\left(x, A_{x, i}\right)\right\}_{i=1}^{d}$ consists of a barycentric spanner of $\{\varphi(x, a): a \in \mathcal{A}\}^{3}$. Note that a different basis $\left\{A_{x, i}\right\}_{i=1}^{d}$ should be computed for each $x$. From

[^2]our previous result (2.4.6) and the fact $\kappa_{x} \geq 1$, for all $T \geq 1$ and all sequences $\left\{a_{i}\right\}_{i=1}^{T}$ that satisfy $\left\{a_{i}\right\}_{i=1}^{d_{x} \wedge T}=\left\{A_{x, i}\right\}_{i=1}^{d_{x} \wedge T}$, $\sum_{t=1}^{T}\left[1 \wedge V_{x}\left(a_{t} \|\left\{a_{i}\right\}_{i=1}^{t-1}\right)\right]=\sum_{t=1}^{T}\left[1 \wedge \kappa_{x}^{2} \varphi(x, a)^{\top}\left(\sum_{i=1}^{t-1}\left[\varphi\left(x, a_{i}\right) \varphi\left(x, a_{i}\right)^{\top}\right]\right)^{-1} \varphi(x, a)\right] \leq \kappa_{x}^{2} 3 d \log T$.

By taking $\mathcal{E}_{x}=\kappa_{x}^{2} d$, and poly $(\log T)=3 \log T$, Assumption 8 holds with $\mathcal{E}=\mathbb{E}_{x}\left[\kappa_{x}^{2}\right] d$.

## Contextual bandits with heterogeneous action set (Example 8)

We consider the problem formulation (2.4.3) where the action set $\mathcal{A}(x)$ is heterogeneous for different $x \in \mathcal{X}$. Note that $\mathcal{A}(x)$ is a compact set contained in a $d_{x}$-dimensional subspace. Given $x \in \mathcal{X}$, we choose $\left\{A_{x, i}\right\}_{i=1}^{d_{x}}$ as the barycentric spanner of $\mathcal{A}(x)$ and take $a_{i}=A_{x, i}$ for $i=1, \ldots, d_{x}$. As stated in Statement 3, given $x \in \mathcal{X}$, the counterfactual action divergence between $a_{t}$ and $\left\{a_{i}\right\}_{i=1}^{t-1}$ evaluated at $x$ is

$$
\left.V_{x}\left(a_{t} \|\left\{a_{i}\right\}_{i=1}^{t-1}\right)\right)=\kappa_{x}^{2} b_{x, a_{t}}^{\top}\left(\sum_{i=1}^{t-1} b_{x, a_{t}} b_{x, a_{t}}^{\top}\right)^{-1} b_{x, a_{t}},
$$

where $b_{x, a_{t}}$ is the coefficient vector of $a_{t}$ with respect to the basis $\left\{A_{x, i}\right\}_{i=1}^{d_{x}}$. From our previous result (2.4.6) and the fact $\kappa_{x} \geq 1$, for all $T \geq 1$ and all sequences $\left\{a_{i}\right\}_{i=1}^{T}$ that satisfy $\left\{a_{i}\right\}_{i=1}^{d_{x} \wedge T}=$ $\left\{A_{x, i}\right\}_{i=1}^{d_{x} \wedge T}$,

$$
\sum_{t=1}^{T} 1 \wedge\left[V_{x}\left(a_{t}| |\left\{a_{i}\right\}_{i=1}^{t-1}\right)\right]=\sum_{t=1}^{T}\left[1 \wedge \kappa_{x}^{2} b_{x, a_{t}}^{\top}\left(\sum_{i=1}^{t-1}\left[b_{x, a_{i}} b_{x, a_{i}}^{\top}\right]\right)^{-1} b_{x, a_{t}}\right] \leq \kappa_{x}^{2} 3 d_{x} \log T
$$

By taking $\mathcal{E}_{x}=\kappa_{x}^{2} d_{x}$, and poly $(\log T)=3 \log T$, we verify Assumption 8 with $\mathcal{E}=\mathbb{E}\left[\kappa_{x}^{2}\right] d_{x}$. We note that under the heterogeneous formulation, Algorithm 2 needs to compute a different basis for each $\mathcal{A}(x)$, and the computation of counterfactual action divergence also requires a coefficient decomposition for each $x \in \mathcal{X}$.

One significant advantage of Algorithm 2 is that the regret does not rely on the full dimension

[^3]$d$-this means that we can increase feature context as long as we can control the average decision entropy $\mathbb{E}\left[\kappa_{x}^{2} d_{x}\right]$.

### 2.5 Using "optimistic subroutines" to generalize randomized algorithms

What is the connection between our proposed optimistic algorithms and existing randomized algorithms? In this section, we show that by combining the idea of counterfactual confidence bounds and a non-trivial "optimistic subroutine," we can also generalize an existing randomized algorithm to the infinite-action setting. However, the analysis and implementation of the resulting randomized algorithm is much more complex than the optimistic algorithm we introduced before. Through this extension, we see the simplicity and importance of the optimism principle for complex settings like infinite-action spaces.

The first least-square-oracle-efficient randomized algorithm in the general realizable contextual bandits, FALCON from [79], is restricted to the finite-action setting. FALCON performs implicit optimization in policy space, but the allocation of policies reduces to a closed-form weighted allocation rule for actions (this design principle also influences the design of UCCB). We find that it becomes more crucial to exploit the counterfactual confidence bounds in the infinite-action setting: the optimization of weighted allocation rules no longer has closed-form solutions, and we need to design a technical "optimistic subroutine" to find feasible weighted allocations.

As the required subroutine is a bit complex, we focus on the linear action model (2.4.1) stated in Example 6 for simplicity. Extensions to more complex models follow similar ideas, and the structure of the proposed algorithm remain mostly unchanged. We assume a deterministic initialization oracle that outputs a barycentric spanner of the compact set $\mathcal{A}$ (e.g. the algorithm in [93]), and an action maximization oracle that outputs the maximizer of the input function over $\mathcal{A}$. The following algorithm extends FALCON to the linear action model (2.4.1), where the step 6 is a novel optimization problem to find the "right" weighted allocation over actions. Here the " $a^{\top}\left(\mathbb{E}_{\widetilde{a} \sim p_{t}}\left[\widetilde{a} \widetilde{a}^{\top}\right]\right)^{-1} a$ " term in (2.5.2) is a continuous analogue to the counterfactual action divergence (2.4.5).

Algorithm 3 runs in an epoch schedule and only calls the least square oracle at the pre-specified

```
Algorithm 3 a generalized version of FALCON for linear action model (2.4.1)
Input epoch schedule \(\left\{\tau_{m}\right\}_{m=1}^{\infty}, \tau_{0}=0\), tuning parameters \(\left\{\beta_{m}\right\}_{m=1}^{\infty}\), an arbitrary function \(\widehat{f_{1}} \in \mathcal{F}\).
    for epoch \(m=1,2, \ldots\) do
        Compute \(\widehat{f_{m}}=\arg \min _{f \in \mathcal{F}} \sum_{t=1}^{\tau_{m}-1}\left(f\left(x_{t}, a_{t}\right)-r_{t}\left(x_{t}, a_{t}\right)\right)^{2}\) via the least square oracle when
        \(m \geq 1\).
        for round \(t=\tau_{m-1}+1, \ldots, \tau_{m}\) do
            Observe context \(x_{t}\).
            Use the action maximization oracle to compute \(\widehat{a}_{t} \in \max _{a \in \mathcal{A}} \widehat{f}_{m}\left(x_{t}, a\right)\).
            Run the algorithm OptimisticSubroutine \(\left(\mathcal{A}, \widehat{a}_{t}, \widehat{f}_{m}\left(x_{t}, \cdot\right), \beta_{m}\right)\) to find a distribu-
            tion \(p_{t}\) over \(\mathcal{A}\) such that,
\[
\begin{array}{ll} 
& \mathbb{E}_{a \sim p_{t}}\left[\widehat{f}_{m}\left(x_{t}, \widehat{a}_{t}\right)-\widehat{f}_{m}\left(x_{t}, a\right)\right] \leq 2 \beta_{m} d, \\
\forall a \in \mathcal{A}, & \widehat{f}_{m}\left(x_{t}, a\right)+\beta_{m} a^{\top}\left(\mathbb{E}_{\widetilde{a} \sim p_{t}}\left[\widehat{a}^{\top}\right]\right)^{-1} a \leq \widehat{f}_{m}\left(x_{t}, \widehat{a}_{t}\right)+2 \beta_{m} d . \tag{2.5.2}
\end{array}
\]
7: \(\quad\) Sample \(a_{t} \sim p_{t}\) and observe reward \(r_{t}\left(a_{t}\right)\).
```

rounds $\tau_{1}, \tau_{2}, \ldots$ We take $\tau_{m}=2^{m}$ for all $m \geq 1$ to simplify the statement of the theorem, though other choices of the epoch schedule are also possible [79].

Theorem 13 (Regret of Algorithm 3) Consider the problem formulation (2.4.1) stated in Example 6, under Assumption 7. Take the epoch schedule $\tau_{m}=2^{m}$ for $m \geq 1$. Let

$$
\beta_{m}=30 \sqrt{\log \left(|\mathcal{F}| \tau_{m-1} / \delta\right) /\left(2 d \tau_{m-1}\right)}
$$

for $m=2, \ldots$, and $\beta_{1}=1$. Then with probability at least $1-\delta$, for all $T \geq 1$, the regret of Algorithm 3 after $T$ rounds is upper bounded by

$$
\operatorname{Regret}(T, \operatorname{Algorithm} 3) \leq 608.5 \sqrt{2 d T \log (|\mathcal{F}| T / \delta)}+2 \sqrt{2 T \log (2 / \delta)}+2
$$

Theorem 13 can be obtained by modifying the regret analysis of the original FALCON algorithm. (We refer the readers to [79] for the background and intuition of the original FALCON algorithm, especially the "Observation 2 " in that paper.) However, the key challenge is to provide an efficient algorithm to find a weighted allocation rule that satisfy both (2.5.1) and (2.5.2) in Algorithm 3.

Algorithm 4 OptimisticSubroutine $(\widehat{a}, \beta, \mathcal{A}, \widehat{h})$
input action set $\mathcal{A}$, greedy action $\widehat{a} \in \mathcal{A}$, function $\widehat{h}: \mathcal{A} \rightarrow[0,1]$, parameter $\beta>0$.
1: Obtain a barycentric spanner $\left\{A_{i}\right\}_{i=1}^{d}$ of $\mathcal{A}$ via the initialization oracle.
2: Set $q_{0}=\sum_{i=1}^{d} \frac{1}{d} \mathbb{1}_{A_{i}}$.
3: for $t=1,2, \ldots$ do
4: Set

$$
\begin{equation*}
q_{t-\frac{1}{2}}=\min \left\{\frac{2 d}{2 d+\mathbb{E}_{a \sim q}[(\widehat{h}(\widehat{a})-\widehat{h}(a)) / \beta]}, 1\right\} \cdot q_{t-1} \tag{2.5.3}
\end{equation*}
$$

5: Use the action maximization oracle to compute

$$
\begin{equation*}
a_{t}=\underset{a \in \mathcal{A}}{\arg \max }\left\{\widehat{h}(a)+\beta a^{\top}\left(\mathbb{E}_{\widetilde{a} \sim q_{t-\frac{1}{2}}}\left[\widetilde{a a}^{\top}\right]\right)^{-1} a\right\} . \tag{2.5.4}
\end{equation*}
$$

6: if $\widehat{h}\left(a_{t}\right)+\beta a_{t}^{\top}\left(\mathbb{E}_{a \sim q_{t-\frac{1}{2}}}\left[a a^{\top}\right]\right)^{-1} a_{t}>\widehat{h}(\widehat{a})+2 \beta d$, then
7: $\quad$ Run the coordinate descent step

$$
\begin{equation*}
q_{t}=q_{t-\frac{1}{2}}+\frac{-2 a_{t}^{\top}\left(\mathbb{E}_{a \sim q}\left[a a^{\top}\right]\right)^{-1} a_{t}+2 d+\left(\widehat{h}(\widehat{a})-\widehat{h}\left(a_{t}\right)\right) / \beta}{\left(a_{t}^{\top}\left(\mathbb{E}_{a \sim q}\left[a a^{\top}\right]\right)^{-1} a_{t}\right)^{2}} \mathbb{1}_{a_{t}} . \tag{2.5.5}
\end{equation*}
$$

8: else
9: $\quad$ Let $q_{t}=q_{t-\frac{1}{2}}$, halt and output

$$
q_{t}+\left(1-\int_{\mathcal{A}} q_{t}(a) \mathrm{d} a\right) \mathbb{1}_{\widehat{a}}
$$

We provide Algorithm 4 as a subroutine to achieve this. The core idea of this algorithm is to use a coordinate descent procedure to compute a sparse distribution over actions, which is motivated by the optimization procedure used in Agarwal et. al. [77]-however, we extend their idea from the finite-action setting to the linear action model, which requires further matrix analysis and may be interesting in its own right. We call this algorithm OptimisticSubroutine as the algorithm is built upon the optimistic step (2.5.4), where the " $a^{\top}\left(\mathbb{E}_{\widetilde{a} \sim q_{t-\frac{1}{2}}}\left[\widetilde{a} \widetilde{a}^{\top}\right]\right)^{-1} a$ " term is a continuous analogue to the counterfactual action divergence (2.4.5) in the linear action model.

Proposition 3 (optimization through SubOpt) At each round within epoch m, Algorithm 4 outputs a probability distribution that satisfies (2.5.1) and (2.5.2) within at most $\left\lceil\frac{4}{\beta_{m}}+8 d(\log d+1)\right\rceil$ iterations.

According to this proposition, the optimistic subroutine outputs an efficient solution that satisfies the requirements (2.5.1) and (2.5.2) within finite number of iterations at every rounds. One advantage of Algorithm 3 is that it requires only $O(\log T)$ calls to the least-square oracle. However, the design and analysis of the optimistic subroutine becomes challenging in the infinite-action setting, especially for complex problem formulations. On the other hand, Algorithm 2 exhibits much cleaner structure and a principled analysis that covers many problem formulations of interest.

### 2.6 Conclusion and future directions

In this paper we propose $U C C B$, a simple generic principle to design optimistic algorithms in the presence of large context spaces. Key ideas underlying UCCB include: 1) confidence bounds in policy space rather than in action space; and 2) the potential function perspective that explains the power of optimism in the contextual setting. We present the first optimal and efficient optimistic algorithm for realizable contextual bandits with general function classes. Besides the traditional finite-action setting, we also discuss the infinite-action setting and provide the first solutions to many interesting models of practical interest.

Moving forward, there are many interesting future directions that may leverage the ideas presented in this work. The principle of optimism in the face of uncertainty plays an essential role in reinforcement learning. Currently the majority of existing provably efficient algorithms are developed for the "tabular" case, and their regret scales with the cardinality of the state space. However, empirical reinforcement learning problems typically have a large state space and rely on function approximation [94]. Motivated by this challenge, a natural next step is to adapt the UCCB principle to reinforcement learning problems with large state space. This paper can be viewed as an initial step towards this goal, as the contextual MAB problem is a special case of episodic reinforcement learning where the episode length is equal to one. Within the scope of bandit problems, UCBtype algorithms are often the "meta-algorithms" for many complex formulations when there is no contextual information. Since UCCB improves over UCB-type algorithms in several fundamental contextual settings, this work may be a building block to combine contextual information and function approximation with more complex formulations such as Gaussian process optimization [95], bandits with long-term constraints [96], and bandits in non-stationary environments [97]. We leave these directions to future work.

## Chapter 3: Bayesian Design Principles for Frequentist Sequential Learning

We develop a general theory to optimize the frequentist regret for sequential learning problems, where efficient bandit and reinforcement learning algorithms can be derived from unified Bayesian principles. We propose a novel optimization approach to create "algorithmic beliefs" at each round, and use Bayesian posteriors to make decisions. This is the first approach to make Bayesian-type algorithms prior-free and applicable to adversarial settings, in a generic and optimal manner. Moreover, the algorithms are simple and often efficient to implement. As a major application, we present a novel algorithm for multi-armed bandits that achieves the "best-of-allworlds" empirical performance in the stochastic, adversarial, and non-stationary environments. And we illustrate how these principles can be used in linear bandits, bandit convex optimization, and reinforcement learning.

### 3.1 Introduction

### 3.1.1 Background

We address a broad class of sequential learning problems in the presence of partial feedback, which arise in numerous application areas including personalized recommendation [98], game playing [99] and control [100]. An agent sequentially chooses among a set of possible decisions to maximize the cumulative reward. By "partial feedback" we mean the agent is only able to observe the feedback of her chosen decision, but does not generally observe what the feedback would be if she had chosen a different decision. For example, in multi-armed bandits (MAB), the agent can only observe the reward of her chosen action, but does not observe the rewards of other actions. In reinforcement learning (RL), the agent is only able to observe her state insofar as the chosen action is concerned, while other possible outcomes are not observed and the underlying
state transition dynamics are unknown. In this paper, we present a unified approach that applies to bandit problems, reinforcement learning, and beyond.

The central challenge for sequential learning with partial feedback is to determine the optimal trade-off between exploration and exploitation. That is, the agent needs to try different decisions to learn the environment; at the same time, she wants to focus on "good" decisions that maximize her payoff. There are two basic approaches to study such exploration-exploitation trade-off: frequentist and Bayesian. One of the most celebrated examples of the frequentist approach is the family of Upper Confidence Bound (UCB) algorithms [101, 102]. Here, the agent typically uses sample average or regression to estimate the mean rewards; and she optimizes the upper confidence bounds of the mean rewards to make decisions. Another widely used frequentist algorithm is EXP3 [74] which was designed for adversarial bandits; it uses inverse probability weighting (IPW) to estimate the rewards, and then applies exponential weighting to construct decisions. One of the most celebrated examples of the Bayesian approach is Thompson Sampling (TS) with a pre-specified, fixed prior [103]. Here, the agent updates the Bayesian posterior at each round to learn the environment, and she uses draws from that posterior to optimize decisions.

The advantage of the frequentist approach is that it does not require a priori knowledge of the environment. However, it heavily depends on a case-by-case analysis exploiting special structure of a particular problem. For example, regression-based approaches can not be easily extended to adversarial problems; and IPW-type estimators are only known for simple rewards/losses such as discrete and linear. The advantage of Bayesian approach is that Bayesian posterior is a generic and often optimal estimator if the prior is known. However, the Bayesian approach requires knowing the prior at the inception, which may not be accessible in complex or adversarial environments. Moreover, maintaining posteriors is computationally expensive for most priors.

In essence, frequentist approach requires less information, but is less principled, or more bottom-up. On the other hand, the Bayesian approach is more principled, or top-down, but requires stronger assumptions. In this paper we focus on the following research:

Can we design principled Bayesian-type algorithms, that are prior-free, computationally effi-
cient, and work well in both stochastic and adversarial/non-stationary environments?

### 3.1.2 Contributions

In this paper, we synergize frequentist and Bayesian approaches to successfully answer the above question, through a novel idea that creates "algorithmic beliefs" that are generated sequentially in each round, and uses Bayesian posteriors to make decisions. Our contributions encompass over theoretical discoveries, novel methodology, and applications thereby. We summarize the main contributions as follows.

Making Bayesian-type algorithms prior-free and applicable to adversarial settings. To the best of our knowledge, we provide the first approach that allows Bayesian-type algorithms to operate without prior assumptions and be applicable in adversarial settings, in a generic, optimal, and often computationally efficient manner. The regret bounds of our algorithms are no worse than the best theoretical guarantees known in the literature. In addition to its applicability in adversarial/non-stationary environments, our approach offers the advantages of being prior-free and often computationally manageable, which are typically not achievable by traditional Bayesian algorithms, except for simple model classes like discrete and linear rewards/losses. It is worth noting that the main ideas underlying our methodology and proofs are quite insightful and can be explained in a succinct manner.

General theory of "Algorithmic Information Ratio" (AIR). We introduce an objective function that depends on an "algorithmic belief" and round-dependent information, which we refer to as "Algorithmic Information Ratio" (AIR). Our approach always selects algorithmic beliefs by (approximately) maximizing AIR, and the regret of our algorithms can always be bounded by the cumulative sum of the values of AIR at each round. We then show that AIR can be upper bounded by previously known complexity measures such as information ratio (IR) [104] and decision-estimation coefficient (DEC) [105]. As an immediate consequence, our machinery converts existing regret upper bounds using information ratio and DEC, into simple frequentist algo-
rithms with tight guarantees. And we provide methods and guarantees to approximately maximize AIR.

Novel algorithm for MAB with "best of all worlds" empirical performance. As a major illustration, we propose a novel algorithm for Bernoulli multi-armed bandits (MAB) that achieves the "best-of-all-worlds" empirical performance in the stochastic, adversarial, and non-stationary environments. This algorithm is quite different from and performs much better than the traditional EXP3 algorithm, which has been the default choice for adversarial MAB for decades. At the same time, the algorithm outperforms UCB and is comparable to Thompson Sampling in the stochastic environment. Moreover, it outperforms traditional Thompson Sampling and "clairvoyant" restarted algorithms in non-stationary environments.

Applications to linear bandits, bandit convex optimization, and reinforcement learning. Our theory can be applied to various settings, including linear and convex bandits, and reinforcement learning, by the principle of approximately maximizing AIR. Specifically, for linear bandits, we derive a modified version of EXP2 based on our framework, which establishes a novel connection between inverse propensity weighting (IPW) and Bayesian posteriors. For bandit convex optimization, we propose the first algorithm that attains the best-known $\tilde{O}\left(d^{2.5} \sqrt{T}\right)$ regret with a finite poly $\left(e^{d} \cdot T\right)$ running time. Lastly, in reinforcement learning, we provide a simple and constructive algorithm that achieves the sharpest result proven through Bayesian Thompson Sampling in the early work [105].

Combining estimation and decision-making. Our approach jointly optimizes the belief of an environment and probability of decision. Most existing algorithms including UCB, EXP3, Estimation-to-Decision (E2D) [105], TS, and Information-Directed Sampling (IDS) [106] maintain a different viewpoint that separates algorithm design into a black-box estimation method (sample average, linear regression, IPW, Bayesian posterior...) and a decision-making rule that makes the estimate as input to an optimization problem. In contrast, by optimizing AIR to generate new beliefs, our
algorithm simultaneously deals with estimation and optimization. This viewpoint is quite powerful and broadens the general scope of bandit algorithms.

### 3.1.3 Related literature

In this subsection we discuss mostly relevant works. [104, 106] propose the concept of "information ratio" to analyze and design Bayesian bandit algorithms. Their work studies Bayesian regret with a known prior rather than the frequentist regret. [107] proposes an algorithm called "Exploration by Optimization (EBO)," which is the first general frequentist algorithm that optimally bounds the frequentist regret of bandit problems using information ratio. However, the EBO algorithm is more of a conceptual construct as it requires intractable optimization over the complete class of "functional estimators," and hence is not implementable in most settings of interest. Our algorithms are inspired by EBO, but are simpler in structure and run in decision and model spaces (rather than intractable functional spaces). In particular, our approach advances EBO by employing explicit construction and randomization of estimators, offering flexibility in selecting updating rules, and providing computation guidelines that come with provable guarantees. The recent work [105] proposes the concept of "decision-estimation coefficient" (DEC) as a general complexity measure for bandit and reinforcement learning problems. Algorithms in this work typically separate black-box estimation method and decision-making rule, and for this reason the proposed E2D algorithm do not generally achieve optimal regret for bandit problems. The subsequent work [108] extends the theory of DEC to adversarial environments. However, their algorithm is an adaptation of EBO in [107], which, as discussed, may present computational challenges.

### 3.2 Preliminaries and definition of AIR

### 3.2.1 Problem formulation

To state our results in the broadest manner, we adopt the general formulation of Adversarial Decision Making with Structured Observation (Adversarial DMSO) [108], which covers broad problems including bandit problems, reinforcement learning, and partial monitoring. For a locally
compact metric space we denote by $\Delta(\cdot)$ the set of Borel probability measures on that space. Let $\Pi$ be a compact decision space. Let $\mathcal{M}$ be a compact model class where each model $M: \Pi \rightarrow O$ is a mapping from the decision space to a locally compact observation space $O$. A problem instance in this protocol can be described by the decision space $\Pi$ and the model class $\mathcal{M}$. We define the mean reward function associated with model $M$ by $f_{M}$.

Consider a $T$-round game played by a randomized player in an adversarial environment. At each round $t=1, \ldots, T$, the agent determines a probability $p_{t}$ over the decisions, and the environment selects a model $M_{t} \in \mathcal{M}$. Then the decision $\pi_{t} \sim p_{t}$ is sampled and an observation $o_{t} \sim M_{t}\left(\pi_{t}\right)$ is revealed to the agent. An admissible algorithm ALG can be described by a sequence of mappings where the $t$-th mapping maps the past decision and observation sequence $\left\{\pi_{i}, o_{i}\right\}_{i=1}^{t-1}$ to a probability $p_{t}$ over decisions. The frequentist regret of the algorithm ALG against the usual target of single best decision in hindsight is defined as

$$
\mathfrak{R}_{T}=\sup _{\pi^{*} \in \Pi} \mathbb{E}\left[\sum_{t=1}^{T} f_{M_{t}}\left(\pi^{*}\right)-\sum_{t=1}^{T} f_{M_{t}}\left(\pi_{t}\right)\right],
$$

where the expectation is taken with respect to the randomness in decisions and observations. There is a large literature that focuses on the so-called stochastic environment, where $M_{t}=M^{*} \in \mathcal{M}$ for all rounds, and the single best decision $\pi^{*} \in \arg \min f_{M^{*}}(\pi)$ is the natural oracle. Regret bounds for adversarial sequential learning problems naturally apply to stochastic problems. We illustrate how the general formulation covers bandit problems, and leave the discussion of reinforcement learning to Section 3.7.

Example 9 (Bernoulli multi-armed bandits (MAB)) We illustrate how the general formulation reduces to the basic MAB problem with Bernoulli reward. Let $\Pi=[K]=\{1, \cdots, K\}$ be a finite set of $K$ actions, and $\mathcal{F}$ be the set of all possible mappings from $[K]$ to $[0,1]$. Take $\mathcal{M}=\left\{M_{f}: f \in\right.$ $\mathcal{F}\}$ as the induced model class, where each $M_{f}$ maps $\pi$ into the Bernoulli distribution Bern $(f(\pi))$. The mean reward function for model $M_{f}$ is $f$ itself. At each round $t$, the environment selects a mean reward function $f_{t}$, and the observation $o_{t}$ is the incurred reward $r_{t} \sim \operatorname{Bern}\left(f_{t}\left(\pi_{t}\right)\right)$.

Example 10 (Structured bandits) We consider bandit problems with general structure of the mean reward function. Let $\Pi$ be a d-dimensional action set, and $\mathcal{F} \subseteq\{f: \Pi \rightarrow[0,1]\}$ be a function class that encodes the structure of the mean reward function. Take $\mathcal{M}=\left\{M_{f}: f \in \mathcal{F}\right\}$ as the induced model class, where each $M_{f}$ maps $\pi$ to the Bernoulli distribution Bern $(f(\pi))$. The mean reward function for model $M_{f}$ is $f$ itself. For example, in $d$-dimensional linear bandits, the mean reward function $f$ is parametrized by some $\theta \in \Theta \subseteq \mathbb{R}^{d}$ such that $f(\pi)=\theta^{T} \pi, \forall \pi \in \Pi$. And in bandit convex optimization, the mean reward (or loss) function class $\mathcal{F}$ is the set of all concave (or convex) mappings from $\Pi$ to $[0,1]$.

### 3.2.2 Algorithmic Information Ratio

Let $v$ be a probability measure of the joint random variable $\left(M, \pi^{*}\right) \in \mathcal{M} \times \Pi$, and $p$ be a distribution of another independent random variable $\pi \in \Pi$. Given a probability measure $v$, let

$$
v_{\pi^{*}}(\cdot)=\int_{\mathcal{M}} v(M, \cdot) d M
$$

be the marginal distribution of $\pi^{*} \in \Pi$. Viewing $v$ as a prior belief over $\left(M, \pi^{*}\right)$, we define $v(\cdot \mid \pi, o)$ as the Bayesian posterior belief conditioned on the decision being $\pi$ and the observation (generated from the distribution $M(\pi)$ ) being $o$; and we define the marginal posterior belief of $\pi^{*}$ conditioned on $\pi$ and $o$ as

$$
v_{\pi^{*} \mid \pi, o}(\cdot)=\int_{\mathcal{M}} v(M, \cdot \mid \pi, o) d M
$$

Denote $\operatorname{KL}(\mathbb{P}, \mathbb{Q})=\int \log \frac{d \mathbb{P}}{d \mathbb{Q}} d \mathbb{P}$ as the KL divergence between two probability measures.
Now we introduce a central definition in this paper-Algorithmic Information Ratio.
Definition 7 (Algorithmic Information Ratio) Given a reference probability $q \in \operatorname{int}(\Delta(\Pi))$ and learning rate $\eta>0$, we define the "Algorithmic Information Ratio" (AIR) for probability $p$ of $\pi$ and belief $v$ of $\left(M, \pi^{*}\right)$ as

$$
\operatorname{AIR}_{q, \eta}(p, v)=\mathbb{E}_{p, v}\left[f_{M}\left(\pi^{*}\right)-f_{M}(\pi)-\frac{1}{\eta} \operatorname{KL}\left(v_{\pi^{*} \mid \pi, o}, q\right)\right],
$$

where the expectation is taken with $\pi \sim p,\left(M, \pi^{*}\right) \sim v$.

The term "Algorithmic Information Ratio" was used to highlight the key difference between AIR and classical information ratio (IR) measures (to be presented shortly in (3.2.1)). Firstly, AIR incorporate a reference probability $q$ in its definition, while classical IR does not. This additional flexibility makes AIR useful for algorithm design and analysis. Secondly, AIR is defined in an offset form, whereas IR is defined in a ratio form. We choose the word "algorithmic" because AIR is particularly suited to designing constructive and efficient frequentist algorithms. We remain the term "ratio" as it is consistent with previous literature on the topic. The formulation of AIR is inspired by the optimization objectives in recent works on EBO [107] and DEC [105], but AIR crucially depends an "algorithmic belief" $v$ rather than taking the maximum with respect to the worst-case deterministic model, and there are multiple differences in the formulations (see the next subsection for details).

Note that AIR is linear with respect to $p$ and concave with respect to $v$, as conditional entropy is always concave with respect to the joint probability measure (see Lemma 36). It will be illustrative to write AIR as the sum of three items:

$$
\begin{array}{r}
\operatorname{AIR}_{q, \eta}(p, v)=\underbrace{\mathbb{E}_{p, v}\left[f_{M}\left(\pi^{*}\right)-f_{M}(\pi)\right]}_{\text {expected regret }} \\
-\frac{1}{\eta} \underbrace{\mathbb{E}_{p, v}\left[\operatorname{KL}\left(v_{\pi^{*} \mid \pi, o}, v_{\pi^{*}}\right)\right]}_{\text {information gain }}-\frac{1}{\eta} \underbrace{\operatorname{KL}\left(v_{\pi^{*}}, q\right)}_{\text {regularization by } q},
\end{array}
$$

where: the "expected regret" measures the difficulty of exploitation; "information gain" is the amount of information gained about $\pi^{*}$ by observing $\pi$ and $o$, and this in fact measures the degree of exploration; and the last "regularization" term forces the marginal distribution of $\pi^{*}$ to be "close" to the reference probability distribution $q$. By maximizing AIR, we generate an "algorithmic belief" that simulates the worst-case environment. This algorithmic belief will automatically balance exploration and exploitation, as well as being close to the chosen reference belief (e.g., a standard reference is the posterior from previous round, as used in traditional Thompson Sampling).

### 3.2.3 Relation to IR and DEC

Notably, our framework allows for the utilization of nearly all existing upper bounds for information ratio (IR) and DEC in practical applications, enabling the derivation of the sharpest regret bounds known, along with the development of constructive algorithms. In this subsection we demonstrate that AIR can be upper bounded by IR and DEC.

We present here the traditional definition of Bayesian information ratio [104]. See [104, 106, 109, 110, 111] for upper bounds of IR in structured bandit and RL problems.

Definition 8 (Information ratio) Given belief $v$ of $\left(M, \pi^{*}\right)$ and decision probability $p$ of $\pi$, the information ratio is defined as

$$
\begin{equation*}
\operatorname{IR}(v, p)=\frac{\left(\mathbb{E}_{v, p}\left[f_{M}\left(\pi^{*}\right)-f_{M}(\pi)\right]\right)^{2}}{\mathbb{E}_{v, p}\left[\operatorname{KL}\left(v_{\pi^{*} \mid \pi, o}, v_{\pi^{*}}\right)\right]} . \tag{3.2.1}
\end{equation*}
$$

Note that the traditional information ratio (3.2.1) does not involve any reference probability distribution $q$ (unlike AIR). By completing the square, it is easy to show that AIR can always be bounded by IR as follows.

Lemma 8 (Bounding AIR by IR) For any $q \in \operatorname{int}(\Delta(\Pi))$, $p \in \Delta(\Pi)$, belief $v \in \Delta(\mathcal{M} \times \Pi)$, and $\eta>0$, we have

$$
\operatorname{AIR}_{q, \eta}(p, v) \leq \frac{\eta}{4} \cdot \operatorname{IR}(v, p)
$$

The recent paper [105] introduced the complexity measure DEC, which aims to unify bandits and many reinforcement learning problems.

Definition 9 (Decision-estimation coefficient) Given a model class $\mathcal{M}$, a nominal model $\bar{M}$ and $\eta>0$, we define the decision-estimation coefficient by

$$
\operatorname{DEC}_{\eta}(\mathcal{M}, \bar{M})=\inf _{p \in \Delta(\Pi)} \sup _{M \in \mathcal{M}} \mathbb{E}_{v, p}\left[f_{M}\left(\pi_{M}\right)-f_{M}(\pi)-\frac{1}{\eta} D_{\mathrm{H}}^{2}(M(\pi), \bar{M}(\pi))\right],
$$

where $D_{\mathrm{H}}^{2}(\mathbb{P}, \mathbb{Q})=\int(\sqrt{d \mathbb{P}}-\sqrt{d \mathbb{Q}})^{2}$ is the squared Hellinger distance between two probability measures.

DEC provides a tighter complexity measure compared to several existing measures in the literature, such as the bilinear dimension [112] and the Eluder Bellman dimension [113], for reinforcement learning (RL) problems. Moreover, a slightly strengthened version of DEC, defined through the KL divergence instead of the Hellinger divergence, can be bounded by the traditional information ratio. This result follows from Proposition 9.1 in [105].

We establish that the worst-case value of AIR under a "maximin" strategy for selecting $p$ is bounded by the worst-case value of the decision-estimation coefficient (DEC) for the convex hull of the model class in the following lemma.

Lemma 9 (Bounding AIR by DEC) Given model class $\mathcal{M}$ and $\eta>0$, we have

$$
\begin{equation*}
\sup _{q \in \operatorname{int}(\Delta(\Pi))} \sup _{v} \inf _{p} \operatorname{AIR}_{q, \eta}(p, v) \leq \sup _{\bar{M} \in \operatorname{conv}(\mathcal{M})} \operatorname{DEC}_{\eta}(\Delta(\mathcal{M}), \bar{M}) . \tag{3.2.2}
\end{equation*}
$$

To prove Lemma 9, we can start by noting that the left-hand side of (3.2.2) is equivalent to the "parametric information ratio," defined as

$$
\begin{equation*}
\max _{v} \min _{p} \mathbb{E}_{v, p}\left[f_{M}\left(\pi^{*}\right)-f_{M}(\pi)-\frac{1}{\eta} \mathrm{KL}\left(v_{\pi \mid \pi, o}, v_{\pi^{*}}\right)\right], \tag{3.2.3}
\end{equation*}
$$

which was introduced in [108]. This equivalence can be shown by using the concavity of AIR to exchange sup over $q$ and min over $p$. Furthermore, the inequality between (3.2.3) and the righthand side of (3.2.2) has been established by Theorem 3.1 in [108]. Therefore, we obtain a proof of Lemma 9.

We highlight that AIR is the tightest complexity measure in the adversarial setting. However, for reinforcement learning problems in the stochastic setting, it is often desirable to remove the convex hull on the right-hand side of (3.2.2). To this end, we introduce a tighter version of AIR, called "Model-index AIR" (MAIR), which allows us to apply most existing regret upper bounds
using DEC to our framework. In Section 3.7, we discuss our theory about MAIR and its application to RL in the stochastic setting.

### 3.3 Algorithms

### 3.3.1 A generic regret bound leveraging AIR

Given an arbitrary admissible algorithm ALG (defined in Section 3.2.1), we can generate a sequence of algorithmic beliefs $\left\{v_{t}\right\}_{t=1}^{T}$ and a corresponding sequence of reference probabilities $\left\{q_{t}\right\}_{t=1}^{T}$ in a sequential manner as shown in Algorithm 5. Maximizing AIR to create algorithmic

```
Algorithm 5 Maximizing AIR to create algorithmic beliefs
Input algorithm ALG and learning rate \(\eta>0\).
Initialize \(q_{1}\) to be the uniform distribution over \(\Pi\).
```

    for round \(t=1,2, \cdots, T\) do
        Obtain \(p_{t}\) from ALG. Find a distribution \(v_{t}\) of \(\left(M, \pi^{*}\right)\) that solves
    $$
\sup _{v \in \Delta(\mathcal{M} \times \Pi)} \operatorname{AIR}_{q_{t}, \eta}\left(p_{t}, v\right) .
$$

3: $\quad$ The algorithm ALG samples decision $\pi_{t} \sim p_{t}$ and observes the feedback $o_{t} \sim M_{t}\left(\pi_{t}\right)$.
4: $\quad$ Update $q_{t+1}=\left(v_{t}\right)_{\pi^{*} \mid \pi_{t}, o_{t}}$.
beliefs is an alternative approach to traditional estimation procedures, as the resulting algorithmic beliefs will simulate the true or worst-case environment. In particular, this approach only stores a single distribution $\left(v_{t}\right)_{\pi^{*} \mid \pi_{t}, o_{t}}$ at round $t$, which is the Bayesian posterior obtained from belief $v_{t}$ and observations $\pi_{t}, o_{t}$, and it is made to forget all the rest information from the past.

Based on these algorithmic beliefs, we can provide regret bound for an arbitrary algorithm. Here we assume $\Pi$ to be finite (but potentially large) for simplicity; this assumption can be relaxed using standard discretization and covering arguments.

Theorem 14 (Generic regret bound for arbitrary learning algorithm) Given a finite decision space $\Pi$, a compact model class $\mathcal{M}$, the regret of an arbitrary learning algorithm ALG is bounded
as follows, for all $T \in \mathbb{N}_{+}$,

$$
\begin{equation*}
\mathfrak{R}_{T} \leq \frac{\log |\Pi|}{\eta}+\sum_{t=1}^{T} \operatorname{AIR}_{q_{t}, \eta}\left(p_{t}, v_{t}\right) \tag{3.3.1}
\end{equation*}
$$

Note that Theorem 14 provides a powerful tool to study the regret of an arbitrary algorithm using the concept of AIR. More importantly, it suggests that the algorithm should choose decision with probability $p_{t+1}$ according to the posterior $\left(\left(v_{t}\right)_{\pi^{*} \mid \pi_{t}, o_{t}}\right.$. Building on this principle to generate algorithmic beliefs, we provide two concrete algorithms: "Adaptive Posterior Sampling" (APS) and "Adaptive Minimax Sampling" (AMS). Surprisingly, their regret bounds are as sharp as the best known regret bounds of existing Bayesian algorithms that require knowledge of a wellspecified prior.

### 3.3.2 Adaptive Posterior Sampling (APS)

When the agent always selects $p_{t+1}$ to be equal to the posterior $q_{t+1}=\left(v_{t}\right)_{\pi^{*} \mid \pi_{t}, o_{t}}$, and optimizes for algorithmic beliefs as in Algorithm 5, we call the resulting algorithm "Adaptive Posterior Sampling" (APS).

```
Algorithm 6 Adaptive Posterior Sampling (APS)
Input learning rate \(\eta>0\).
Initialize \(p_{1}=\operatorname{Unif}(\Pi)\).
    for round \(t=1,2, \cdots, T\) do
        Find a distribution \(v_{t}\) of \(\left(M, \pi^{*}\right)\) that solves
```

$$
\sup _{v \in \Delta(\mathcal{M} \times \Pi)} \operatorname{AIR}_{p_{t}, \eta}\left(p_{t}, v\right) .
$$

3: $\quad$ Sample decision $\pi_{t} \sim p_{t}$ and observe $o_{t} \sim M_{t}\left(\pi_{t}\right)$.
4: $\quad$ Update $p_{t+1}=\left(v_{t}\right)_{\pi^{*} \mid \pi_{t}, o_{t}}$.

At round $t$, APS inputs $p_{t}$ to the objective $\operatorname{AIR}_{p_{t}, \eta}\left(p_{t}, v\right)$ to optimize for the algorithmic belief $v_{t}$, and it sets $p_{t+1}$ to be the Bayesian posterior obtained from belief $v_{t}$ and observations $\pi_{t}, o_{t}$. Unlike traditional TS, APS does not require knowing the prior or stochastic environment; instead, APS creates algorithmic beliefs "on the fly" to simulate the worst-case environment. We can prove
the following theorem using the regret bound (3.3.1) in Theorem 14 and the relationship between AIR and IR established in Lemma 8.

Theorem 15 (Regret of APS) Assume that $f_{M}(\pi) \in[0,1]$ for all $M \in \mathcal{M}$ and $\pi \in \Pi$. The regret of Algorithm 6 with $\eta=\sqrt{2 \log |\Pi| /\left(\mathrm{IR}_{\mathrm{H}}(\mathrm{TS}) \cdot T+4 T\right)}$ is bounded as follows, for all $T \geq$ $2 \log |\Pi| \mathrm{IR}_{\mathrm{H}}(\mathrm{TS})+4$,

$$
\mathfrak{R}_{T} \leq \sqrt{\log |\Pi|\left(\mathrm{IR}_{\mathrm{H}}(\mathrm{TS}) / 2+2\right) T}
$$

where $\operatorname{IR}_{\mathrm{H}}(\mathrm{TS}):=\sup _{v} \operatorname{IR}_{\mathrm{H}}\left(v, v_{\pi^{*}}\right)^{1}$ is the maximal value of information ratio for Thompson Sampling.

For $K$-armed bandits, APS achieves the near-optimal regret $O(\sqrt{K T \log K})$ because $\mathrm{IR}_{\mathrm{H}}(\mathrm{TS}) \leq$ $K$; for $d$-dimensional linear bandits, APS recovers the optimal regret $O\left(\sqrt{d^{2} T}\right)$ because $I \mathrm{R}_{\mathrm{H}}(\mathrm{TS}) \leq$ d.

The main messages about APS and Theorem 15 are: 1) the regret bound of APS is no worse than the standard regret bound of TS [104], but in contrast to the latter, does not rely on any knowledge needed to specify a prior! 2) Because APS only keeps the marginal beliefs of $\pi^{*}$ but forgets beliefs of the models, it is robust to adversarial and non-stationary environments. And 3) Experimental results in Section 3.4 show that APS achieves "best-of-all-worlds" empirical performance for Bernoulli MAB in different environments.

To the best of our knowledge, Theorem 15 is the first generic result to make TS prior-free and applicable to adversarial environment. To that end, we note that Corollary 19 in [107] only applies to $K$-armed bandits because of their truncation procedure.

[^4]
### 3.3.3 Adaptive Minimax Sampling (AMS)

When the agent selects decision $p_{t}$ by solving the minimax problem

$$
\inf _{p_{t}} \sup _{v} \operatorname{AR}_{q_{t}, \eta}(p, v),
$$

and optimizes for algorithmic beliefs as in Algorithm 5, we call the resulting algorithm "Adaptive Minimax Sampling" (AMS). By the regret bound (3.3.1) in Theorem 14 and the relationship

```
Algorithm 7 Adaptive Minimax Sampling (AMS)
Input learning rate \(\eta>0\).
Initialize \(q_{1}=\operatorname{Unif}(\Pi)\).
```

    for round \(t=1,2, \cdots, T\) do
    2: Find a distribution \(p\) of \(\pi\) and a distribution \(v_{t}\) of \(\left(M, \pi^{*}\right)\) that solves the saddle point of
                        \(\inf _{p \in \Delta(\Pi)} \sup _{v \in \Delta(\mathcal{M} \times \Pi)}\) AIR \(_{q_{t}, \eta}(p, v)\).
    3: \(\quad\) Sample decision \(\pi_{t} \sim p_{t}\) and observe \(o_{t} \sim M_{t}\left(\pi_{t}\right)\).
    4: Update $q_{t+1}=\left(v_{t}\right)_{\pi^{*} \mid \pi_{t}, o_{t}}$.
between AIR and IR established in Lemma 8, it is straightforward to prove the following theorem.

Theorem 16 (Regret of AMS) For a finite decision space $\Pi$ and a compact model class $\mathcal{M}$, the regret of Algorithm 7 with $\eta=2 \sqrt{\log |\Pi| /(\operatorname{IR(IDS)} \cdot T)}$ is always bounded by

$$
\mathfrak{R}_{T} \leq \frac{1}{2} \sqrt{\log |\Pi| \cdot \operatorname{IR}(\mathrm{IDS}) \cdot T}
$$

where $\operatorname{IR}(\operatorname{IDS}):=\sup _{v} \inf _{p} \operatorname{IR}(v, p)$ is the maximal information ratio of Information-Directed Sampling.

Theorem 16 shows that the regret bound of AMS is always no worse than that of IDS [106]. By showing implicit equivalence and making clean-ups, Algorithm 7 can also be explained as a much simplified implementation of the key ideas in the EBO algorithm from [107], but AMS runs in
computationally tractable spaces (rather than intractable functional spaces) and does not require unnecessary truncation.

### 3.3.4 Using approximate maximizers

In Algorithm 5, we ask for the algorithmic beliefs to maximize AIR. In order to give computationally efficient algorithms in practical applications (MAB, linear bandits, RL, ...), we will require the algorithmic beliefs to approximately maximize AIR. This argument is made rigorous in the following theorem, which uses the first-order optimization error of AIR to represent the regret bound.

Theorem 17 (Generic regret bound using approximate maximizers) Given a finite $\Pi$, a compact $\mathcal{M}$, an arbitrary algorithm ALG that produces decision probability $p_{1}, \ldots, p_{T}$, and a sequence of beliefs $v_{1}, \ldots, v_{T}$ where $q_{t}=\left(v_{t-1}\right)_{\pi^{*} \mid \pi, o} \in \operatorname{int}(\Delta(\Pi))$ for all rounds, we have

$$
\begin{array}{r}
\mathfrak{R}_{T} \leq \frac{\log |\Pi|}{\eta}+\sum_{t=1}^{T}\left(\operatorname{AIR}_{q_{t}, \eta}\left(p_{t}, v_{t}\right)\right. \\
\left.+\sup _{v^{*}}\left(\left.\frac{\partial \operatorname{AIR}_{q_{t}, \eta}\left(p_{t}, v\right)}{\partial v}\right|_{v=v_{t}}\right)^{\top}\left(v^{*}-v_{t}\right)\right) .
\end{array}
$$

Thus we give a concrete approach towards computationally efficient algorithms with rigorous guarantees-making the gradient of AIR small to approximately maximize AIR.

### 3.4 Application to Bernoulli MAB

Our Bayesian design principles give rise to a novel algorithm for the Bernoulli multi-armed bandits (MAB) problem. It is well-known that every bounded-reward MAB problem can equivalently be reduced to the Bernoulli MAB problem, so our algorithm and experimental results actually apply to all bounded-reward MAB problems. The reduction is very simple: assuming the rewards are always bounded in $[a, b]$, then after receiving $r_{t}$ at each round, the agent re-samples a binary reward $\tilde{r}_{t} \sim \operatorname{Bern}\left(\left(r_{t}-a\right) / b-a\right)$ so that $\tilde{r}_{t} \in\{0,1\}$.

### 3.4.1 Simplified APS for Bernoulli MAB

In Example $9, \Pi=[K]=\{1, \cdots, K\}$ is a set of $K$ actions, and each model $\mathcal{M}$ is a mapping from actions to Bernoulli distributions. Given belief $v \in \Delta(\mathcal{M} \times[K])$, we introduce the following parameterization: $\forall i, j \in[K]$,

$$
\begin{aligned}
\theta_{i}(j) & :=\mathbb{E}\left[r(j) \mid \pi^{*}=i\right], & \text { (conditional mean reward) } \\
\alpha(i) & :=v_{\pi^{*}}(i), & \text { (marginal belief) } \\
\beta_{i}(j) & :=\alpha(i) \cdot \theta_{i}(j) . & \text { (guarantees concavity) }
\end{aligned}
$$

Then we have a concave parameterization of AIR by the $K(K+1)$-dimensional vector $(\alpha, \beta)=$ $\left(\alpha, \beta_{1}, \cdots, \beta_{K}\right):$

$$
\begin{array}{r}
\operatorname{AIR}_{q, \eta}(p, v)=\sum_{i \in[K]} \beta_{i}(i)-\sum_{i, j \in[K]} p(j) \beta_{i}(j) \\
-\frac{1}{\eta} \sum_{i, j \in[K]} p(j) \alpha(i) \mathrm{kl}\left(\frac{\beta_{i}(j)}{\alpha(j)}, \sum_{i \in[K]} \beta_{i}(j)\right)-\frac{1}{\eta} \mathrm{KL}(\alpha, q),
\end{array}
$$

where $\mathrm{kl}(x, y):=x \log \frac{x}{y}+(1-x) \log \frac{1-x}{1-y}$ for all $x, y \in(0,1)$. By setting the gradients of AIR with respect to all $K^{2}$ coordinates in $\beta$ to be exactly zero, and choosing $\alpha=p$ (which results in the gradient of AIR with respect to $\alpha$ being suitbly bounded), we are able to write down a simplified APS algorithm in closed form (see Algorithm 8). We apply Theorem 17 to show that the algorithm achieves near-optimal $O(\sqrt{K T \log K})$ regret in the general adversarial setting. We leave the detailed derivation and analysis of the Algorithm 8 to Appendix C.2.2.

At each round, Algorithm 8 increases the weight of the selected action $\pi_{t}$ if $r_{t}=1$, and decreases the weight if $r_{t}=0$. The algorithm also maintains the "relative weight" between all unchosen actions $\pi \neq \pi_{t}$, allocating probabilities to these actions proportionally to $p_{t}$. Algorithm 8 is clearly very different from the well-known EXP3 algorithm, which instead updates $p_{t+1}$ by the

```
Algorithm 8 Simplified APS for Bernoulli MAB
Input learning rate \(\eta>0\).
Initialize \(p_{1}=\operatorname{Unif}(\Pi)\).
    for round \(t=1,2, \cdots, T\) do
        Sample action \(\pi_{t} \sim p_{t}\) and receives \(r_{t}\).
        Update \(p_{t+1}\) by
\[
\begin{aligned}
p_{t+1}\left(\pi_{t}\right) & =\left\{\begin{array}{ll}
\frac{1-\exp (-\eta)}{1-\exp \left(-\eta / p_{t}\left(\pi_{t}\right)\right)}, & \text { if } r_{t}=1 \\
\frac{1-\exp (\eta)}{1-\exp \left(\eta / p_{t}\left(\pi_{t}\right)\right)}, & \text { if } r_{t}=0
\end{array},\right. \text { and } \\
p_{t+1}(\pi) & =p_{t}(\pi) \cdot \frac{1-p_{t+1}\left(\pi_{t}\right)}{1-p_{t}\left(\pi_{t}\right)}, \quad \forall \pi \neq \pi_{t} .
\end{aligned}
\]
```

formula

$$
p_{t+1}(\pi)=p_{t}(\pi) \exp \left(\eta \cdot \frac{r_{t} \mathbb{1}\left\{\pi=\pi_{t}\right\}}{p_{t}\left(\pi_{t}\right)}\right), \quad \forall \pi \in \Pi .
$$

In Section 3.6.2 we recover a modified version of EXP3 by Bayesian principle assuming Gaussian reward. We conclude that Algorithm 6 uses a precise posterior for Bernoulli reward, while EXP3 estimates worst-case Gaussian reward. This may explain why Algorithm 8 performs much better in all of our experiments.

### 3.4.2 Numerical experiments

We implement Algorithm 8 (with the legend "APS" in the figures) in the stochastic, adversarial and non-stationary environments. We plot expected regret (average of 100 runs) for different choices of $\eta$, and set $\gamma=0.001$ in all experiments. We find APS 1) outperforms UCB and matches TS in the stochastic environment; 2) outperforms EXP3 in the adversarial environment; and 3) outperforms EXP3 and is comparable to the "clairvoyant" benchmarks (that have prior knowledge of the changes) in the non-stationary environment. For this reason we say Algorithm 8 (APS) achieves the "best-of-all-worlds" performance. We note that the optimized choice of $\eta$ in APS differ instance by instance, but by an initial tuning we typically see good results, whether we tune $\eta$ optimally or not optimally.


Figure 3.1: Sensitivity analysis in a stochastic bandit problem.

## Stochastic Bernoulli MAB

In Figure 3.1 we report the expected regret for APS with different choices of $\eta$, TS with different Beta priors, and the UCB 1 algorithm, in a stochastic 16-armed Bernoulli bandit problem. We refer to this as "sensitivity analysis" because the red, semi-transparent, area reports the regret of APS when learning rates $\eta$ are chosen across a range of values drawn from the interval $[0.05,0.5]$ (the interval is specified by an initial tuning); and the priors of TS are chosen from $\operatorname{Beta}(c, 1)$ where $c \in[0.5,5]$. In particular, the bottom curve of the red (or blue) area is the regret curve of APS (or TS) using optimally tuned $\eta$ (respectively, prior). The conclusion is that APS outperforms UCB 1 , and is comparable to TS in this stochastic environment.

## Adversarial Bernoulli MAB

We equidistantly take 16 horizontal lines from an abstract art piece by Jackson Pollock to simulate the rewards (pre-specified) in an adversarial environment, and study this via a 16-armed bandit problem. Figure 3.2 shows the sensitivity analysis for APS and EXP3 when both the learning rates are chosen from $[0.1,5]$ (the interval is specified by an initial tuning). In particular, the red and


Figure 3.2: Sensitivity analysis in an adversarial bandit problem.
green lower curves compare the optimally tuned versions of APS and EXP3. The conclusion is that APS outperforms EXP3 whether $\eta$ is tuned optimally or not.

## Non-stationary Bernoulli MAB (with change points)

We study a 16-armed Bernoulli bandit problem in a non-stationary environment. We generate 4 batches of i.i.d. sequences, where the changes in the environment occur after round 1000 , round 2000, and round 3000. We consider a stronger notion of regret known as the dynamic regret [114], which compares the cumulative reward of an algorithm to the cumulative reward of the best nonstationary policy (rather than a single arm) in hindsight. In this particular setting, the benchmark is to select the best arm in all the 4 batches. In Figure 3.3 we perform sensitivity analysis for APS and EXP3, where the learning rates are chosen across $[0,05,5]$. Since the agent will not know when and how the adversarial environment changes in general, it is most reasonable to compare APS with EXP3 without any knowledge of the environment as in Figure 3.3. We observe that APS dramatically improves the dynamic regret by several times.

In Figure 3.4, we compare APS to three "clairvoyant" restarted algorithms, which require


Figure 3.3: Sensitivity analysis in a "change points" environment.


Figure 3.4: Comparing APS to "clairvoyant" restarted algorithms in a "change points" environment.
knowing that the environment consists of 4 batches of i.i.d. sequences, as well as knowing the exact change points. We tune the parameters in these algorithms optimally. Without knowledge of the environment, APS performs better than restarted EXP3 and restarted UCB 1, and is comparable to restarted TS. (It is important to emphasize again that the latter algorithms are restarted based on foreknowledge of the change points.)

## Non-stationary Bernoulli MAB (with "sine curve" reward sequences)

We generate a 4-armed bandit problem with the mean-reward structure shown in Figure 3.5. The four sine curves (with different colors) in Figure 3.5 represent the mean reward sequences of the 4 arms. We tune the parameters in all the algorithms to optimal and report their regret curves in Figure 3.6. As shown in Figure 3.6, APS achieves the best performance, while TS fails in this non-stationary environment. This experiment shows the vulnerability of TS if the environment is not stationary, such as the sine curve structure shown here.

To better illustrate the smartness of APS compared with TS in the non-stationary environment, we track the selected arms and the best arms throughout the process. In Figure 3.7 and Figure 3.8, the horizontal line represents the 4000 rounds, and the vertical lines represent the 4 arms (indexed as 1, 2, 3, and 4). In Figure 3.7, the red points show the selected arms of APS, and the black


Figure 3.5: "Sine curve" reward sequences for 4 arms.

Figure 3.6: Regret curves in a "sine curve" environment.


Figure 3.7: Tracking selected arms of APS in a "sine curve" environment.

Figure 3.8: Tracking selected arms of TS in a "sine curve" environment.
points represent the best arms at each round in this "sine curve" non-stationary environment. In Figure 3.8, the blue points show the selected arms of TS. The more consistent the selected arms are with the best arms (black points), the better choices an algorithm makes. Comparing Figure 3.7 and Figure 3.8, we can see that APS is highly responsive to changes in the best arm, whereas TS is relatively sluggish in this regard. The implication of this experiment is that creating a new algorithmic belief at each round has the potential to significantly improve performance and be a game changer in many problem settings.

These experiments provide some numerical evidence indicating that APS achieves the "best-of-all-worlds" across stochastic, adversarial, and non-stationary environments.

### 3.5 Key intuitions of the proof

It is worth noting that the proof of Theorem 14 is quite insightful and parsimonious. The two major steps in the proof may be interesting on its own. The first step is a succinct analysis to bound the cumulative regret by sum of AIR (see Section 3.5.1); and the second step is to extend the classical minimax theory of "exchanging values" into a constructive approach to design minimax decisions (estimators, algorithms, etc.), which will be presented in Section 3.5.2.

### 3.5.1 Bounding regret by sum of AIR

For every $\bar{\pi} \in \Pi$, we have

$$
\begin{equation*}
\sum_{t=1}^{T}\left[\log \frac{q_{t+1}(\bar{\pi})}{q_{t}(\bar{\pi})}\right]=\log \frac{q_{T}(\bar{\pi})}{q_{1}(\bar{\pi})} \leq \log |\Pi| . \tag{3.5.1}
\end{equation*}
$$

Taking $q_{t+1}=\left(v_{t}\right)_{\pi^{*} \mid \pi_{t}, o_{t}}$ as in Algorithm 5, and taking expectation on the left hand side of (3.5.1), we have

$$
\begin{equation*}
\sum_{t=1}^{T} \mathbb{E}_{\pi_{t}, o_{t}}\left[\log \frac{\left(v_{t}\right)_{\pi^{*}}\left(\bar{\pi} \mid \pi_{t}, o_{t}\right)}{q_{t}(\bar{\pi})}\right] \leq \log |\Pi| \tag{3.5.2}
\end{equation*}
$$

By subtracting the additive elements on the left-hand side of (3.5.2) (divided by $\eta$ ) from the perround regrets against $\bar{\pi}$, we obtain

$$
\begin{align*}
& \mathbb{E}\left[\sum_{t=1}^{T} f_{M_{t}}(\bar{\pi})-\sum_{t=1}^{T} f_{M_{t}}\left(\pi_{t}\right)\right]-\frac{1}{\eta} \cdot \sum_{t=1}^{T} \mathbb{E}_{\pi_{t}, o_{t}}\left[\log \frac{\left(v_{t}\right)_{\pi^{*}}\left(\bar{\pi} \mid \pi_{t}, o_{t}\right)}{q_{t}(\bar{\pi})}\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[f_{M_{t}}(\bar{\pi})-f_{M_{t}}\left(\pi_{t}\right)-\frac{1}{\eta} \log \frac{\left(v_{t}\right)_{\pi^{*} \mid \pi_{t}, o_{t}}(\bar{\pi})}{q_{t}(\bar{\pi})}\right] \\
& \leq \sum_{t=1}^{T} \sup _{M, \bar{\pi}} \mathbb{E}\left[f_{M}(\bar{\pi})-f_{M}\left(\pi_{t}\right)-\frac{1}{\eta} \log \frac{\left(v_{t}\right)_{\pi^{*} \mid \pi_{t}, o_{t}}(\bar{\pi})}{q_{t}\left(\pi^{*}\right)}\right] \\
& \stackrel{(*)}{=} \sum_{t=1}^{T} \operatorname{AIR}_{q_{t}, \eta}\left(p_{t}, v_{t}\right) \tag{3.5.3}
\end{align*}
$$

where the inequality is by taking supremum at each rounds; and the last equality $(*)$ in (3.5.3) is by Lemma 11, an important identity to be explained in Section 3.5.2, which is derived from the fact that the pair of maximizer $v_{t}$ and posterior functional is a Nash equilibrium of a convex-concave function.

Combining (3.5.3) and (3.5.2), we obtain the following inequality for every $\bar{\pi} \in \Pi$ :

$$
\mathbb{E}\left[\sum_{t=1}^{T} f_{M_{t}}(\bar{\pi})-\sum_{t=1}^{T} f_{M_{t}}\left(\pi_{t}\right)\right]-\frac{\log |\Pi|}{\eta} \leq \sum_{t=1}^{T} \operatorname{AIR}_{q_{t}, \eta}\left(p_{t}, v_{t}\right)
$$

By taking the supremum over $\bar{\pi} \in \Pi$ on the left-hand side of the inequality, we are able to prove Theorem 14:

$$
\mathfrak{R}_{T} \leq \frac{\log |\Pi|}{\eta}+\sum_{t=1}^{T} \operatorname{AIR}_{q_{t}, \eta}\left(p_{t}, v_{t}\right)
$$

3.5.2 Mimimax theory: from value to construction

Consider a decision space $\mathcal{X}$, a space $\mathcal{Y}$ of the adversary's outcome, and a convex-concave function $\psi(x, y)$ defined in $\mathcal{X} \times \mathcal{Y}$. The classical minimax theorem [115] says that, under regularity
conditions, the minimax and maximin values of $\psi(x, y)$ are equal:

$$
\min _{\mathcal{X}} \max _{y} \psi(x, y)=\max _{y} \min _{\mathcal{X}} \psi(x, y) .
$$

We refer to $\arg \min _{\mathcal{X}} \max y \psi(x, y)$ as the set of "minimax decisions," as they are optimal in the worst-case scenario. And we say $\tilde{x} \in \arg _{\min }^{\mathcal{X}} \psi(x, \bar{y})$ is "maximin decision" if $\bar{y} \in \arg \max _{y} \min _{\mathcal{X}} \psi(x, y)$ is "maximin adversary's outcome." One natural and important question is, when will the "maximin decision" $\tilde{x}$ also be a "minimax decision?" The study to this question may provide a constructive way to design frequentist estimators and algorithms through worst-case Bayesian posteriors and regularization. Making use of strong convexity, we extends the classical minimax theorem for values into the following minimax theorem for decisions:

Lemma 10 (Constructing minimax decisions) Let $\mathcal{X}$ and $\mathcal{Y}$ be convex and compact sets, and $\psi: X \times Y \rightarrow \mathbb{R}$ a function which for all $y$ is strongly convex and continuous in $x$ and for all $x$ is concave and continuous in $y$. For each $y \in \mathcal{Y}$, let $x_{y}=\min _{x \in X} \psi(x, y)$ be the corresponding unique minimizer. Then by maximizing the concave objective

$$
\bar{y} \in \max _{y \in \mathcal{Y}} \psi\left(x_{y}, y\right)
$$

the pair $\left(x_{\bar{y}}, \bar{y}\right)$ will be a Nash equilibrium that solves the minimax optimization problem $\min _{\mathcal{X}} \max y \psi(x, y)$.

Applying Lemma 10 to our framework, we can show: 1) Bayesian posterior $v_{\pi^{*} \mid \pi, o}$ is the optimal functional to make decision under belief $v$; and 2) by choosing worst-case belief $\bar{v}$, we construct a Nash equilibrium. As a result, we can establish the following per-round identity, which leads to the key identity denoted as $(*)$ in (3.5.3). This identity plays a crucial role in the proof of Theorem 14.

Lemma 11 (Identity by Nash equilibrium) Given $q \in \operatorname{int}(\Delta(\Pi)), \eta>0$ and $p \in \Delta(\Pi)$, denote
$\bar{v} \in \arg \max \operatorname{AIR}_{q, \eta}(p, v)$. Then we have

$$
\sup _{M, \bar{\pi}} \mathbb{E}\left[f_{M}(\bar{\pi})-f_{M}(\pi)-\frac{1}{\eta} \log \frac{\bar{\nu}_{\pi^{*} \mid \pi, o}(\bar{\pi})}{q(\bar{\pi})}\right]=\operatorname{AIR}_{q, \eta}(p, \bar{v}) .
$$

### 3.6 Applications to infinite-armed bandits

Our design principles can be applied in many sequential learning and decision making environments. In order to maximize AIR in practical applications, we parameterize the belief $v$, and make the gradient of AIR with respect to such parameter small. Going beyond multi-armed bandits (MAB), we often need to constrain the search of algorithmic belief within a tractable subspace; and we study useful concave relaxations of AIR towards efficient algorithm design. We will present our results for linear bandits and bandit convex optimization in this section and present our results for reinforcement learning in Section 3.7. We give a high-level overview of the applications to linear bandits and bandit convex optimization here.

Application to linear bandits. A classical algorithm for adversarial linear bandits (described in Example 10) is the EXP2 algorithm [116], which uses IPW for linear loss as a black-box estimation method, and combines it with continuous exponential weight. We derive a modified version of EXP2 from our framework, establishing interesting connection between IPW and Bayesian posteriors.

Application to bandit convex optimization. Bandit convex optimization (described in Example 10) is a notoriously challenging problem, and much effort has been put to understanding its minimax regret and algorithm design. The best known result, which is of order $\tilde{O}\left(d^{2.5} \sqrt{T}\right)$, is derived through the non-constructive information-ratio analysis in [117]. As a corollary of Theorem 16, Adaptive Minimax Sampling (AMS) recovers the best known regret bound with a constructive algorithm, which can be computed in poly $\left(e^{d} \cdot T\right)$ time. To the best of our knowledge, this is the first finite-running-time algorithm that attains the best known $\tilde{O}\left(d^{2.5} \sqrt{T}\right)$ regret.

### 3.6.1 Maximization of AIR for structured bandits

Consider the structured bandit problems described in Example 10. We consider the computation complexity of the optimization problem

$$
\begin{equation*}
\sup _{v \in \Delta(\mathcal{M} \times \Pi)} \operatorname{AIR}_{q, \eta}(p, v) . \tag{3.6.1}
\end{equation*}
$$

The computational complexity of (3.6.1) may be $O(\operatorname{poly}(\exp (\exp (d))))$ in the worst case as the size of $\mathcal{M} \times \Pi$. However, when the mean reward function class $\mathcal{F}$ is a convex function class, the computational complexity will be $O(\operatorname{poly}(|\Pi|))$ which is efficient for $K$-armed bandits and is no more than $O\left(\operatorname{poly}\left(e^{d}\right)\right)$ in general (by standard discretization and covering arguments, we may assume $\Pi \subset \mathbb{R}^{d}$ to have finite cardinality $O\left(e^{d}\right)$ for the simplicity of theoretical analysis). Moreover, we also give efficient algorithm for linear bandits with exponential-many actions. We refer to Appendix C.2.1 for the detailed discussion on the parameterization method and computational complexity.

### 3.6.2 Application to Gaussian linear bandits

We consider the adversarial linear bandit problem with Gaussian reward. In such a MAB problem, $\Pi=\mathcal{A} \subseteq \mathbb{R}^{d}$ is an action set with dimension $d$. The model class $\mathcal{M}$ can be parameterized by a $d$-dimensional vector $\theta \in \mathbb{R}^{d}$ that satisfies $\theta^{\top} a \in[0,1]$ for all $a \in \mathcal{A}$. The reward $r(a)$ for each action $a \in \mathcal{A}$ is independently drawn from a Gaussian distribution that has mean $\theta^{\top} a$ and variance $\sigma^{2}$, and we assume that $\sigma \leq 1$. Here we use the notations $\mathcal{A}$ (as action set), $a$ (as action) and $a^{*}$ (as optimal action) to follow the tradition of literature about linear bandits.

As discussed in Section 3.6.1, we restrict our attention to sparse $v$ where for each $\pi^{*} \in \Pi$ there is only one model $M$, which corresponds to the Gaussian distribution $r(\pi) \sim N\left(\theta_{\pi^{*}}(\pi), 1\right)$. We parameterize the prior $v$ by vectors $\left\{\beta_{a}^{*}\right\}_{a^{*} \in \mathcal{A}}$ and $\alpha \in \Delta(\Pi)$, where $\alpha=\mathbb{P}_{v}\left(a^{*}\right)$ and $\beta_{a^{*}}=$ $\alpha\left(a^{*}\right) \cdot \theta_{a^{*}}$. As discussed in (C.2.7) in Appendix C.2.3, we propose to define a surrogate version
of AIR by

$$
\begin{align*}
\overline{\operatorname{AIR}}_{q, \eta}(p, v)= & \int_{\mathcal{A}} \beta_{a^{*}}^{\top} a^{*} d a^{*}-\int_{\mathcal{A}} \int_{\mathcal{A}} p(a) \beta_{a^{*}}^{\top} a d a^{*} d a \\
& -\frac{1}{2 \eta} \int_{\mathcal{A}} \int_{\mathcal{A}} p(a) \alpha\left(a^{*}\right)\left(\frac{\beta_{a^{*}}^{\top} a}{\alpha\left(a^{*}\right)}-\int_{\mathcal{A}} \beta_{a^{*}}^{\top} a d a^{*}\right)^{2} d a-\frac{1}{\eta} \operatorname{KL}(\alpha, q) . \tag{3.6.2}
\end{align*}
$$

As discussed in Section 3.6.1, It can be shown that approximate maximizers of this surrogate lead to rigorous regret bounds. Note that the surrogate defined in (3.6.2) can be bounded by the classical information ratio bounds defined by square loss (see, e.g., [104, 117]).

By making all the gradients of (3.6.2) with respect to $\left\{\beta_{a^{*}}\right\}_{a^{*} \in \mathcal{A}}$ to be exactly zero, and taking $\alpha=p$, we obtain an approximate maximizer of AIR in (3.6.2). We calculate the Bayesian posterior, and find that the resulting algorithm is an exponential weight algorithm with a modified IPW estimator: at each round $t$, the agent update $p_{t+1}$ by

$$
\tilde{p}_{t+1}(a) \propto p_{t}(a) \exp \left(\eta \hat{r}_{t}(a)\right),
$$

where $\hat{r}_{t}$ is the modified IPW estimator for linear loss,

$$
\begin{equation*}
\hat{r}_{t}(a)=a^{\top}\left(\mathbb{E}_{a \sim p_{t}}\left[a a^{\top}\right]\right)^{-1} a_{t} r_{t}\left(a_{t}\right)-\frac{\eta}{2}\left(a^{\top}\left(\mathbb{E}_{a \sim p_{t}}\left[a a^{\top}\right]\right)^{-1} a_{t}\right)^{2} . \tag{3.6.3}
\end{equation*}
$$

Note that in order to avoid boundary conditions in our derivation, we require forced exploration to ensure $\lambda_{\min }\left(\mathbb{E}_{a \sim p}\left[a a^{\top}\right]\right) \geq \eta$. This can be done with the help of the volumetric spanners constructed in [118]. The use of volumetric spanner makes our final proposed algorithm (Algorithm 9) to be slightly more involved, but we only use the volumetric spanner in a "black-box" manner. We highlight that the algorithm is computationally efficient as the reward estimator (3.6.3) is concave, so one can apply log-concave sampling when executing exponential weighting. The additional term in (3.6.3) is ignorable from a regret analysis perspective, so the standard analysis for exponential weight algorithms applies to Algorithm 9 to establish the optimal $O\left(\sqrt{d^{2} T}\right.$ regret bound. One may also analyze Algorithm 9 within our algorithmic belief framework through The-

```
Algorithm 9 Simplified APS for Gaussian linear bandits
Input learning rate \(\eta>0\), forced exploration rate \(\gamma\), and action set \(\mathcal{A}\).
Initialize \(p_{1}=\operatorname{Unif}(\Pi)\).
    for round \(t=1,2, \cdots, T\) do
        Let \(S_{t}^{\prime}\) be a \(\left(p_{t}, \exp (-(4 \sqrt{d}+\log (2 T)))\right)\)-exp-volumetic spanner of \(\mathcal{A}\),
        Let \(S_{t}^{\prime \prime}\) be a \(2 \sqrt{d}\)-ratio-volumetric spanner of \(\mathcal{A}\).
        Set \(S_{t}\) as the union of \(S_{t}^{\prime}\) and \(S_{t}^{\prime \prime}\).
3: \(\quad\) Sample action \(a_{t} \sim p_{t}\) and receives \(r_{t}\).
4: Calculate \(\tilde{p}_{t+1}\) by
\[
\tilde{p}_{t+1}(a) \propto p_{t}(a) \exp \left(\eta \hat{r}_{t}(a)\right),
\]
```

where $\hat{r}_{t}$ is the modified IPW estimator for linear loss,

$$
\hat{r}_{t}(a)=a^{\top}\left(\mathbb{E}_{a \sim p_{t}}\left[a a^{\top}\right]\right)^{-1} a_{t} r_{t}\left(a_{t}\right)-\frac{\eta}{2}\left(a^{\top}\left(\mathbb{E}_{a \sim p_{t}}\left[a a^{\top}\right]\right)^{-1} a_{t}\right)^{2} .
$$

5: $\quad$ Update $p_{t+1}$ by $p_{t+1}(a)=(1-\gamma) \tilde{p}_{t}(a)+\frac{\gamma}{\left|S_{t}\right|} \mathbb{1}\left\{a \in S_{t}\right\}$
orem 17, as we did for Algorithm 8 in Section C.2.2; we omit the analysis here. Finally, we note that the algorithm reduces to a modified version of EXP3 for finite armed bandits, a connection we mentioned at the end of Section 3.4.1.

### 3.6.3 Application to bandit convex optimization

We consider the bandit convex optimization problem described in Example 10. In bandit convex optimization, $\Pi \subseteq \mathbb{R}^{d}$ is a $d$-dimensional action set whose diameter is bounded by diam $(\Pi)$, and the mean reward (or loss) function is required to be concave (respectively, convex) with respect to actions:

$$
\mathcal{F}=\{f: \Pi \rightarrow[0,1]: f \text { is concave w.r.t. } \pi \in \Pi\} .
$$

The problem is often formed with finite (but exponentially large) action set by standard discretization arguments [117]. Bandit convex optimization is a notoriously challenging problem, and much effort has been put to understanding its minimax regret and algorithm design. The best known result, which is of order $\tilde{O}\left(d^{2.5} \sqrt{T}\right)$, is derived through the non-constructive information-ratio analy-
sis in [117]. By the information ratio upper bound for the non-constructive Bayesian IDS algorithm in [117], Lemma 8 that bounds AIR by IR, and Theorem 16 (regret of AMS), we immediately have that Algorithm 7 (AMS) with optimally tuned $\eta$ achieves

$$
\mathfrak{R}_{T} \leq O\left(d^{2.5} \sqrt{T} \cdot \operatorname{polylog}(d, \operatorname{diam}(\mathcal{A}), T)\right)
$$

As a result, AMS recovers the best known $\tilde{O}\left(d^{2.5} \sqrt{T}\right)$ regret with a constructive algorithm. By our discussion on the computational complexity in Appendix C.2.1, AMS solves convex optimization in a poly $(|\Pi|)$-dimensional space, so it can be computed in poly $\left(e^{d} \cdot T\right)$ time for bandit convex optimization. To the best of our knowledge, this is the first algorithm with a finite running time that attains the best known $\tilde{O}\left(d^{2.5} \sqrt{T}\right)$ regret. We note that the EBO algorithm in [107] has given a constructive algorithm that achieves the same $\tilde{O}\left(d^{2.5} \sqrt{T}\right)$ regret derived by Bayesian nonconstructive analysis. However, EBO operates in an abstract functional space, so it is less clear how to execute the computation.

### 3.7 Model-index AIR and application to RL

In the stochastic environment, where $M_{t}=M^{*} \in \mathcal{M}$ for all rounds, we want to find the optimal decision $\pi_{M^{*}}$ that minimizes the mean reward function $f_{M^{*}}(\pi)$. Unlike the adversarial setting, where algorithmic beliefs are formed over pairs of models and optimal decisions, in the stochastic setting, we only need to search for algorithmic beliefs regarding the underlying model. This distinction allows us to develop a strengthened version of AIR, which we call "Model-index AIR" (MAIR), particularly suited for studying reinforcement learning problems.

Crucially, we can construct a generic and closed-form sequence of algorithmic beliefs that approximate the maximization of MAIR at each round. By leveraging these beliefs, we develop a model-based APS algorithm that achieves the sharpest known bounds for RL problems within the bilinear class [112, 105]. Our algorithm features a generic and closed-form updating rule, making it potentially well-suited for efficient implementation through efficient sampling oracles.

### 3.7.1 Model-index Algorithmic Information Ratio

We denote decision $\pi_{M} \in \arg \min _{\Pi} f_{M}(\pi)$ be the induced optimal decision of model $M$. We introduce the following definition of Model-index AIR (MAIR).

Definition 10 (Model-index AIR) Denote $\rho \in \operatorname{int}(\Delta(\mathcal{M})$ ) be a reference distribution of models, and $\mu \in \operatorname{int}(\Delta(\mathcal{M}))$ be a prior belief of models, we define the "Model-index Algorithmic Information Ratio" as

$$
\operatorname{MAIR}_{\rho, \eta}(p, \mu)=\mathbb{E}_{\mu, p}\left[f_{M}\left(\pi_{M}\right)-f_{M}(\pi)-\frac{1}{\eta} \operatorname{KL}(\mu(\cdot \mid \pi, o), \rho)\right],
$$

where $\mu(\cdot \mid \pi, o)$ is the Bayesian posterior belief of models induced by the prior belief $\mu$.

It can be seen from the definition that KL divergence between two model distributions will be no smaller than KL divergence between the two induced decision distributions. Thus we have the following Lemma.

Lemma 12 (MAIR smaller than AIR) When $q$ is the decision distribution of $\pi_{M}$ induced by the model distribution $\rho$, and $v$ is the distribution of $\left(M, \pi_{M}\right)$ induced by the model distribution $\mu$, we have

$$
\operatorname{MAIR}_{\rho, \eta}(p, \mu) \leq \operatorname{AIR}_{q, \eta}(p, v)
$$

Lemma 9 has shown that the worst-case value of AIR under the "maximin" strategy is smaller than DEC of the convex hull of $\mathcal{M}$. Now we demonstrate that the worst-case value of MAIR under a "maximin" strategy is smaller than the worst-case value of DEC, which does not uses the convex hull of model class in its first argument.

Lemma 13 (Bounding MAIR by DEC) Given model class $\mathcal{M}$ and $\eta>0$, we have

$$
\sup _{\rho \in \operatorname{int}(\Delta(\mathcal{M}))} \sup _{\mu} \inf _{p} \operatorname{MAIR} R_{\rho, \eta}(p, v) \leq \sup _{\bar{M} \in \operatorname{conv}(\mathcal{M})} \operatorname{DEC}_{\eta}(\mathcal{M}, \bar{M}) .
$$

Moreover, when the reference distribution $\rho$ is centered at $M^{*}$ and has "small" variance, we may completely removes the convex hull in the expression of DEC (unlike Lemma 13 still leaving a convex hull restriction in the subscribe). This enable us to match the tightest possible version of DEC, and is discussed in Section 3.7.3.

Comparing AIR and MAIR. We have seen that 1) Maximin AIR can be bounded by DEC of the convex hull $\Delta(\mathcal{M}) ; 2$ ) Maximin MAIR can be bounded by DEC of the original class $\mathcal{M}$; and 3) MAIR is "smaller" than AIR as illustrated in Lemma 12. However, as we will later show in Theorem 14 and Theorem 18, the regret bound using AIR will scale with a $\log |\Pi|$ factor (estimation complexity of decision space), while the regret bound using MAIR will scale with a bigger $\log |\mathcal{M}|$ factor (estimation complexity of model class). We explain their difference as follows.

When to use AIR versus MAIR? First, AIR is useful for both stochastic and adversarial bandit learning problems, while MAIR may only be useful for stochastic environments. Second, using AIR will result in a $\log |\Pi|$ factor along with information ratio (or DEC), while MAIR will result in a bigger $\log |\mathcal{M}|$ factor, so AIR is often the tighter option for bandit problems. For example, AIR provides optimal regret for $K$-armed bandits and $\sqrt{T}$-type regret bound for the challenging problem bandit convex optimization, while MAIR may not. On the other hand, MAIR can achieve optimal regret for stochastic linear bandits and stochastic model-based contextual bandits [78], and it is more useful than AIR for reinforcement learning problems where taking convex hull to the model class may greatly increase the richness of model class. For example, the model class (especially the state transition dynamic) in reinforcement learning problems may not satisfy convexity. In general, AIR is more useful for "infinite divisible" problems where taking convex hull does not greatly increase the complexity of model class; while MAIR is more useful for stochastic model-based bandit and reinforcement learning problems where one wants to avoid taking convex hull.
3.7.2 Near-optimal algorithmic beliefs in closed form

For any fixed decision probability $p$, it is illustrative to write MAIR as

$$
\begin{align*}
\operatorname{MAIR}_{\rho, \eta}(p, \mu) & =\mathbb{E}\left[f_{M}\left(\pi_{M}\right)-f_{M}(\pi)-\frac{1}{\eta} \operatorname{KL}(\mu(M \mid \pi, o), \rho)\right] \\
& =\mathbb{E}\left[f_{M}\left(\pi_{M}\right)-f_{M}(\pi)-\frac{1}{\eta} \operatorname{KL}(\mu(M \mid \pi, o), \mu)-\frac{1}{\eta} \operatorname{KL}(\mu, \rho)\right] \\
& =\mathbb{E}\left[f_{M}\left(\pi_{M}\right)-f_{M}(\pi)-\frac{1}{\eta} \operatorname{KL}\left(M(\pi), \mu_{o \mid \pi}\right)-\frac{1}{\eta} \operatorname{KL}(\mu, \rho)\right], \tag{3.7.1}
\end{align*}
$$

where $\mu_{o \mid \pi}=\mathbb{E}_{M \sim \mu}[M(\pi)]$ is the induced distribution of $o$ conditioned on $\pi$, and the third equality is by property of mutual information. We would like to give a sequence of algorithmic beliefs that approximately maximize MAIR at each rounds, as well as have closed-form expression.

We consider the following algorithmic priors at each round:

$$
\mu_{t}(M) \propto \rho_{t}(M) \cdot \exp \left(\eta\left(f_{M}\left(\pi_{M}\right)-\mathbb{E}_{p_{t}}\left[f_{M}(\pi)\right]\right)\right)
$$

and use their corresponding posteriors to update the sequence of reference probabilities:

$$
\mu_{t+1}=\mu_{t}\left(M \mid \pi_{t}, o_{t}\right) \propto \mu_{t}(M)\left[M\left(\pi_{t}\right)\right]\left(o_{t}\right) .
$$

This results in the following update of $\rho$ :

$$
\begin{equation*}
\rho_{t+1}(M)=\exp (\sum_{s=1}^{t}(\underbrace{\log \left[M\left(\pi_{s}\right)\right]\left(o_{s}\right)}_{\text {log likelihood }}+\underbrace{\eta\left(f_{M}\left(\pi_{M}\right)-\mathbb{E}_{p_{s}}\left[f_{M}(\pi)\right]\right)}_{\text {adaptive algorithmic belief }})) . \tag{3.7.2}
\end{equation*}
$$

Our algorithm (3.7.2) updates both the log likelihood term and an adaptive algorithmic belief term at each iteration, whereas traditional (fixed-prior) Thompson Sampling only updates the log likelihood term and relies on a fixed prior term.

In Lemma 13, we demonstrate how an upper bound on DEC can automatically translate into

```
Algorithm 10 Model-index AIR generation
Input algorithm ALG and learning rate \(\eta>0\).
Initialize \(\rho_{1}\) to be the uniform distribution over \(\mathcal{M}\).
```

    for round \(t=1,2, \cdots, T\) do
        Obtain \(p_{t}\) from ALG. The algorithm ALG samples \(\pi_{t} \sim p_{t}\) and observe the feedback \(o_{t} \sim\)
        \(M_{t}\left(\pi_{t}\right)\).
    3: Update
    $$
\rho_{t+1}(M) \propto \exp \left(\sum_{s=1}^{t}\left(\log \left[M\left(\pi_{s}\right)\right]\left(o_{s}\right)+\eta\left(f_{M}\left(\pi_{M}\right)-\mathbb{E}_{p_{s}}\left[f_{M}(\pi)\right]\right)\right)\right) .
$$

an upper bound on the maximin value of MAIR. However, the variable $\bar{M}$ in DEC is maximized within the convex hull $\Delta(\mathcal{M})$ rather than $\mathcal{M}$. Therefore, to directly apply the upper bounds on DEC proved in [105], we need to establish a stronger regret bound that completely eliminates convex hull from the expression of DEC. To achieve this, we prove that when the prior distribution $\mu$ is sufficiently "centered," we can bound MAIR using a DEC-type quantity that does not involve taking convex hull at all. Specifically, $\bar{M}$ takes values from $\mathcal{M}$ instead of $\Delta(\mathcal{M})$. We are motivated to prove this result by the fact that the update (3.7.2) will converge to $M^{*}$ over time.

Theorem 18 (Generic regret bound in the stochastic setting) Given a finite model class $\mathcal{M}$ where the underlying true model is $M^{*} \in \mathcal{M}$, and $f_{M}(\pi) \in[0,1]$ for every $M \in \mathcal{M}$ and $\pi \in \Pi$. For an arbitrary algorithm ALG, the regret of algorithm ALG is bounded by
$\mathfrak{R}_{T} \leq \frac{\log (|\mathcal{M}| T)+1}{\eta}+2+\sum_{t=1}^{T} \mathbb{E}_{\mu_{t}, p_{t}}\left[5\left(f_{M}\left(\pi_{M}\right)-f_{M}(\pi)\right)-\frac{1}{\eta} D_{\mathbf{H}}^{2}\left(M(\pi), M^{*}(\pi)\right)-\frac{1}{\eta} \operatorname{KL}\left(\mu_{t}, \rho_{t}\right)\right]$,
where $\mu_{t}(M) \propto \exp \left(\sum_{s=1}^{t}\left(\log \left[M\left(\pi_{s}\right)\right]\left(o_{s}\right)+\eta\left(f_{M}\left(\pi_{M}\right)-\mathbb{E}_{p_{s}}\left[f_{M}(\pi)\right]\right)\right)\right)$.

### 3.7.3 Model-index APS

In our applications, we often use a simple posterior sampling strategy for which we always induce the distribution of optimal decisions from the posterior distribution of models. We refer to the resulting algorithm, Algorithm 11, as "Model-index Adaptive Posterior Sampling."

```
Algorithm 11 Model-index Adaptive Posterior Sampling
Input learning rate \(\eta\) and forced exploration rate \(\gamma\).
Initialize \(\rho_{1}\) to be the uniform distribution over \(\mathcal{M}\).
    for round \(t=1,2, \cdots, T\) do
        Sample \(\pi_{t} \sim p_{t}\) where \(p_{t}(\pi)=\sum_{\pi=\pi_{M}} \rho_{t}(M)\), and observe the feedback \(o_{t} \sim M_{t}\left(\pi_{t}\right)\).
        Update
\[
\rho_{t+1}(M) \propto \exp \left(\sum_{s=1}^{t}\left(\log \left[M\left(\pi_{s}\right)\right]\left(o_{s}\right)+\eta\left(f_{M}\left(\pi_{M}\right)-\mathbb{E}_{p_{s}}\left[f_{M}(\pi)\right]\right)\right)\right) .
\]
```

Algorithm 11 is inspired by and closely related to the optimistic posterior sampling algorithm proposed in [119] (also termed as feel-good Thompson sampling in [120]). Our analysis of sequential estimation (see Appendix C.3.2) is built on the analysis in [119, 120]. However, our approach has adaptive terms in our algorithmic beliefs rather than using a pre-specified optimistic prior. Moreover, our regret bounds can be applied to both on-policy bilinear class as well as the general bilinear class (as we will explain shortly in Theorem 19 and Section 3.7.4), while the theoretical results of optimistic posterior sampling in [119] are only proved for on-policy bilinear class.

For model class $\mathcal{M}$, a nominal model $\bar{M}$, and the posterior sampling strategy $p(\pi)=\mu(\{M$ : $\left.\pi_{M}=\pi\right\}$ ), we can define the Bayesian decision-estimation coefficient of Thompson Sampling by

$$
\begin{equation*}
\operatorname{DEC}_{\eta}^{\mathrm{TS}}(\mathcal{M}, \bar{M})=\sup _{\mu \in \Delta(\mathcal{M})} \mathbb{E}_{v, p}\left[f_{M}\left(\pi_{M}\right)-f_{M}(\pi)-\frac{1}{\eta} D_{\mathrm{H}}^{2}(M(\pi), \bar{M}(\pi))\right] . \tag{3.7.3}
\end{equation*}
$$

This value is bigger than the minimax DEC in Definition 9, but often easier to use in model-based RL problems.

Theorem 19 (Regret of Model-index Adaptive Posterior Sampling) Given a finite model class $\mathcal{M}$ where $f_{M}(\pi) \in[0,1]$ for every $M \in \mathcal{M}$ and $\pi \in \Pi$. The regret of Algorithm 11 with $\eta \leq 1 / 10$ is bounded by

$$
\mathfrak{R}_{T} \leq \frac{\log (|\mathcal{M}| T)+1}{\eta}+5 \cdot \sup _{\bar{M} \in \mathcal{M}} \operatorname{DEC}_{2 \eta}^{\mathrm{TS}}(\mathcal{M}, \bar{M}) \cdot T+2 .
$$

### 3.7.4 Application to reinforcement learning

By using Algorithm 10 and Algorithm 11, we are able to recover several results in [105] that bound the regret of RL by DEC and the estimation complexity $\log |\mathcal{M}|$ of the model class. Note that we are able to prove such results for the Model-index APS (Algorithm 6), which has the potential to be efficiently implemented through efficient sampling oracles. In contrast, the E2D algorithm in [105] is not in closed form and requires minimax optimization, and the sharp regret bounds are proved through the non-constructive Bayesian Thompson Sampling. The paper also presents regret bounds for a constructive algorithm using the so-called "inverse gap weighting" updating rules, but that algorithm has worse regret bounds than those proved through the non-constructive approach (by a factor of the bilinear dimension). As a result, Algorithm 11 makes an improvement because its simplicity and achieving the sharpest regret bound proved in [105] for RL problems in the bilinear class.

We discuss how our general problem formulation in Section 3.2.1 covers RL problems as follows.

Example 11 (Reinforcement learning) An episodic finite-horizon reinforcement learning problems is defined as follows. Let $H$ be the horizon and $\mathcal{A}$ be a finite action set. Each model $M \in \mathcal{M}$ specifies a non-stationary Markov decision process (MDP) $\left\{\left\{S^{(h)}\right\}_{h=1}^{H}, \mathcal{A},\left\{P_{M}^{(h)}\right\}_{h=1}^{H},\left\{R_{M}^{(h)}\right\}_{h=1}^{H}, \mu\right\}$, where $\mu$ is the initial distribution over states; and for each layer $h, S^{(h)}$ is a finite state space, $P_{M}^{(h)}: S^{(h)} \times \mathcal{A} \rightarrow\left(S^{(h+1)}\right)$ is the probability transition kernel, and $R_{M}^{(h)}: S^{(h)} \times \mathcal{A} \rightarrow \Delta([0,1])$ is the reward distribution. We allow the transition kernel and loss distribution to be different for different $M \in \mathcal{M}$ but assume $\mu$ to be fixed for simplicity. Let $\Pi_{N S}$ be the space of all deterministic non-stationary policies $\pi=\left(u^{(1)}, \ldots, u^{(H)}\right)$, where $\left.u^{(h)}: S^{( } h\right) \rightarrow \mathcal{A}$. Given an MDP M and policy $\pi$, the MDP evolves as follows: beginning from $s^{(1)} \sim \mu$, at each layer $h=1, \ldots, H$, the action $a^{(h)}$ is sampled from $u^{(h)}\left(s^{(h)}\right)$, the loss $r^{(h)}\left(a^{(h)}\right)$ is sampled from $R_{M}\left(s^{(h)}, a^{(h)}\right)$ and the state $s^{(h+1)}$ is sampled from $P_{M}\left(\cdot \mid s^{(h)}, a^{(h)}\right)$. Define $f_{M}(\pi)=\mathbb{E}\left[\Sigma_{h=1}^{H} r^{(h)}\left(a^{(h)}\right)\right]$ to be the expected reward under MDP M and policy $\pi$. The general framework covers episodic reinforcement learning
problems by taking the observation $o_{t}$ to be the trajectory $\left(s_{t}^{(1)}, a_{t}^{(1)}, r_{t}^{(1)}\right), \ldots,\left(s_{t}^{(H)}, a_{t}^{(H)}, r_{t}^{(H)}\right)$ and $\Pi$ be a subspace of $\Pi_{N S}$. While our framework and complexity measures allow for agnostic policy classes, recovering existing results often requires us to make realizability-type assumptions.

We now focus on a broad class of structured reinforcement learning problems called "bilinear class" $[112,105]$. The following definition of the bilinear class is from [105].

Definition 11 (Bilinear class) A model class $\mathcal{M}$ is said to be bilinear relative to reference model $\bar{M}$ if:

1. There exist functions $W_{h}(\cdot ; \bar{M}): \mathcal{M} \rightarrow \mathbb{R}^{d}, X_{h}(\cdot ; \bar{M}): \mathcal{M} \times \mathbb{R}^{d}$ such that for all $M \in \mathcal{M}$ and $h \in[H]$,

$$
\left|\mathbb{E}^{\bar{M}, \pi_{M}}\left[Q_{h}^{M, *}\left(s_{h}, a_{h}\right)-r_{h}-V_{h}^{M, *}\left(s_{h+1}\right)\right]\right| \leq\left|\left\langle W_{h}(M ; \bar{M}), X_{h}(M ; \bar{M})\right\rangle\right| .
$$

We assume that $W_{h}(M: \bar{M})=0$.
2. Let $z_{h}=\left(s_{h}, a_{h}, r_{h}, s_{h+1}\right)$. There exists a collection of estimation policies $\left\{\pi_{M}^{\mathrm{est}}\right\}_{M \in \mathcal{M}}$ and estimation functions $\left\{\ell_{M}^{\text {est }}(\cdot ; \cdot)\right\}_{M \in \mathcal{M}}$ such that for all $M, M^{\prime} \in \mathcal{M}$ and $h \in[H]$,

$$
\left\langle X_{h}(M ; \bar{M}), W_{h}\left(M^{\prime}: \bar{M}\right)\right\rangle=\mathbb{E}^{\bar{M}, \pi_{M} \circ h \pi_{M}^{\text {est }}}\left[\ell_{M}^{\text {est }}\left(M^{\prime} ; z_{h}\right)\right] .
$$

If $\pi_{M}^{\mathrm{est}}=\pi_{M}$, we say that estimation is on-policy.
If $M$ is bilinear relative to all $\bar{M} \in \mathcal{M}$, we say that $\mathcal{M}$ is a bilinear class. We let $d_{\mathrm{bi}}(\mathcal{M}, \bar{M})$ denote the minimal dimension $d$ for which the bilinear class property holds relative to $\bar{M}$, and define $d_{\mathrm{bi}}(\mathcal{M})=\sup _{\bar{M} \in \mathcal{M}} d_{\mathrm{bi}}(\mathcal{M}, \bar{M})$. We let $L_{\mathrm{bi}}(\mathcal{M} ; \bar{M}) \geq 1$ denote any almost sure upper bound on $\left|\ell_{M}^{\text {est }}\left(M^{\prime} ; z_{h}\right)\right|$ under $\bar{M}$, and let $L_{\mathrm{bi}}(\mathcal{M})=\sup _{\bar{M} \in \mathcal{M}} L_{\mathrm{bi}}(\mathcal{M} ; \bar{M})$.

For $\gamma \in[0,1]$, let $\pi_{M}^{\gamma}$ be the randomized policy that—for each $h —$ plays $\pi_{M, h}$ with probability $1-\gamma / H$ and $\pi_{M, h}^{\text {est }}$ with probability $\gamma / H$. As an application of Theorem 7.1 in [105], we have upper bounds for $\mathrm{DEC}_{\eta}^{\mathrm{TS}}$ as follows.

Proposition 4 (Upper bounds for bilinear class reinforcement learning) Let $\mathcal{M}$ be a bilinear class and let $\bar{M} \in \mathcal{M}$. Let $\mu \in \Delta(\mathcal{M})$ be given, and consider the modified Bayesian posterior sampling strategy that samples $M \sim \mu$ and plays $\pi_{M}^{\alpha}$, where $\alpha \in[0,1]$ is a parameter.

1. If $\pi_{M}^{\mathrm{est}}=\pi_{M}$ (i.e., estimation is on-policy), this strategy with $\alpha=0$ certifies that

$$
\operatorname{DEC}_{\eta}^{\mathrm{TS}}(\mathcal{M}, \bar{M}) \leq 4 \eta H^{2} L_{\mathrm{bi}}^{2}(\mathcal{M}) d_{\mathrm{bi}}(\mathcal{M} ; \bar{M})
$$

for all $\eta>0$.
2. For general estimation policies, this strategy with $\gamma=\left(8 \eta H^{3} L_{\mathrm{bi}}^{2}(\mathcal{M}) d_{\mathrm{bi}}(\mathcal{M}, \bar{M})\right)^{1 / 2}$ certifies that

$$
\operatorname{DEC}_{\eta}^{\mathrm{TS}_{\gamma}}(\mathcal{M}, \bar{M}) \leq\left(32 \eta H^{3} L_{\mathrm{bi}}^{2}(\mathcal{M}) d_{\mathrm{bi}}(\mathcal{M} ; \bar{M})\right)^{1 / 2}
$$

whenever $\gamma \geq 32 H^{3} L_{\mathrm{bi}}^{2}(\mathcal{M}) d_{\mathrm{bi}}(\mathcal{M}, \bar{M})$.

By applying the upper bounds on $\mathrm{DEC}^{\mathrm{TS}} \eta$ from Proposition 4 to Theorem 19, we can immediately obtain regret guarantees for RL problems in the bilinear class. In the on-policy case, Algorithm 11 with optimally tuned $\eta$ achieves regret

$$
\mathfrak{R}_{T} \leq O\left(H^{2} L \mathrm{bi}^{2} d_{\mathrm{bi}}(\mathcal{M}) \cdot T \cdot \log |\mathcal{M}|\right) .
$$

In the general case, Algorithm 11 with forced exploration rate $\gamma=\left(8 \eta H^{3} L_{\mathrm{bi}}^{2}(\mathcal{M}) d_{\mathrm{bi}}(\mathcal{M}, \bar{M})\right)^{1 / 2}$ and optimally tuned $\eta$ achieves regret

$$
\Re_{T} \leq O\left(\left(H^{3} L_{\mathrm{bi}}^{2} d_{\mathrm{bi}}(\mathcal{M}) \log |\mathcal{M}|\right)^{1 / 3} \cdot T^{2 / 3}\right)
$$

As a closed-form algorithm that may be computed through sampling techniques, Algorithm 11 matches the sharp results for the non-constructive Bayesian Posterior Sampling algorithm proved in [105], and it achieves better regret bounds than the closed-form "inverse gap weighting" al-
gorithm provided in the same paper. Its regret bound for RL problems in the bilinear class also match the E2D algorithms in $[105,121]$ that are not in closed-form and require more challenging minimax optimization.

Our results in this section apply to reinforcement learning problems where the DEC is easy to upper bound, but bounding the information ratio may be more challenging, particularly for complex RL problems where the model class $\mathcal{M}$ may not be convex and the average of two MDPs may not belong to the model class. Specifically, we propose MAIR and provide a generic algorithm that uses DEC and the estimation complexity of the model class $(\log |\mathcal{M}|)$ to bound the regret. Another promising research direction is to extend our general results for AIR and the tools from Section 3.6.2 to reinforcement learning problems with suitably bounded information ratios, such as tabular MDPs and linear MDPs, as suggested in [111]. We anticipate that our tools can pave the way for developing constructive algorithms that provide regret bounds scaling solely with the estimation complexity of the value function class, which is typically smaller than that of the model class.

### 3.8 Conclusion and future directions

In this work, we propose a novel approach to solve sequential learning problems by generating "algorithmic beliefs." We optimize the Algorithmic Information Ratio (AIR) to generate these beliefs. Surprisingly, our algorithms achieve regret bounds that are as good as those assuming prior knowledge, even in the absence of such knowledge, which is often the case in adversarial or complex environments. Our approach results in simple and often efficient algorithms for various problems, such as multi-armed bandits, linear and convex bandits, and reinforcement learning.

Our work provides a new perspective on designing and analyzing bandit and reinforcement learning algorithms. Our theory applies to any algorithm through the notions of AIR and algorithmic beliefs, and it provides a simple and constructive understanding of the duality between frequentist regret and Bayesian regret in sequential learning. Optimizing AIR is a key principle to design effective and efficient bandit and RL algorithms. We demonstrate the effectiveness of our frame-
work empirically via experiments on Bernoulli MAB and show that our derived algorithm achieves "best-of-all-worlds" empirical performance. Specifically, our algorithm outperforms UCB and is comparable to TS in stochastic bandits, outperforms EXP3 in adversarial bandits, and outperforms TS as well as clairvoyant restarted algorithms in non-stationary bandits.

Our study suggests several future research directions. First, we aim to provide computational guidelines for optimizing algorithmic beliefs, including techniques for selecting belief subspaces, parameterization, and surrogate objective functions. Second, we plan to develop efficient algorithm designs for infinite-armed bandit and reinforcement learning problems. As a first step, we aim to explore the Bayesian interpretation of frequentist approaches, such as gaining a deeper understanding of the inverse probability weighting (IPW) estimators and existing computationallyefficient algorithms for infinite-armed bandits (such as SCRiBLe [122]). Third, we aim to simulate average-case or non-stationary environments through constraint optimization for algorithmic beliefs. Fourth, we plan to investigate the essential features of the offset and constraint formulations in the algorithmic belief approach and explore possible connections with localized complexity in statistical learning theory [123] (offset formulation of DEC has been recently studied in [124]). Lastly, we aim to study instance-dependent bounds by leveraging AIR and algorithmic beliefs, which, to the best of our knowledge, is currently lacking in the context of information ratio.

## Conclusion

As the field of machine learning rapidly evolves, the demand for creating intelligent systems that can handle large-scale and complex data sets grows. This requires developing new mathematical frameworks to understand the behavior of learning algorithms, including statistical and sequential complexities, constructive optimization procedures, and uniform convergence and localization. This thesis aimed to contribute to the emerging field of learning theory by advancing principled algorithm design and localized statistical complexity. The proposed frameworks and algorithms studied uniform convergence and localization in statistical learning theory, developed efficient algorithms using the optimism principle for contextual bandits, and created Bayesian design principles for bandit and reinforcement learning problems. Comprising three parts, each focused on a different but related area of research, Chapter 1 introduced a principled framework for uniform localized convergence, resolving limitations of existing approaches and characterizing sharp problem-dependent generalization error bounds for central statistical learning problems. Chapter 2 proposed a simple and generic principle to design optimistic algorithms for contextual bandits, which proved to be efficient and optimal even in the presence of large context and action spaces. Chapter 3 developed a novel optimization approach to create "algorithmic beliefs" for sequential learning problems, resulting in simple and often efficient algorithms for multi-armed bandits, linear and convex bandits, and reinforcement learning. The contributions of this thesis establish a strong foundation for future research towards developing more potent learning algorithms and a deeper understanding of the fundamental principles that underlie intelligent behavior.

## References

[1] P. L. Bartlett, O. Bousquet, and S. Mendelson, "Local Rademacher complexities," The Annals of Statistics, vol. 33, no. 4, pp. 1497-1537, 2005.
[2] V. Koltchinskii and D. Panchenko, "Rademacher processes and bounding the risk of function learning," in High Dimensional Probability II, Springer, 2000, pp. 443-457.
[3] M. J. Wainwright, High-dimensional statistics: A non-asymptotic viewpoint. Cambridge University Press, 2019, vol. 48.
[4] H. Namkoong and J. C. Duchi, "Variance-based regularization with convex objectives," in Advances in Neural Information Processing Systems, 2017, pp. 2971-2980.
[5] D. J. Foster and V. Syrgkanis, "Orthogonal statistical learning," arXiv preprint arXiv:1901.09036, 2019.
[6] A. Maurer and M. Pontil, "Empirical Bernstein bounds and sample variance penalization," arXiv preprint arXiv:0907.3740, 2009.
[7] A. Swaminathan and T. Joachims, "Counterfactual risk minimization: Learning from logged bandit feedback," in International Conference on Machine Learning, 2015, pp. 814-823.
[8] S. Mendelson, "Learning without concentration," in Conference on Learning Theory, 2014, pp. 25-39.
[9] S. Mendelson, "Learning without concentration for general loss functions," Probability Theory and Related Fields, vol. 171, no. 1-2, pp. 459-502, 2018.
[10] T. Liang, A. Rakhlin, and K. Sridharan, "Learning with square loss: Localization through offset rademacher complexity," in Conference on Learning Theory, 2015, pp. 1260-1285.
[11] E. Giné and V. Koltchinskii, "Concentration inequalities and asymptotic results for ratio type empirical processes," The Annals of Probability, vol. 34, no. 3, pp. 1143-1216, 2006.
[12] N. Zhivotovskiy and S. Hanneke, "Localization of vc classes: Beyond local Rademacher complexities," in International Conference on Algorithmic Learning Theory, Springer, 2016, pp. 18-33.
[13] V. Koltchinskii, Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008. Springer Science \& Business Media, 2011, vol. 2033.
[14] D. Pollard, Convergence of Stochastic Processes. Springer Science \& Business Media, 2012.
[15] A. W. Van der Vaart and J. A. Wellner, Weak convergence and empirical processes. Springer, 1996.
[16] A. B. Tsybakov, "Optimal rates of aggregation," in Learning Theory and Kernel Machines, Springer, 2003, pp. 303-313.
[17] P. Massart and É. Nédélec, "Risk bounds for statistical learning," The Annals of Statistics, vol. 34, no. 5, pp. 2326-2366, 2006.
[18] N. Zhivotovskiy, "Optimal learning via local entropies and sample compression," arXiv preprint arXiv:1706.01124, 2017.
[19] S. Mei, Y. Bai, and A. Montanari, "The landscape of empirical risk for nonconvex losses," The Annals of Statistics, vol. 46, no. 6A, pp. 2747-2774, 2018.
[20] D. J. Foster, A. Sekhari, and K. Sridharan, "Uniform convergence of gradients for nonconvex learning and optimization," in Advances in Neural Information Processing Systems, 2018, pp. 8745-8756.
[21] S. Balakrishnan, M. J. Wainwright, and B. Yu, "Statistical guarantees for the EM algorithm: From population to sample-based analysis," The Annals of Statistics, vol. 45, no. 1, pp. 77120, 2017.
[22] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk, A distribution-free theory of nonparametric regression. Springer Science \& Business Media, 2006.
[23] W. Rudin, Principles of mathematical analysis. McGraw-hill New York, 1964, vol. 3.
[24] K. Sridharan, "Note on refined dudley integral covering number bound," Unpublished Manuscript, https://www.cs.cornell.edu/ sridharan/dudley.pdf, 2010.
[25] G. Lugosi, "Pattern classification and learning theory," in Principles of Nonparametric Learning, Springer, 2002, pp. 1-56.
[26] A. W. Van der Vaart, Asymptotic statistics. Cambridge university press, 2000, vol. 3.
[27] L. Devroye and G. Lugosi, "Lower bounds in pattern recognition and learning," Pattern Recognition, vol. 28, no. 7, pp. 1011-1018, 1995.
[28] S. Athey and S. Wager, "Efficient policy learning," arXiv preprint arXiv:1702.02896, 2017.
[29] D. Davis and D. Drusvyatskiy, "Graphical convergence of subgradients in nonconvex optimization and learning," arXiv preprint arXiv:1810.07590, 2018.
[30] S. Mendelson, "On aggregation for heavy-tailed classes," Probability Theory and Related Fields, vol. 168, no. 3-4, pp. 641-674, 2017.
[31] S. Mendelson and N. Zhivotovskiy, "Robust covariance estimation under $L_{4}-L_{2}$ norm equivalence," arXiv preprint arXiv:1809.10462, 2018.
[32] H. Karimi, J. Nutini, and M. Schmidt, "Linear convergence of gradient and proximalgradient methods under the Polyak-łojasiewicz condition," in Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2016, pp. 795811.
[33] Y. Li and Y. Yuan, "Convergence analysis of two-layer neural networks with relu activation," in Advances in Neural Information Processing Systems, 2017, pp. 597-607.
[34] R. Kleinberg, Y. Li, and Y. Yuan, "An alternative view: When does SGD escape local minima?" arXiv preprint arXiv:1802.06175, 2018.
[35] Y. Zhou, J. Yang, H. Zhang, Y. Liang, and V. Tarokh, "SGD converges to global minimum in deep learning via star-convex path," arXiv preprint arXiv:1901.00451, 2019.
[36] O. Hinder, A. Sidford, and N. Sohoni, "Near-optimal methods for minimizing star-convex functions and beyond," in Conference on Learning Theory, PMLR, 2020, pp. 1894-1938.
[37] M. Hardt, T. Ma, and B. Recht, "Gradient descent learns linear dynamical systems," The Journal of Machine Learning Research, vol. 19, no. 1, pp. 1025-1068, 2018.
[38] H. Liu, W. Su, and A. M.-C. So, "Quadratic optimization with orthogonality constraints: Explicit Lojasiewicz exponent and linear convergence of line-search methods.," in International Conference on Machine Learning, 2016, pp. 1158-1167.
[39] M. Hardt and T. Ma, "Identity matters in deep learning," arXiv preprint arXiv:1611.04231, 2016.
[40] J. Sun, Q. Qu, and J. Wright, "A geometric analysis of phase retrieval," Foundations of Computational Mathematics, vol. 18, no. 5, pp. 1131-1198, 2018.
[41] X. Li, S. Ling, T. Strohmer, and K. Wei, "Rapid, robust, and reliable blind deconvolution via nonconvex optimization," Applied and computational harmonic analysis, vol. 47, no. 3, pp. 893-934, 2019.
[42] S. Ghadimi and G. Lan, "Stochastic first-and zeroth-order methods for nonconvex stochastic programming," SIAM Journal on Optimization, vol. 23, no. 4, pp. 2341-2368, 2013.
[43] S. J. Reddi, A. Hefny, S. Sra, B. Póczos, and A. Smola, "Stochastic variance reduction for nonconvex optimization," in International Conference on Machine Learning, 2016, pp. 314-323.
[44] A. Van der Vaart, "On the asymptotic information bound," The Annals of Statistics, pp. 14871500, 1989.
[45] J. Hájek, "Local asymptotic minimax and admissibility in estimation," in Proceedings of the sixth Berkeley symposium on mathematical statistics and probability, vol. 1, 1972, pp. 175-194.
[46] L Le Cam, "Limits of experiments," in Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Theory of Statistics, The Regents of the University of California, 1972.
[47] A. E. Litvak, A. Pajor, M. Rudelson, and N. Tomczak-Jaegermann, "Smallest singular value of random matrices and geometry of random polytopes," Advances in Mathematics, vol. 195, no. 2, pp. 491-523, 2005.
[48] R. Vershynin, High-dimensional probability: An introduction with applications in data science. Cambridge University Press, 2018, vol. 47.
[49] O. Chapelle, C. B. Do, C. H. Teo, Q. V. Le, and A. J. Smola, "Tighter bounds for structured estimation," in Advances in Neural Information Processing Systems, 2009, pp. 281-288.
[50] J. N. Laska and R. G. Baraniuk, "Regime change: Bit-depth versus measurement-rate in compressive sensing," IEEE Transactions on Signal Processing, vol. 60, no. 7, pp. 34963505, 2012.
[51] T. Nguyen and S. Sanner, "Algorithms for direct 0-1 loss optimization in binary classification," in International Conference on Machine Learning, 2013, pp. 1085-1093.
[52] S. Shalev-Shwartz, "Stochastic convex optimization.," in Conference on Learning Theory, 2009.
[53] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan, "Learnability, stability and uniform convergence," Journal of Machine Learning Research, vol. 11, no. Oct, pp. 26352670, 2010.
[54] A. Shapiro, D. Dentcheva, and A. Ruszczyński, Lectures on stochastic programming: modeling and theory. SIAM, 2009.
[55] S. M. Kakade and A. Tewari, "On the generalization ability of online strongly convex programming algorithms," in Advances in Neural Information Processing Systems, 2009, pp. 801-808.
[56] L. Zhang, T. Yang, and R. Jin, "Empirical risk minimization for stochastic convex optimization: $O(1 / n)$-and $O\left(1 / n^{2}\right)$-type of risk bounds," arXiv preprint arXiv:1702.02030, 2017.
[57] L. Zhang and Z.-H. Zhou, "Stochastic approximation of smooth and strongly convex functions: Beyond the $O(1 / T)$ convergence rate," arXiv preprint arXiv:1901.09344, 2019.
[58] M. Liu, X. Zhang, L. Zhang, R. Jin, and T. Yang, "Fast rates of ERM and stochastic approximation: Adaptive to error bound conditions," in Advances in Neural Information Processing Systems, 2018, pp. 4678-4689.
[59] Y. Lei and Y. Ying, "Sharper generalization bounds for learning with gradient-dominated objective functions," in International Conference on Learning Representations, 2021.
[60] G. J. McLachlan and T. Krishnan, The EM algorithm and extensions. John Wiley \& Sons, 2007, vol. 382.
[61] Z. Wang, Q. Gu, Y. Ning, and H. Liu, "High dimensional expectation-maximization algorithm: Statistical optimization and asymptotic normality," arXiv preprint arXiv:1412.8729, 2014.
[62] J. Kwon, W. Qian, C. Caramanis, Y. Chen, and D. Davis, "Global convergence of EM algorithm for mixtures of two component linear regression," arXiv preprint arXiv:1810.05752, 2018.
[63] N. Ho, K. Khamaru, R. Dwivedi, M. J. Wainwright, M. I. Jordan, and B. Yu, "Instability, computational efficiency and statistical accuracy," arXiv preprint arXiv:2005.11411, 2020.
[64] J. Blanchet, Y. Kang, and K. Murthy, "Robust Wasserstein profile inference and applications to machine learning," Journal of Applied Probability, vol. 56, no. 3, pp. 830-857, 2019.
[65] P. L. Bartlett, D. J. Foster, and M. J. Telgarsky, "Spectrally-normalized margin bounds for neural networks," in Advances in Neural Information Processing Systems, 2017, pp. 62406249.
[66] N. Golowich, A. Rakhlin, and O. Shamir, "Size-independent sample complexity of neural networks," in Conference On Learning Theory, PMLR, 2018, pp. 297-299.
[67] V. Nagarajan and J. Z. Kolter, "Uniform convergence may be unable to explain generalization in deep learning," in Advances in Neural Information Processing Systems, 2019, pp. 11 611-11 622.
[68] P. L. Bartlett and P. M. Long, "Failures of model-dependent generalization bounds for least-norm interpolation," arXiv preprint arXiv:2010.08479, 2020.
[69] L. Zhou, D. Sutherland, and N. Srebro, "On uniform convergence and low-norm interpolation learning," arXiv preprint arXiv:2006.05942, 2020.
[70] W. Chu, L. Li, L. Reyzin, and R. Schapire, "Contextual bandits with linear payoff functions," in Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, 2011, pp. 208-214.
[71] L. Li, Y. Lu, and D. Zhou, "Provably optimal algorithms for generalized linear contextual bandits," in Proceedings of the 34th International Conference on Machine LearningVolume 70, JMLR. org, 2017, pp. 2071-2080.
[72] A. Bietti, A. Agarwal, and J. Langford, "A contextual bandit bake-off," arXiv preprint arXiv:1802.04064, 2018.
[73] D. J. Foster, A. Agarwal, M. Dudík, H. Luo, and R. E. Schapire, "Practical contextual bandits with regression oracles," arXiv preprint arXiv:1803.01088, 2018.
[74] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The nonstochastic multiarmed bandit problem," SIAM journal on computing, vol. 32, no. 1, pp. 48-77, 2002.
[75] M. Dudik et al., "Efficient optimal learning for contextual bandits," arXiv preprint arXiv:1106.2369, 2011.
[76] A. Agarwal, M. Dudík, S. Kale, J. Langford, and R. Schapire, "Contextual bandit learning with predictable rewards," in Artificial Intelligence and Statistics, 2012, pp. 19-26.
[77] A. Agarwal, D. Hsu, S. Kale, J. Langford, L. Li, and R. Schapire, "Taming the monster: A fast and simple algorithm for contextual bandits," in International Conference on Machine Learning, 2014, pp. 1638-1646.
[78] D. J. Foster and A. Rakhlin, "Beyond ucb: Optimal and efficient contextual bandits with regression oracles," arXiv preprint arXiv:2002.04926, 2020.
[79] D. Simchi-Levi and Y. Xu, "Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability," arXiv preprint arXiv:2003.12699, 2020.
[80] J. Langford and T. Zhang, "The epoch-greedy algorithm for multi-armed bandits with side information," Advances in neural information processing systems, vol. 20, pp. 817-824, 2007.
[81] A. Krishnamurthy, A. Agarwal, T.-K. Huang, H. Daumé III, and J. Langford, "Active learning for cost-sensitive classification.," Journal of Machine Learning Research, vol. 20, no. 65, pp. 1-50, 2019.
[82] A. Goldenshluger and A. Zeevi, "A linear response bandit problem," Stochastic Systems, vol. 3, no. 1, pp. 230-261, 2013.
[83] D. Russo and B. Van Roy, "Eluder dimension and the sample complexity of optimistic exploration," in Advances in Neural Information Processing Systems, 2013, pp. 2256-2264.
[84] P. Rigollet and A. Zeevi, "Nonparametric bandits with covariates," arXiv preprint arXiv:1003.1630, 2010.
[85] A. Slivkins, "Contextual bandits with similarity information," The Journal of Machine Learning Research, vol. 15, no. 1, pp. 2533-2568, 2014.
[86] J. Abernethy, K. Amin, M. Kearns, and M. Draief, "Large-scale bandit problems and kwik learning," in International Conference on Machine Learning, 2013, pp. 588-596.
[87] A. Krishnamurthy, J. Langford, A. Slivkins, and C. Zhang, "Contextual bandits with continuous actions: Smoothing, zooming, and adapting," arXiv preprint arXiv:1902.01520, 2019.
[88] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, "Improved algorithms for linear stochastic bandits," in Advances in Neural Information Processing Systems, 2011, pp. 2312-2320.
[89] V. Dani, T. P. Hayes, and S. M. Kakade, "Stochastic linear optimization under bandit feedback," 2008.
[90] S. Filippi, O. Cappe, A. Garivier, and C. Szepesvári, "Parametric bandits: The generalized linear case," in Advances in Neural Information Processing Systems, 2010, pp. 586-594.
[91] R. D. Kleinberg, "Nearly tight bounds for the continuum-armed bandit problem," in Advances in Neural Information Processing Systems, 2005, pp. 697-704.
[92] V. Chernozhukov, M. Demirer, G. Lewis, and V. Syrgkanis, "Semi-parametric efficient policy learning with continuous actions," in Advances in Neural Information Processing Systems, 2019, pp. 15039-15049.
[93] B. Awerbuch and R. D. Kleinberg, "Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches," in Proceedings of the thirty-sixth annual ACM symposium on Theory of computing, 2004, pp. 45-53.
[94] N. Jiang, A. Krishnamurthy, A. Agarwal, J. Langford, and R. E. Schapire, "Contextual decision processes with low bellman rank are pac-learnable," in International Conference on Machine Learning, 2017, pp. 1704-1713.
[95] N. Srinivas, A. Krause, S. M. Kakade, and M. Seeger, "Gaussian process optimization in the bandit setting: No regret and experimental design," arXiv preprint arXiv:0912.3995, 2009.
[96] A. Badanidiyuru, R. Kleinberg, and A. Slivkins, "Bandits with knapsacks," in 2013 IEEE 54th Annual Symposium on Foundations of Computer Science, IEEE, 2013, pp. 207-216.
[97] A. Garivier and E. Moulines, "On upper-confidence bound policies for non-stationary bandit problems," arXiv preprint arXiv:0805.3415, 2008.
[98] L. Li, W. Chu, J. Langford, and R. E. Schapire, "A contextual-bandit approach to personalized news article recommendation," in Proceedings of the 19th international conference on World wide web, 2010, pp. 661-670.
[99] D. Silver et al., "Mastering the game of go with deep neural networks and tree search," nature, vol. 529, no. 7587, pp. 484-489, 2016.
[100] V. Mnih et al., "Human-level control through deep reinforcement learning," nature, vol. 518, no. 7540, pp. 529-533, 2015.
[101] T. L. Lai, H. Robbins, et al., "Asymptotically efficient adaptive allocation rules," Advances in applied mathematics, vol. 6, no. 1, pp. 4-22, 1985.
[102] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," Machine learning, vol. 47, no. 2, pp. 235-256, 2002.
[103] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," Biometrika, vol. 25, no. 3-4, pp. 285-294, 1933.
[104] D. Russo and B. Van Roy, "An information-theoretic analysis of thompson sampling," The Journal of Machine Learning Research, vol. 17, no. 1, pp. 2442-2471, 2016.
[105] D. J. Foster, S. M. Kakade, J. Qian, and A. Rakhlin, "The statistical complexity of interactive decision making," arXiv preprint arXiv:2112.13487, 2021.
[106] D. Russo and B. Van Roy, "Learning to optimize via information-directed sampling," Operations Research, vol. 66, no. 1, pp. 230-252, 2018.
[107] T. Lattimore and A. Gyorgy, "Mirror descent and the information ratio," in Conference on Learning Theory, PMLR, 2021, pp. 2965-2992.
[108] D. J. Foster, A. Rakhlin, A. Sekhari, and K. Sridharan, "On the complexity of adversarial decision making," arXiv preprint arXiv:2206.13063, 2022.
[109] T. Lattimore and C. Szepesvári, "An information-theoretic approach to minimax regret in partial monitoring," in Conference on Learning Theory, PMLR, 2019, pp. 2111-2139.
[110] T. Lattimore and C. Szepesvári, Bandit algorithms. Cambridge University Press, 2020.
[111] B. Hao and T. Lattimore, "Regret bounds for information-directed reinforcement learning," arXiv preprint arXiv:2206.04640, 2022.
[112] S. Du et al., "Bilinear classes: A structural framework for provable generalization in rl," in International Conference on Machine Learning, PMLR, 2021, pp. 2826-2836.
[113] C. Jin, Q. Liu, and S. Miryoosefi, "Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms," Advances in Neural Information Processing Systems, vol. 34, 2021.
[114] O. Besbes, Y. Gur, and A. Zeevi, "Stochastic multi-armed-bandit problem with non-stationary rewards," Advances in neural information processing systems, vol. 27, 2014.
[115] M. Sion, "On general minimax theorems.," Pacific Journal of mathematics, vol. 8, no. 1, pp. 171-176, 1958.
[116] V. Dani, S. M. Kakade, and T. Hayes, "The price of bandit information for online optimization," Advances in Neural Information Processing Systems, vol. 20, 2007.
[117] T. Lattimore, "Improved regret for zeroth-order adversarial bandit convex optimisation," Mathematical Statistics and Learning, vol. 2, no. 3, pp. 311-334, 2020.
[118] E. Hazan and Z. Karnin, "Volumetric spanners: An efficient exploration basis for learning," The Journal of Machine Learning Research, vol. 17, no. 1, pp. 4062-4095, 2016.
[119] A. Agarwal and T. Zhang, "Model-based rl with optimistic posterior sampling: Structural conditions and sample complexity," arXiv preprint arXiv:2206.07659, 2022.
[120] T. Zhang, "Feel-good thompson sampling for contextual bandits and reinforcement learning," arXiv preprint arXiv:2110.00871, 2021.
[121] D. J. Foster, N. Golowich, J. Qian, A. Rakhlin, and A. Sekhari, "A note on model-free reinforcement learning with the decision-estimation coefficient," arXiv preprint arXiv:2211.14250, 2022.
[122] J. Abernethy, E. Hazan, and A. Rakhlin, "Competing in the dark: An efficient algorithm for bandit linear optimization," in 21 st Annual Conference on Learning Theory, COLT 2008, 2008.
[123] Y. Xu and A. Zeevi, "Towards optimal problem dependent generalization error bounds in statistical learning theory," arXiv preprint arXiv:2011.06186, 2020.
[124] D. J. Foster, N. Golowich, and Y. Han, "Tight guarantees for interactive decision making with the decision-estimation coefficient," arXiv preprint arXiv:2301.08215, 2023.
[125] E. Hazan, T. Koren, and K. Y. Levy, "Logistic regression: Tight bounds for stochastic and online optimization," in Conference on Learning Theory, 2014, pp. 197-209.
[126] U. Marteau-Ferey, D. Ostrovskii, F. Bach, and A. Rudi, "Beyond least-squares: Fast rates for regularized empirical risk minimization through self-concordance," arXiv preprint arXiv:1902.03046, 2019.
[127] G. Lecué and S. Mendelson, "Sparse recovery under weak moment assumptions," arXiv preprint arXiv:1401.2188, 2014.
[128] V. Koltchinskii and S. Mendelson, "Bounding the smallest singular value of a random matrix without concentration," International Mathematics Research Notices, vol. 2015, no. 23, pp. 12991-13 008, 2015.
[129] M. Rudelson and R. Vershynin, "Small ball probabilities for linear images of high-dimensional distributions," International Mathematics Research Notices, vol. 2015, no. 19, pp. 95949617, 2015.
[130] G. Lecué and S. Mendelson, "Regularization and the small-ball method i: Sparse recovery," The Annals of Statistics, vol. 46, no. 2, pp. 611-641, 2018.
[131] R. Tibshirani, "Regression shrinkage and selection via the lasso," Journal of the Royal Statistical Society: Series B (Methodological), vol. 58, no. 1, pp. 267-288, 1996.
[132] M. Bogdan, E. Van Den Berg, C. Sabatti, W. Su, and E. J. Candès, "SLOPE—adaptive variable selection via convex optimization," The Annals of Applied Statistics, vol. 9, no. 3, p. 1103, 2015.
[133] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, "A unified framework for high-dimensional analysis of $M$-estimators with decomposable regularizers," Statistical Science, vol. 27, no. 4, pp. 538-557, 2012.
[134] P.-L. Loh and M. J. Wainwright, "Regularized $M$-estimators with nonconvexity: Statistical and algorithmic theory for local optima," in Advances in Neural Information Processing Systems, 2013, pp. 476-484.
[135] S. Mendelson, "Improving the sample complexity using global data," IEEE Transactions on Information Theory, vol. 48, no. 7, pp. 1977-1991, 2002.
[136] S. Mendelson, "An optimal unrestricted learning procedure," arXiv preprint arXiv:1707.05342, 2017.
[137] G. Raskutti, M. J. Wainwright, and B. Yu, "Minimax-optimal rates for sparse additive models over kernel classes via convex programming," Journal of Machine Learning Research, vol. 13, no. Feb, pp. 389-427, 2012.
[138] R. Meir and T. Zhang, "Generalization error bounds for Bayesian mixture algorithms," Journal of Machine Learning Research, vol. 4, no. Oct, pp. 839-860, 2003.
[139] S. Dirksen, "Tail bounds via generic chaining," Electronic Journal of Probability, vol. 20, 2015.
[140] I. Pinelis, "Optimum bounds for the distributions of martingales in Banach spaces," The Annals of Probability, vol. 22, no. 4, pp. 1679-1706, 1994.
[141] I. Pinelis, "Correction:"Optimum bounds for the distributions of martingales in Banach spaces"[ann. probab. 22 (1994), no. 4, 1679-1706; mr 96b: 60010]," The Annals of Probability, vol. 27, no. 4, pp. 2119-2119, 1999.
[142] M. Talagrand, "Majorizing measures: The generic chaining," The Annals of Probability, vol. 24, no. 3, pp. 1049-1103, 1996.
[143] P. L. Bartlett, V. Dani, T. Hayes, S. Kakade, A. Rakhlin, and A. Tewari, "High-probability regret bounds for bandit online linear optimization," in Proceedings of the 21st Annual Conference on Learning Theory, 2008.
[144] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," arXiv preprint arXiv:1011.3027, 2010.
[145] V. I. Bogachev and M. A. S. Ruas, Measure theory. Springer, 2007, vol. 1.
[146] M. Song, "Proving that the conditional entropy of a probability measure is concave," Mathematics Stack Exchange, vol. https://math.stackexchange.com/q/3080334, 2019.
[147] R. T. Rockafellar, "Convex analysis," in Convex analysis, Princeton university press, 2015.

## Appendix A: Appendix for Chapter 1

## Appendix A is organized as follows:

- Appendix A. 1 presents additional results of our general framework related to the topic of "learning without concentration" (proofs are deferred to Appendix A.4).
- Appendix A. 2 and A. 3 contain the proofs for all theoretical claims in the main paper. These can be read in a selective manner; key steps of each proof will be indicated shortly.
- Appendix A.2.3 and A.2.5 discuss how to compute loss-dependent and variance-dependent rates directly from data, filling in details that were omitted from the main paper for space consideration.

For convenience of readers, we will present a guide for reading Appendix A.

## Overview for Appendix A. 1

Appendix A. 1 applies our general framework to study supervised learning problems with structured convex cost. The relationship between this setting and other parts of the paper has been explained in Section 1.2.4 and Section 1.5.1. In essence, the treatment presented here allows for nonparametric and heavy-tailed hypothesis classes, whereas the main paper (Section 1.5-1.7) focuses on parametric models and assumes a sub-exponential type assumption (Assumption 1). Unlike the main paper which focuses primally on iterative optimization algorithms, popular non-convex learning problems, and generic-form stochastic optimization problems, the focus here is on better studied "classical" problem to best illustrate the key points.

The appendix presents work related to a stream of work pioneered by Mendelson and others referred to as "learning without concentration" [8, 9]. This line of work has motivated by the
need to find new localization approaches to replace the traditional "local Rademacher complexity" analysis. An interesting open question is whether one can achieve the same goals by directly strengthening the traditional concentration-based approaches. Our investigation shows that one of the core limitations of traditional localization approaches in this paradigm is its requirement of "sub-root" surrogate functions. Since our new approach completely remove this requirement, we are able to answer this question in the affirmative; the results from $[8,9]$ can be recovered via a concentration-based analysis. Moreover, we are able to show some technical improvements-our approach does not require the "star-hull" of the hypothesis class that may increase complexity, and there are concrete examples showing that the improvement may be meaningful for non-convex classes. Appendix A.1.3 focuses on these new findings.

## Overview of Appendix A. 2 and A. 3

Appendix A. 2 and A. 3 provide proofs for all the theoretical claims in the main paper. Readers may read them in a selective manner, and to that end, we present a high level guide here.

Per our "uniform localized convergence" principle, a proof for problem-dependent generalization error bounds contain two steps: 1) obtaining "localized uniform convergence" arguments; and 2) subsequent analysis that is customized to the problem setting. Among the major theoretical results in the main paper, the following are dedicated to the first step:

- Proposition 1 provides a general tool to prove "uniform localized convergence" arguments. The proof of this result is "one-shot" via a surprisingly simple observation explained in Section 1.2.2. The formal proof is given in Appendix A.2.1 (we actually prove a more general version, Proposition 5), which is succinct and straightforward to verify.
- Proposition 2, the "uniform localized convergence of gradients" argument, is the foundation for all our results in the parametric "fast rate" regime. The proof (which is presented in Appendix A.3.2) crucially relies on a careful choice of concentrated function (1.5.3), a novel chaining analysis (see lines from (A.3.1) to (A.3.3)), and the application of Proposition 1.
- Theorem 2 (using empirical moment penalization to achieve optimal variance-dependent rate) crucially requires a "uniform localized convergence" argument where the measurement functional is data-dependent. The argument, Lemma 15 in Appendix A.2.4, uses tools from empirical processes theory and is somewhat technical in nature.

As for the second step (subsequent analysis that is customized to the problem setting and the learning algorithm), the paper presents three different approaches: 1) using the definition of the estimator to establish an inequality and then calculating the fixed point (this is used in most traditional approaches); 2) adding a regularization term and then directly using the definition of the regularized estimator; and 3) coupling the statistical error with analysis of an iterative optimization algorithm. Below are some of the key points.

- Theorem 1 (using empirical risk minimization to achieve optimal loss-dependent rate) uses the "fixed point analysis" approach. The core step is to establish an inequality where a "measurement" functional of $\hat{h}_{\text {ERM }}$ appears in both sides of the inequality (see inequality (A.2.8)). Then one can use the definition of the fixed point to prove loss-dependent generalization error bounds. The proof is presented in Appendix A.2.2.
- For Theorem 2 (using empirical moment penalization to achieve optimal variance-dependent rate), the core message is that the definition of the proposed moment-regularized estimator directly leads to variance-dependent generalization error bounds. The proof is somewhat lengthy, but the readers may focus on "Part III" in Appendix A.2.4 (in particular, the lines from (A.2.32) to (A.2.35)) for the main message.
- Theorem 3 ("fast rate" of approximate stationary points) uses the "fixed point analysis" approach, and the core step is to establish the inequality (A.3.10). The readers can parse (A.3.10) in a rather simple manner: the right hand side is mostly due to Proposition 2 (the "uniform localized convergence of gradients" argument); and the left hand side is due to the property of the Polyak-Lojasiewisz (PL) condition. The full proof is presented in Appendix A.3.3.
- The proofs for Theorem 4 (Appendix A.3.4) and Theorem 6 (Appendix A.3.6) are very similar. The core idea is to couple the statistical error from Proposition 2 to the optimization analysis of an iterative algorithm. The major difference is that Theorem 4 discusses samplebased gradient descent (before coupling the statistical error, its optimization analysis leads to (A.3.11)); and Theorem 6 discusses sample-based first-order Expectation-Maximization (before coupling the statistical error, its optimization analysis leads to (A.3.15)).

Additional corollaries. Besides the above, there are two corollaries in the main paper. Corollary 7 is the application of Theorem 6 to Example 4 (mixture of two Gaussians) and Example 5 (Mixture of two component linear regression). The explicit calculation of problem-dependent parameters here is quite novel (Appendix A.3.7), but the informal explanation at the end of Section 1.7.3 (from (1.7.11) to (1.7.12)) should serve as a better source to understand the main message. Corollary 5 is the application of Theorem 3 and Theorem 4 to Example 3 (non-convex regression with non-linear activation), and the verification of the assumptions here are mostly technical in nature (Appendix A.3.5).

Data-dependent bounds. Lastly but importantly, we would like to highlight Appendix A.2.3 and Appendix A.2.5-they discuss how to estimate the loss-dependent and variance-dependent rates from data, which are mentioned in the main paper (see remarks after Theorem 1 and Theorem 2) but details are omitted there. A central challenge is to replace the loss $\mathcal{L}^{*}$ and the variance $\mathcal{V}^{*}$ (which depends on the unknown "best hypothesis" $h^{*}$ ) by suitable empirical estimates. Readers who are interested in fully data-dependent generalization error bounds may find this of interest.

## A. 1 Fast rates in supervised learning with structured convex cost

The main purpose of this section is to recover the problem-dependent rates in [9, 8] for (possibly non-parametric and heavy-tailed) supervised learning problems with structured convex cost functions. While $[9,8]$ propose an approach they call "learning without concentration," our approach emphasizes the use of surrogate functions that are not "sub-root," and relates one-sided
uniform inequalities to two-sided concentration of "truncated" functions. Besides providing a unification, there are some technical improvements as well. For example, our approach does not require the "star-hull" of the hypothesis class that may increase complexity, and there are concrete examples showing that the improvement may be meaningful for non-convex classes. See Appendix A.1.3 for contributions of our method, and detailed comparison with existing approaches.

## A.1.1 Background

Problem formulation and assumptions. Let the data $z$ be a feature-label pair $(x, y)$ where $x \in \mathcal{X}$ and $y \in \mathcal{Y} \subseteq \mathbb{R}$. We assume every hypothesis $h$ in the hypothesis class $\mathcal{H}$ is a mapping from $\mathcal{X}$ to $\mathbb{R}$. In supervised learning, the loss function is of the form $\ell(h ;(x, y))=\ell_{\mathrm{sv}}(h(x), y)$ where the deterministic bivariate function $\ell_{\mathrm{SV}}: \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ is called the cost function. We assume that the cost function is differentiable, globally convex with respect to its first argument, and the population risk is smooth.

Assumption 9 (differentiability, convexity and smoothness) The partial derivative of $\ell_{\mathrm{su}}$ with respect to its first argument, denoted $\partial_{1} \ell_{\mathrm{su}}$, exists and is continuous everywhere, and $\ell_{\mathrm{sv}}$ is a convex function with respect to its first argument, i.e., $\forall u_{1}, u_{2}, y \in \mathbb{R}$,

$$
\ell_{\mathrm{su}}\left(u_{1}, y\right)-\ell_{\mathrm{su}}\left(u_{2}, y\right)-\partial_{1} \ell_{\mathrm{sv}}\left(u_{2}, y\right)\left(u_{1}-u_{2}\right) \geq 0 .
$$

In addition, the population risk is smooth, i.e., there exists a constant $\beta_{\mathrm{Sv}}>0$ such that $\forall h_{1}, h_{2} \in$ $\mathcal{H}$,
$\mathbb{P} \ell_{\mathrm{sv}}\left(h_{1}(x), y\right)-\mathbb{P} \ell_{\mathrm{sv}}\left(h_{2}(x), y\right)-\mathbb{P}\left[\partial_{1} \ell_{\mathrm{sv}}\left[\left(h_{2}(x), y\right)\left(h_{1}(x)-h_{2}(x)\right)\right] \leq \frac{\beta_{\mathrm{sv}}}{2} \mathbb{P}\left[\left(h_{1}(x)-h_{2}(x)\right)^{2}\right]\right.$.

Given a cost function that is globally convex and locally strongly convex, we define $\{\alpha(v)\}_{v \geq 0}$ as follows.

Definition 12 (strong convexity parameter) For a fixed $v>0$, let $\alpha(v)$ be the largest constant
such that for all $y \in \mathcal{Y}, \ell_{\mathrm{Sv}}(u+y, y)$ is $\alpha(v)-$ strongly convex with respect to $u$ when $u \in[-v, v]$. That is,
$\ell_{\mathrm{sv}}\left(u_{1}+y, y\right)-\ell_{\mathrm{sv}}\left(u_{2}+y, y\right)-\partial_{1} \ell_{\mathrm{sv}}\left(u_{2}+y, y\right)\left(u_{1}-u_{2}\right) \geq \frac{\alpha(v)}{2}\left(u_{1}-u_{2}\right)^{2}, \quad \forall u_{1}, u_{2} \in[-v, v], \forall y \in \mathcal{y}$.

Clearly $\{\alpha(v)\}_{v \geq 0}$ is non-increasing with respect to $v$, and we denote $\alpha(0)=\lim \sup _{v \rightarrow 0} \alpha(v)$.

When $\ell_{\mathrm{SV}}$ is second-order continuously differentiable, we have the simple relation

$$
\alpha(v)=\sup _{|u| \leq v, y \in \mathcal{Y}} \partial_{1,1}^{2} \ell_{\mathrm{sv}}(u+y, y), \quad, \forall v \geq 0,
$$

where $\partial_{1,1}^{2} \ell_{\mathrm{sv}}$ is the second order partial derivative of $\ell_{\mathrm{sV}}$ with respect to its first argument. Moreover, to accommodate popular choices of robust costs, Definition 12 also allows $\partial_{1} \ell_{\mathrm{sv}}$ to be nondifferentiable at certain points in its domain. We list three widely used cost functions, their strong convexity parameters $\{\alpha(v)\}_{v \geq 0}$, and the smoothness parameters $\beta_{\mathrm{Sv}}$ of the corresponding population risks.

- Square cost: consider the regression setting $\mathbb{E}[y \mid x]=h_{\text {true }}(x)$, where $h_{\text {true }}$ is the function we want to estimate (not necessarily in $\mathcal{H}$ ). It is natural to consider the square cost function

$$
\ell_{\mathrm{SV}}(h(x), y)=\frac{1}{2}(h(x)-y)^{2} .
$$

Here $\ell_{\mathrm{sv}}(u+y, y)=u^{2}$. Thus $\alpha(v)=\frac{1}{2}, \forall v \geq 0$. The smoothness parameter of the population risk is $\beta_{\mathrm{SV}}=\frac{1}{2}$. In this example, one does not need to localize the strong convexity parameter $\alpha(v)$ as it is a constant.

- Huber cost: consider the regression setting $\mathbb{E}[y \mid x]=h_{\text {true }}(x)$, where $h_{\text {true }}$ is the function we want to estimate (not necessarily in $\mathcal{H}$ ). When the conditional distribution of $y$ is "heavy
tailed," one often considers the Huber cost function as follows. For $\gamma>0$, let

$$
\ell_{\mathrm{sv}, \gamma}(h(x), y)= \begin{cases}\frac{1}{2}(h(x)-y)^{2} & \text { for }|h(x)-y| \leq \gamma  \tag{A.1.1}\\ \gamma|h(x)-y|-\frac{\gamma^{2}}{2} & \text { for }|h(x)-y|>\gamma\end{cases}
$$

Here $\alpha(v)=\frac{1}{2}$ whenever $v \leq \gamma$ but $\alpha(v)=0$ for all $v>\gamma$. The smoothness parameter of the population risk is $\beta_{\mathrm{Sv}}=\frac{1}{2}$. Localization analysis of $\alpha(v)$ is required for this loss, and the key is to avoid its inverse diverging to infinity.

- Logistic cost: consider the standard logistic regression setting, where $y \in\{-1,1\}$ and one models the "log odd ratio" as

$$
\begin{equation*}
\log (\operatorname{Prob}(y=1 \mid x) / \operatorname{Prob}(y=-1 \mid x))=h_{\text {true }}(x) \tag{A.1.2}
\end{equation*}
$$

Here $h_{\text {true }}$ is the discriminant function to be estimated (not necessarily in $\mathcal{H}$ ). The maximum likelihood estimation problem corresponds to using the cost function

$$
\ell_{\mathrm{sV}}(h(x), y)=\log (1+\exp (-y h(x))) .
$$

Here $\partial_{1,1}^{2} \ell_{\mathrm{Sv}}(u+y, y)=\frac{\exp (1+u y)}{(1+\exp (1+u y))^{2}}$, so we have $\alpha(v)=\frac{\exp (v+1)}{(\exp (v+1)+1)^{2}}, \forall v \geq 0$, and the smoothness parameter of the population risk is $\beta_{\mathrm{Sv}}=\frac{1}{4}$. The issue is that $\frac{1}{\alpha(v)}$, a complexity constant that will appear in the generalization error bound, grows exponentially with $v$ [125, 126]. This issue strongly motivate us to localize the parameter $v$ within $\alpha(v)$ to avoid large exponential constants.

The following assumption is usually invoked in the most representative literature on this topic [8, 9].

Assumption 10 (optimality condition) Recall that $h^{*} \in \mathbb{P} \ell_{\mathrm{SV}}(h(x), y)$ is the population risk min-
imizer. Assume for all $h \in \mathcal{H}$,

$$
\mathbb{P}\left[\partial_{1} \ell_{\mathrm{sv}}\left(h^{*}(x), y\right)\left(h(x)-h^{*}(x)\right)\right] \geq 0
$$

We summarize the two primary settings where Assumption 10 holds true.

- Well-specified models: for certain problems, as long as the model is well-specified, then $\partial_{1} \ell_{\mathrm{sV}}\left(h^{*}(x), y\right)$ is independent of $x$ and $\mathbb{E} \partial_{1} \ell_{\mathrm{sv}}\left(h^{*}(x), y\right)=0$. Thus Assumption 10 will hold. Examples include 1) the settings studied in [9] where $\ell_{\mathrm{sv}}$ is a univariate function of $(h(x)-y)$ and $\partial_{1} \ell_{\mathrm{sv}}\left(h^{*}(x), y\right)$ is odd with respect to $y$, such as applications that use the square cost or the Huber cost; and 2) generalized linear models where the conditional distribution of $y$ belongs to the exponential family, such as the the logistic regression problem (A.1.2).
- $\mathcal{H}$ is a convex class of functions: in this case, we verify Assumption 10 as follows. If there exists some $h_{1} \in \mathcal{H}$ such that Assumption 10 is not true, then by considering $h_{\lambda}=$ $\lambda h_{1}+(1-\lambda) h^{*} \in \mathcal{H}$ with $\lambda$ sufficiently close to 0 , we find $\mathbb{P} \ell_{\mathrm{sv}}\left(h_{\lambda}(x), y\right)<\mathbb{P} \ell_{\mathrm{Sv}}\left(h^{*}(x), y\right)$, in contradiction, as $h^{*}$ is the population risk minimizer.

We call the random variable $\partial_{1} \ell_{\mathrm{sv}}\left(h^{*}(x), y\right)$ the "noise multiplier" as it often characterizes the "effective noise" of the learning problem when using a particular cost function. We define another random variable $\xi:=h^{*}(x)-y$. In some applications, $\xi$ is closely related to the "noise multiplier" (e.g., they are equivalent when one uses the square cost). And the notation $\xi$ is useful in other applications as well, because one always seeks to localize the parameter $v$ in $\alpha(v)$ to the order of $\|\xi\|_{L_{2}}$. We denote $\Delta=\sup _{h \in \mathcal{H}}\|h(x)-y\|_{L_{2}}$ and $\Delta_{\infty}=\sup _{h, x, y}|h(x)-y|$ as the worst-case $L_{2}$ distance and $L_{\infty}$ distance between $h(x)$ and $y$. respectively. It is clear that we typically have $\|\xi\|_{L_{2}} \ll \Delta \ll \Delta_{\infty}$ in practical applications.

Our analysis requires a very weak distributional assumption:

Assumption 11 ("small ball" property) There exist constants $\kappa>0$ and $c_{\kappa} \in(0,1)$ such that
for all $h \in \mathcal{H}$,

$$
\operatorname{Prob}\left(\left|h(x)-h^{*}(x)\right| \geq \kappa\left\|h-h^{*}\right\|_{L_{2}}\right) \geq c_{\kappa} .
$$

Assumption 11 is often referred to as "minimal" in the literature, and there are many examples in which it can be verified for $\kappa$ and $c_{\kappa}$ that are absolute constants [8, 9, 127, 128, 129, 130]. The scope of Assumption 11 subsumes and is much broader than the "sub-Gaussian" setting. For example, it is naturally satisfied when the class $\left\{h-h^{*}: h \in \mathcal{H}\right\}$ satisfies any sort of moment equivalence (see, e.g., Lemma 4.1 in [8]).

Main challenges. Let us first examine limitations of the results obtained using the traditional "local Rademacher complexity" analysis (Statement 1), which includes the results from [1, 3, 5] in the fast-rate regime. Assuming the cost function to be $L_{\mathrm{SV}}$-Lipchitz continuous with respect to its first argument and setting $f(z)=\ell_{\mathrm{sv}}(h(x), y)-\ell_{\mathrm{sv}}\left(h^{*}(x), y\right), T(f)=\mathbb{P}\left[f^{2}\right]$, and $B_{e}=$ $L_{\mathrm{sv}}^{2} / \alpha\left(\Delta_{\infty}\right)$, following Statement 1, one can prove that the empirical risk minimizer $\hat{h}$ satisfies

$$
\begin{equation*}
\mathcal{E}(\hat{h}) \leq O\left(\frac{r^{*}}{B_{e}}\right), \tag{A.1.3}
\end{equation*}
$$

where $r^{*}$ is the fixed point of $B_{e} \psi$, and $\psi$ is a sub-root surrogate function that governs $\sup _{\mathbb{P}\left[f^{2}\right] \leq r}\left(\mathbb{P}-\mathbb{P}_{n}\right) f$. Denote by $r_{1}^{*}$ the fixed point of $\psi$. From the sub-root property of $\psi$ we know that $r^{*} \geq B_{e}^{2} r_{1}^{*}$, so the generalization error bound (A.1.3) is at least of order

$$
\begin{equation*}
\frac{r^{*}}{B_{e}} \geq B_{e} r_{1}^{*}=\frac{L_{\mathrm{sv}}^{2}}{\alpha\left(\Delta_{\infty}\right)} r_{1}^{*} \tag{A.1.4}
\end{equation*}
$$

The main message here is that the traditional result (A.1.3) is often loose and not problem-dependent. As indicated by Mendelson in a series of papers [8, 9], the traditional result (A.1.3) has the following limitations.

- The global Lipchitz constant $L_{\mathrm{sv}}$ is not problem-dependent and potentially unbounded.
$L_{\mathrm{sv}}$ is effectively the worst-case value $\sup _{h, x, y}\left|\partial_{1} \ell(h(x), y)\right|$. For the square cost, this is $\Delta_{\infty}=\sup _{h, x, y}|h(x)-y|$ and is unbounded when either the hypothesis class or noise are unbounded. It would be beneficial to have a bound that only scales with a measure related to the "noise multiplier" $\partial_{1} \ell\left(h^{*}(x), y\right)$, because we usually have $\left|\partial_{1} \ell\left(h^{*}(x), y\right)\right| \ll L_{\mathrm{Sv}}$ in practical applications.
- The global strong convexity parameter $\alpha\left(\Delta_{\infty}\right)$ is often very small for the logistic cost and the Huber cost, so its inverse is often large (and potentially unbounded). The challenge here is to sharpen this to the inverse of a localized strongly convex parameter $\alpha\left(O\left(\|\xi\|_{L_{2}}\right)\right)$. Since we usually have $\sigma \ll \Delta_{\infty}$, the inverse of $\alpha\left(O\left(\|\xi\|_{L_{2}}\right)\right)$ can be much smaller than the inverse of $\alpha\left(\Delta_{\infty}\right)$.

The "small ball method" and beyond. The breakthrough papers [8, 9] propose the "small ball method" (also referred to as "learning without concentration") to provide problem-dependent rates that overcome the limitations mentioned above. Their proofs builds on structural results of $0-1$ valued indicator functions under the small-ball condition, whose connection to the traditional localization analysis may not be completely obvious. Moving the focal point from indicator functions to "truncated" functions, we provide the following perspectives.

1) A simple interpretation to the "small-ball" condition is that, suitably "truncated" quadratic forms are of the same scale as the original quadratic forms. Under the "small-ball" condition, one can trivially show that uniformly over all $h \in \mathcal{H}$,

$$
\begin{aligned}
& \mathbb{P}\left[\min \left\{\left(h(x)-h^{*}(x)\right)^{2}, \kappa^{2}\left\|h-h^{*}\right\|_{L_{2}}^{2}\right\}\right] \geq \operatorname{Prob}\left(\left|h(x)-h^{*}(x)\right| \geq \kappa\left\|h-h^{*}\right\|_{L_{2}}\right) \kappa^{2}\left\|h-h^{*}\right\|_{L_{2}}^{2} \\
& \geq c_{\kappa} \kappa^{2} \mathbb{P}\left[\left(h(x)-h^{*}(x)\right)^{2}\right] .
\end{aligned}
$$

This suggests that one only needs to concentrate simple "truncated" functions to derive generalization error bounds.
2) One-sided uniform inequalities are contained in the "uniform localized convergence" frame-
work and are often derived from concentration of truncated functions. Many one-sided uniform inequalities can be equivalently written as "uniform localized convergence" arguments. Consider the uniform "lower isomorphic bound" (which plays a central role in the "small-ball" method): for some constant $c>0$, with high probability, uniformly over all $h \in \mathcal{H}$,

$$
\mathbb{P}_{n}\left[\left(h(x)-h^{*}(x)\right)^{2}\right] \geq c \mathbb{P}\left[\left(h(x)-h^{*}(x)^{2}\right] .\right.
$$

The above argument is equivalent with the following "uniform localized convergence" argument:

$$
\left(\mathbb{P}-\mathbb{P}_{n}\right)\left[\left(h(x)-h^{*}(x)\right)^{2}\right] \leq(1-c) T(h), \quad \forall h \in \mathcal{H}
$$

where the measurement functional $T(h)$ is set to be $\left\|h-h^{*}\right\|_{L_{2}}^{2}$. A more flexible perspective may directly view the truncated quadratic forms as the concentrated functions, making traditional twosided uniform convergence tools applicable in a straightforward manner.

Motivated by the above observations, an interesting open question is to recover the results in [8, 9] by directly strengthening the traditional concentration framework, explicitly figuring out which component of the excess loss contributes to which part of the surrogate function. In what follows, we will present such an analysis. While our error bounds roughly follow the same form as the results in [8, 9], we obtain several technical improvements; see Appendix A.1.3 for the novel implications and methodological contributions of our approach.

## A.1.2 Main results and illustrative examples

We assume some regularity conditions that hold for non-pathological choices of surrogate functions.

Assumption 12 (regularity conditions on surrogate functions) Assume there is a non-decreasing,
non-negative and bounded function $\varphi(r)$ such that $\forall r>0$,

$$
\begin{equation*}
\mathfrak{R}\left\{h-h^{*}: h \in \mathcal{H},\left\|h-h^{*}\right\|_{L_{2}}^{2} \leq r\right\} \leq \varphi(r) \tag{A.1.5}
\end{equation*}
$$

and there is a meaningful surrogate function $\varphi_{\text {noise }}(r, \delta)$ that is non-decreasing w.r.t. $r$, and satisfies that $\forall \delta \in(0,1)$, with probability at least $1-\delta$,

$$
\begin{equation*}
\sup _{h \in \mathcal{H},\left\|h-h^{*}\right\|_{L_{2}}^{2} \leq r}\left\{\left(\mathbb{P}-\mathbb{P}_{n}\right)\left[\partial_{1} \ell_{\text {sv }}\left(h^{*}(x), y\right)\left(h-h^{*}\right)\right]\right\} \leq \varphi_{\text {noise }}(r, \delta) . \tag{A.1.6}
\end{equation*}
$$

Given any fixed $\delta \in(0,1)$ and $r_{0} \in\left(0,4 \Delta^{2}\right)$, denote $C_{r_{0}}=2+\left(\frac{16}{c_{\kappa}}+2\right) \log \frac{4 \Delta^{2}}{r_{0}}$. Assume there is a positive integer $\bar{N}_{\delta, r_{0}}$ such that for all $n \geq \bar{N}_{\delta, r_{0}}$,

$$
\begin{equation*}
\varphi_{\text {noise }}\left(8 \Delta^{2} ; \frac{\delta}{C_{r_{0}}}\right) \leq \frac{\alpha\left(4\|\xi\|_{L_{2}} / \sqrt{c_{\kappa}}\right)\|\xi\|_{L_{2}}^{2}}{2} \quad \text { and } \quad \varphi\left(8 \Delta^{2}\right) \leq \frac{\sqrt{2 c_{\kappa}}\|\xi\|_{L_{2}}^{2}}{16 \Delta} \tag{A.1.7}
\end{equation*}
$$

We note that the requirements do not place meaningful restrictions on the choice of surrogate function. The main requirement, condition (A.1.7), asks for uniform errors over $\mathcal{H}$ to be smaller than some fixed values that are independent of $n$. For non-pathological choices of surrogate functions, this will always be satisfied as long as the sample size $n$ is larger than some positive integer $\bar{N}_{\delta, r_{0}}$. The boundedness requirement for $\varphi$ (and $\varphi_{\text {noise }}$ ) can always be met by setting $\varphi(r)=\varphi\left(4 \Delta^{2}\right)$ (and $\left.\varphi_{\text {noise }}(r ; \delta)=\varphi_{\text {noise }}\left(4 \Delta^{2} ; \delta\right)\right)$ for all $r \geq 4 \Delta^{2}$, because $\left\|h-h^{*}\right\|_{L_{2}} \leq 2 \Delta$ for all $h \in \mathcal{H}$.

Theorem 20 (supervised learning with structured convex cost) Let Assumptions 9, 10, 11, 12 hold and $\alpha\left(4\|\xi\|_{L_{2}} / \sqrt{c_{\kappa}}\right)>0$. Let $r_{v e r}^{*}$ be the fixed point of the function

$$
\begin{equation*}
\frac{4}{c_{\kappa} \kappa^{2} \cdot \alpha\left(4\|\xi\|_{L_{2}} / \sqrt{c_{\kappa}}\right)} \varphi_{\text {noise }}\left(2 r ; \frac{\delta}{C_{r_{0}}}\right) . \tag{A.1.8}
\end{equation*}
$$

Given any fixed $\delta \in(0,1)$ and $r_{0} \in\left(0,4 \Delta^{2}\right)$, let $r_{\text {noise }}^{*}$ be the fixed point of the function

$$
\begin{equation*}
\frac{8}{c_{\kappa} K} \sqrt{2 r} \varphi(2 r) \tag{A.1.9}
\end{equation*}
$$

$$
\begin{aligned}
&\left\|\hat{h}-h^{*}\right\|_{L_{2}(\mathbb{P})}^{2} \leq \max \left\{r_{\text {noise }}^{*}, r_{\text {ver }}^{*}, r_{0}\right\} \quad \text { and } \\
& \mathcal{E}(\hat{h}) \leq \frac{\beta_{\mathrm{sv}}}{2} \max \left\{r_{\text {noise }}^{*}, r_{\text {ver }}^{*}, r_{0}\right\},
\end{aligned}
$$

provided that $n>\max \left\{\bar{N}_{\delta, r_{0}}, \frac{72}{c_{\kappa}^{2}} \log \frac{C_{r_{0}}}{\delta}\right\}$.

Remarks. 1) The term $r_{0}$ is negligible since it can be arbitrarily small. One can simply set $r_{0}=1 / n^{4}$, which will be much smaller than $r_{\text {noise }}^{*}$ for typical applications. In high-probability bounds, $C_{r_{0}}$ will only appear in the form $\log \left(C_{r_{0}} / \delta\right)$, which is of a negligible $O(\log \log n)$ order. In the subsequent discussion, we will hide parameters that only depend on $\kappa$ and $c_{\kappa}$ in the big $O$ notation, as they are often absolute constants in practical applications.
2) The two fixed points $r_{\text {noise }}^{*}$ and $r_{\text {ver }}^{*}$ correspond to the two sources of complexities: the uniform errors characterized by the two surrogate functions in (A.1.8) and (A.1.9). Recall that a fundamental limitation of the traditional "local Rademacher complexity analysis" is that it requires a "sub-root" surrogate function that can not differentiate the two sources of complexity. In contrast, the surrogate function in (A.1.8) (which we write as $O(\sqrt{r} \varphi(r))$ for simplicity) is obviously a "super-root" function, thus our analysis overcomes that limitation and provides more precise upper bounds. The key point is that $O(\sqrt{r} \varphi(r))$ is a benign "super-root" surrogate function, in the sense that its fixed point $r_{\text {ver }}^{*}$ is "very small" when the sample size is large enough; in other words, when the problem is learnable. For example, for a $d$-dimensional linear classes, where $\varphi=O(\sqrt{d r / n})$, $r_{\text {ver }}^{*}$ will be the fixed point of $O(d r / n)$. Thus $r_{\text {ver }}^{*}$ will be 0 as long as the sample size $n$ is larger than $O(d)$. Therefore, the typical generalization error derived by Theorem 20 is of order

$$
\mathcal{E}(\hat{h}) \leq \frac{\beta_{\mathrm{sv}}}{2} r_{\text {noise }}^{*}
$$

where $r_{\text {noise }}^{*}$ is the fixed point of the function in (A.1.9). Clearly, $r_{\text {noise }}^{*}$ only depends on the noise multiplier at $h^{*}$ and the local strong convexity parameter, and it is independent of the worst-case
parameters of the cost function.
At a high level, the subscripts "ver" and "noise" have the the meaning of "version space" and "noise multiplier," respectively. Intuitively, $r_{\mathrm{ver}}$ is the estimation error of the noise-free realizable problem, which reflects the complexity of version space-the random subset of $\mathcal{H}$ that consists of all $h$ that agree with $h^{*}$ on $\left\{x_{i}\right\}_{i=1}^{n}$. On the other hand, $r_{\text {noise }}$ is the estimation error induced by the interaction of $\mathcal{H}$ and noise multiplier $\partial_{1} \ell_{\mathrm{sv}}\left(h^{*}(x), y\right)$. We refer to [8] for a more detailed discussion on the source of these two fixed points.

Now we present some representative applications of Theorem 20.

Example 12 (localization of unfavorable parameters) In practical applications, one often wants to avoid the global Lipchitz constant and the inverse of the global strong convexity parameter. For example, in regression with square cost, the global Lipchitz constant is equal to $\Delta_{\infty}$ and is often unbounded, so it is desirable to convert it to $\|\xi\|_{L_{2}}$; and in logistic regression, the inverse of global strong convexity parameter is an exponential constant $e^{O\left(\Delta_{\infty}\right)}$, which we hope to convert to $e^{O\left(\|\xi\| L_{2}\right)}$. These goals are achieved in Theorem 20: since the right hand side of (A.1.8) contains an extra $\sqrt{r}$ factor, $r_{v e r}^{*}$ is typically much smaller than $r_{\text {noise }}^{*}$ for sufficiently large $n$ (see remark 2 after Theorem 20). Therefore, the generalization error bound will be determined by the fixed point $r_{\text {noise }}^{*}$, which only depends on the noise multiplier at $h^{*}$ and the local strong convexity parameter.

Example 13 (regression with heavy-tailed noise) We consider the problem of predicting y using $h(x)$, and allow the "noise" $\xi=h^{*}(x)-y$ to be heavy-tailed. To illustrate the main message of this example, we consider the d-dimensional linear class with sub-Gaussian features. That is, $h(x)=\theta^{T} x$ where $\theta \in \mathbb{R}^{d}$, and the random feature $x \in \mathbb{R}^{d}$ is sub-Gaussian. In this setting, the Huber cost is preferred to the square cost.

- For the Huber cost and truncation parameter $\gamma=O\left(\|\xi\|_{L_{2}}\right)$ in the definition (A.1.1), Theorem 20 implies that the parameter $v$ will be localized to the region where the strong convexity parameter $\alpha(v)$ is non-zero. As a result, the strong convexity parameter in the generalization error bound will be $\frac{1}{2}$ rather than the problematic value 0 (since the generalization
error scales with the inverse of $\alpha(v)$, the value 0 will make the bound vacuous). For the $d$-dimensional linear class, $r_{\text {ver }}^{*}$ will be 0 as long as $n \geq O(d)$. Since $\partial_{1} \ell_{\mathrm{sv}}\left(h^{*}(x), y\right)$ will be uniformly bounded by $O(\sigma)$, we obtain

$$
r_{\text {noise }}^{*} \leq O\left(\frac{\|\xi\|_{L_{2}}^{2}\left(d+\log \frac{1}{\delta}\right)}{n}\right)
$$

which recovers the problem-dependent rate in [9].

- For the square cost, the fixed point $r_{\text {noise }}$ will often cause the generalization error to be suboptimal. For the $d$-dimensional linear class, $r_{\text {noise }}$ will have a polynomial dependence on $1 / \delta$ as explained in [9]. The reason is that in the definition of $\varphi_{\text {noise }}(r, \delta)$ in A.1.6, the noise multiplier $\partial_{1} \ell_{\mathrm{sv}}\left(h^{*}(x), y\right)$ is equal to $\xi$ for the square cost. For "heavy-tailed" $\xi$, this will cause the rate $r_{\text {noise }}$ to be sub-optimal.

We note that the condition that $\hat{h}$ is the empirical risk minimizer is not essential to the proof of Theorem 20. Similar to the prior work [130], we can extend the result to more general learning rules that are based on regularization (e.g., LASSO [131], SLOPE [132], etc.) as follows.

Corollary 21 (extension to general regularized learning rules) Let Assumptions 9 10, 11 hold. Let $\hat{h}$ be the solution of

$$
\begin{equation*}
\min _{\mathcal{H}} \mathbb{P}_{n} \ell_{\mathrm{sv}}(h(x), y)+\Psi(h), \tag{A.1.10}
\end{equation*}
$$

where $\Psi(h)$ is a non-negative regularization term. Let $\mathcal{H}_{0}$ be a subset of $\mathcal{H}$ that is independent of the samples. If inequality (A.1.5) is modified to

$$
\mathfrak{R}\left\{h-h^{*}: h \in \mathcal{H}_{0},\left\|h-h^{*}\right\|_{L_{2}}^{2} \leq r\right\} \leq \varphi(r),
$$

and inequality (A.1.6) is modified to

$$
\sup _{h \in \mathcal{H}_{0},\left\|h-h^{*}\right\|_{L_{2}}^{2} \leq r}\left\{\left(\mathbb{P}-\mathbb{P}_{n}\right)\left[\partial_{1} \ell_{\mathrm{su}}\left(h^{*}(x), y\right)\left(h-h^{*}\right)\right]\right\}+\Psi\left(h^{*}\right) \leq \varphi_{\text {noise }}(r ; \delta),
$$

then under Assumption 12, conditioned on the event $\left\{\hat{h} \in \mathcal{H}_{0}\right\}$, the conclusion of Theorem 20 remains true.

As illustrated in the following example, Corollary 21 is able to recover several important results in the high-dimensional statistics literature.

Example 14 (high-dimensional estimation and LASSO) Consider the linear regression set-up $\mathbb{E}[y \mid x]=x^{T} \theta^{*}$ where $\theta \in \Theta \subseteq \mathbb{R}^{d}, d \gg n$ and $\left\|\theta^{*}\right\|_{0} \leq s \ll d$. Consider the LASSO estimator $\hat{\theta}$, which is the solution of the $\ell_{1}-$ norm regularized risk minimization problem, where the regularization term is $\Phi(h)=\lambda\|\theta\|_{1}$ and $\lambda>0$ is the regularization parameter, i.e.,

$$
\hat{\theta} \in \underset{\Theta}{\arg \min } \mathbb{P}_{n} \ell_{\mathrm{sv}}\left(\theta^{T} x, y\right)+\lambda\|\theta\|_{1} .
$$

Assume $\ell_{\mathrm{sv}}$ is the square cost and $\xi$ is $\sigma-s u b-G a u s s i a n$, or $\ell_{\mathrm{sv}}$ is the Huber cost with truncation parameter $\gamma=O(\sigma)$. Assume the feature $x \in \mathbb{R}^{d}$ is sub-Gaussian. Following standard analysis (see, e.g., Lemma 1 in [133]), by setting $\lambda$ to be of order $\sqrt{\sigma^{2} \log (d / \delta) / n}$, the Lasso estimator $\hat{\theta}$ will lie in a sparse cone $\Theta_{S}$ (with high probability), where it can be proven [134] that $\varphi(r)=$ $O(\sqrt{r s \log d / n})$ and $\varphi_{\text {noise }}(r ; \delta)=O\left(\sqrt{r \sigma^{2} s \log (d / \delta) / n}\right)$ (ignoring dependence on the parameters $C$ and $p$ described in Assumption 11). Applying Corollary 21 with $\mathcal{H}_{0}=\left\{x \mapsto \theta^{T} x: \theta \in \Theta_{S}\right\}$ and $n \geq \Omega(s \log d)$, we have $r_{\text {ver }}^{*}=0$ and

$$
r_{\text {noise }}^{*} \leq O\left(\frac{\sigma^{2} s \log \frac{d}{\delta}}{n}\right) .
$$

## A.1.3 Contributions relative to previous approaches

So far we have recovered the main results in the prior works [8, 9], which are valid for unbounded regression problems and thus improve the traditional "local Rademacher complexity" analysis. Now we would like to illustrate how Theorem 20 improves the results in [8, 9] by removing a "star-shape" requirement. That is, we do not need to assume the hypothesis class is star-shaped/convex, or consider the star-hull of it which may increase complexity.

To be specific, [8, 9] assumes that $\mathcal{H}$ is a convex class (and thus star-shaped). When $\mathcal{H}$ is not star-shaped, the results in $[8,9]$ are still valid by taking the star-hull of $\mathcal{F}$ and considering the local Rademacher/Gaussian complexity of the star-hull. The increase in complexity is quite moderate for traditional hypothesis classes (e.g., those chracterized by covering number conditions; see Lemma 4.6 in [135] for more details). However, taking the star-hull may significantly increase the local Rademacher complexity of modern non-convex and overparameterized classes. Here we show that, even for very simple function classes (e.g., linear classes with non-convex support), our approach improves on what can be achieved using the star-hulls.

Note that the improvement brought by our approach is systematic and may carry over to more complicated learning procedures as well. A more comprehensive comparison with existing localization approaches will be presented after the following example.

Example 15 (overparameterized linear class with growing sparsity) Consider the linear regression model

$$
y \sim N\left(x^{T} \theta^{*}, \sigma^{2}\right), \quad x \sim N\left(0, I_{d \times d}\right),
$$

where $\theta^{*} \in \Theta \subseteq \mathbb{R}^{d}$ and $d \gg n$ (i.e., the model is overoarameterized). Assume the feasible parameter set $\theta$ satisfies that for all $\theta \in \Theta$,

$$
\begin{equation*}
\left\|\theta-\theta^{*}\right\|_{0} \leq\left\lfloor\left\|\theta-\theta^{*}\right\|_{2}^{2}\right\rfloor . \tag{A.1.11}
\end{equation*}
$$

In other words, the sparsity of $\theta$ increases the more $\theta$ deviates from $\theta^{*}$. The maximum likelihood estimation problem corresponds to minimize the empirical average of the square cost with respect to $\mathcal{H}=\left\{x \mapsto x^{T} \theta: \theta \in \Theta\right\}$. For this problem, the surrogate function $\varphi_{\text {noise }}$ need to satisfy (with probability at $1-\delta$ )

$$
\begin{equation*}
\sup _{\theta \in \Theta,\left\|\theta-\theta^{*}\right\|_{2}^{2} \leq r}\left(\mathbb{P}-\mathbb{P}_{n}\right)\left[\xi \cdot x^{T}\left(\theta-\theta^{*}\right)\right] \leq \varphi_{\text {noise }}(r ; \delta) \tag{A.1.12}
\end{equation*}
$$

where the left hand side of (A.1.12) is the localized Gaussian complexity of $\mathcal{H}$. Thanks to the sparsity condition (A.1.11), it can be tightly controlled by

$$
\begin{equation*}
\varphi_{\text {noise }}(r ; \delta)=O\left(\sqrt{\frac{\sigma^{2}\left(\left\|\theta^{*}\right\|_{0}+r\right) r \log \frac{d}{\delta}}{n}}\right)=\underbrace{O\left(\sqrt{\frac{\sigma^{2}\left\|\theta^{*}\right\|_{0} r \log \frac{d}{\delta}}{n}}\right)}_{\text {problem-dependent component }}+\underbrace{O\left(\sqrt{\frac{\sigma^{2} \log \frac{d}{\delta}}{n}} \cdot r\right)}_{\text {benign "super-root" component }} . \tag{A.1.13}
\end{equation*}
$$

Here, the benign "super-root" component" in $\varphi_{\text {noise }}(r ; \delta)$ does not affect the order of its fixed point $r_{\text {noise }}^{*}$ : when $n \geq \Omega\left(4 \sigma^{2} \log \frac{d}{\delta}\right)$, the "super-root" component" in (A.1.13) will be less than $\frac{1}{2} r$ so that $r_{\text {noise }}^{*}$ is of order $\sigma^{2}\left\|\theta^{*}\right\|_{0} \log \frac{d}{\delta} / n$. In other words, only the problem-dependent component in $\varphi_{\text {noise }}(r ; \delta)$ matters.

In contrast, if one takes the star-hull (e.g., expanding $\Theta$ to $\operatorname{star}(\Theta)=\left\{\theta^{*}+\lambda\left(\theta-\theta^{*}\right): \theta \in\right.$ $\Theta, \lambda \in[0,1]\}$, then it is straightforward to verify that $\varphi_{\text {noise }}$ has to be a "sub-root" function. A sub-root surrogate function that governs (A.1.13) will be at least of order

$$
\overline{\varphi_{\text {noise }}}(r ; \delta)=O\left(\sqrt{\frac{\sigma^{2}\left(\left\|\theta^{*}\right\|_{0}+\Delta\right) r \log \frac{d}{\delta}}{n}}\right),
$$

whose fixed point unavoidably scales with the worst-case $L_{2}$ distance $\Delta$. Here we do not consider computational issues, and the key message is that if the complexity (e.g., the "effective dimension") of an overparameterized non-convex class grows very rapidly with respect to its localization scale,
then some "fast growing components" may still be benign and they may not necessarily increase the complexity. It is an open question whether such phenomena manifests in more practical applications.

Comparison with the "small ball method." In a series of pioneering works, Mendelson [8, 9, 30,136 ] proposes the "small ball method" as an alternative approach to the traditional "concentrationcontraction" framework. Under the "small ball" condition, that approach establishes one-sided uniform inequalities through structural results on binary valued indicator functions. Motivated by these works, we seek to refine the traditional concentration framework. Our approach brings added flexibility to concentration by emphasizing the use of surrogate functions that are not "subroot," and relates one-sided uniform inequalities to two-sided concentration of simple "truncated" functions. Following are the main contributions relative to the "small ball method."

First, our approach does not require the hypothesis class to be star-shaped/convex (or to consider the star-hull of the hypothesis class). This improvement is particularly relevant for nonconvex hypothesis classes whose complexity can grow rapidly when "away" from the optimal hypothesis. In Example 15 (and its discussion) we show that the improvement may be meaningful for some non-convex, overparametrized classes; and the phenomenon of "benign fast growing" components in overparameterized models may be of independent interest.

To the best of our knowledge, the "small ball method" cannot overcome the star-shape requirement in a straightforward manner, without additional uniform convergence arguments. The "small ball method" is able to prove one-sided inequalities that hold uniformly over a fixed sphere $\left\{h \in \mathcal{H}:\left\|h-h^{*}\right\|_{L_{2}}^{2}=r\right\}$, and by assuming the class $\mathcal{H}$ to be star-shaped around $h^{*}$, it circumvents the need to have a uniform bound that holds simultaneously for all possible values of $r$. However, without the star-shape assumption and additional uniform convergence arguments, it is not clear how to uniformly extend the bound to all $r$ using peeling. In our analysis, we introduce some new tricks to address this issue. In particular, we use "adaptive truncation levels" and concentration over "rings." Combining these with the "uniform localized convergence" procedure, we
completely circumvent the need for star-hulls (see "Part II" in Appendix A.4.1 for details).
The discussion here is orthogonal to lifting the star-shape/convexity assumptions using aggregation [10], whose primary goal is to remove Assumption 10 (recall that this assumption implicitly asks the hypothesis class to be convex/star-shaped when the model is mis-specified). When using aggregation and improper learning procedures, it is natural to consider the complexity of the enlarged class. Still, we suspect that taking the star-hull may be unnecessary if the enlarged class need not to be star-shaped [30,136], and our analysis may be useful there as well. We note in passing that aggregation procedures are often computationally demanding.

Lastly, the formulation of supervised costs is slightly broader here compared with [9]. In that paper, the loss is assumed to be a univariate function of $(h(x)-y)$, so costs involving the term $y h(x)$ (e.g., the canonical logistic cost and the costs in some other generalized linear models) are not permitted ([9] instead analyzes a modified version of the logistic cost).

Comparison with offset Rademacher complexity. Under the square cost and assuming the socalled "lower isometry bound" as an a priori condition (see Definition 5 in [10]), offset Rademahcer complexity [10] is also able to provide problem-dependent rates. However, establishing such a "lower isometry bound" is typically challenging, so this approach may still need to rely on the "small ball method" (or our analysis) for unbounded regression problems. Moreover, this tool is tailored to the setting of supervised learning with square cost, and it is unclear how to extend the analysis to more general losses.

## Comparison with the "restricted strong convexity" framework in high-dimensional statis-

tics. In the high-dimensional statistics literature, the "restricted strong convexity" framework [133, 3] provides analytical tools to prove problem-dependent rates, but only when such condition is assumed as an a priori (see Definition 2 in [133]). To achieve this, [137, 133, 134] develop a truncation-based analysis that can establish "restricted strong convexity" for sparse kernel regression and sparse generalized linear models. Those works also indicate that one-sided uniform inequalities can be established by two-sided concentration of the "truncated" functions. There are
several differences between their analysis and ours. First, those proofs rely on linearity/star-shape of the hypothesis class and thus only need to prove the "restricted strong convexity" on a fixed sphere (similar to what we have discussed in comparison with the "small-ball method"). In contrast, our framework does not put any geometric restriction on the hypothesis class, by passing this through the use of "adaptive truncation levels" and concentration over "rings," tools that may be of independent interest from a technical perspective. Second, when seeking problem-dependent generalization error bounds, the proposed $L_{2}-L_{4}$ moment equivalence condition [133,134] is stronger than the "small ball" condition used in our analysis. Third, the analysis does not fully localize the strong convexity parameter, and does not cover interesting supervised costs that may have zero curvature, e.g., the Huber cost.

## A. 2 Proofs for Section 1.2 and Section 1.3

In all the proofs we consider a fixed sample size $n$. In order to distinguish "probability of events" and "expectation with respect to $\mathbb{P}$," we will use the notation $\operatorname{Prob}(\mathcal{A})$ to denote the probability of the event $\mathcal{A}$.

## A.2.1 Proofs for Proposition 1 and its variants

We prove a more general version of of Proposition 1. The differences are that 1 ) here we use a more general "peeling scale" $\lambda$ which can be any value larger than 1 , while in Proposition 1 we simply set $\lambda$ to be 2 ; and 2 ) we only ask $\psi(r ; \delta)$ to be a high-probability surrogate function of the uniform error over the "ring" $\{f \in \mathcal{F}: r / \lambda \leq T(f) \leq r\}$ rather than the "bigger" localized area $\{f \in \mathcal{F}: 0 \leq T(f) \leq r\}$.

Proposition 5 (a more general "uniform localized convergence" argument) For a function class $\mathcal{G}=\left\{g_{f}: f \in \mathcal{F}\right\}$ and functional $T: \mathcal{F} \rightarrow[0, R]$, assume there is a function $\psi(r ; \delta)$ (possibly depending on the samples), which is non-decreasing with respect to $r$ and satisfies that
$\forall \delta \in(0,1), \forall r \in[0, R]$, with probability at least $1-\delta$,

$$
\sup _{f \in \mathcal{F}: \frac{r}{\lambda} \leq T(f) \leq r}\left(\mathbb{P}-\mathbb{P}_{n}\right) g_{f} \leq \psi(r ; \delta) .
$$

Then, given any $\delta \in(0,1), r_{0} \in(0, R]$ and $\lambda>1$, with probability at least $1-\delta$, for all $f \in \mathcal{F}$, either $T(f) \leq r_{0}$ or

$$
\left(\mathbb{P}-\mathbb{P}_{n}\right) g_{f} \leq \psi\left(\lambda T(f) ; \delta\left(\log _{\lambda} \frac{\lambda R}{r_{0}}\right)^{-1}\right)
$$

Proof of Proposition 5: we apply a "peeling" technique. Given any $r_{0} \in(0, R]$, take $r_{k}=\lambda^{k} r_{0}$, $k=1, \cdots,\left\lceil\log _{\lambda} \frac{R}{r_{0}}\right\rceil$. Note that $\left\lceil\log _{\lambda} \frac{R}{r_{0}}\right\rceil \leq \log _{\lambda} \frac{\lambda R}{r_{0}}$.

We use a union bound to establish that $\sup _{\frac{r}{\lambda} T(f) \leq r}\left(\mathbb{P}-\mathbb{P}_{n}\right) g_{f} \leq \psi(r ; \delta)$ holds for all these $r_{k}$ simultaneously: $\forall \delta \in(0,1)$, with probability at least $1-\delta$,

$$
\sup _{r_{k-1} \leq T(f) \leq r_{k}}\left(\mathbb{P}-\mathbb{P}_{n}\right) g_{f} \leq \psi\left(r_{k} ; \frac{\delta}{\log _{2} \frac{2 R}{r_{0}}}\right), \quad k=1, \cdots,\left\lceil\log _{2} \frac{R}{r_{0}}\right\rceil .
$$

For any fixed $f \in \mathcal{F}$, if $T(f) \leq r_{0}$ is false, then let $k$ be the non-negative integer such that $\lambda^{k} r_{0}<T(f) \leq \lambda^{k+1} r_{0}$, and we further know that $r_{k+1}=\lambda^{k+1} r_{0} \leq \lambda T(f)$. Therefore, with probability at least $1-\delta$,

$$
\begin{aligned}
\left(\mathbb{P}-\mathbb{P}_{n}\right) g_{f} & \leq \sup _{\tilde{f} \in \mathcal{F}: r_{k} \leq T(\tilde{f}) \leq r_{k+1}}\left(\mathbb{P}-\mathbb{P}_{n}\right) g_{\tilde{f}} \\
& \leq \psi\left(r_{k+1} ; \frac{\delta}{\log _{\lambda} \frac{\lambda R}{r_{0}}}\right) \\
& \leq \psi\left(\lambda T(f) ; \frac{\delta}{\log _{\lambda} \frac{\lambda R}{r_{0}}}\right)
\end{aligned}
$$

Therefore, with probability at least $1-\delta, \forall f \in \mathcal{F}$, either $T(f) \leq r_{0}$ or

$$
\left(\mathbb{P}-\mathbb{P}_{n}\right) g_{f} \leq \psi\left(\lambda T(f) ; \frac{\delta}{\log _{\lambda} \frac{\lambda R}{r_{0}}}\right)
$$

This completes the proof of Proposition 5.

Clearly, Proposition 1 can be viewed as a corollary of Proposition 5. We now present an implication of Proposition 1, which may be more convenient to use for some problems.

Proposition 6 (a variant of the "uniform localized convergence" argument) For a function class $\mathcal{G}=\left\{g_{f}: f \in \mathcal{F}\right\}$ and functional $T: \mathcal{F} \rightarrow[0, R]$, assume there is a function $\psi(r ; \delta)$ (possibly depending on the samples), which is non-decreasing with respect to $r$ and satisfies that $\forall \delta \in(0,1)$, $\forall r \in[0, R]$, with probability at least $1-\delta$,

$$
\sup _{f \in \mathcal{F}: T(f) \leq r}\left(\mathbb{P}-\mathbb{P}_{n}\right) g_{f} \leq \psi(r ; \delta) .
$$

Then, given any $\delta \in(0,1)$ and $r_{0} \in(0, R]$, with probability at least $1-\delta$, for all $f \in \mathcal{F}$,

$$
\left(\mathbb{P}-\mathbb{P}_{n}\right) g_{f} \leq \psi\left(2 T(f) \vee r_{0} ; \frac{\delta}{C_{r_{0}}}\right),
$$

where $C_{r_{0}}=2 \log _{2} \frac{2 R}{r_{0}}$.

Proof of Proposition 6: From Proposition 1 we know that with probability at least $1-\frac{\delta}{2}$, for all $f \in \mathcal{F}$, either $T(f) \leq r_{0}$ or

$$
\begin{equation*}
\left(\mathbb{P}-\mathbb{P}_{n}\right) g_{f} \leq \psi\left(2 T(f) ; \frac{\delta}{2}\left(\log _{2} \frac{2 R}{r_{0}}\right)^{-1}\right)=\psi\left(2 T(f) ; \frac{\delta}{C_{r_{0}}}\right) . \tag{A.2.1}
\end{equation*}
$$

We denote the event

$$
\mathcal{A}_{1}=\left\{\text { there exists } f \in \mathcal{F} \text { such that } T(f) \geq r_{0} \text { and }\left(\mathbb{P}-\mathbb{P}_{n}\right) g_{f}>\psi\left(2 T(f) ; \frac{\delta}{C_{r_{0}}}\right)\right\} .
$$

Then from (A.2.1), we have

$$
\begin{equation*}
\operatorname{Prob}\left(\mathcal{A}_{1}\right) \leq \frac{\delta}{2} \tag{A.2.2}
\end{equation*}
$$

We denote the event

$$
\mathcal{A}_{2}=\left\{\text { there exists } f \in \mathcal{F} \text { such that } T(f)>r_{0} \text { and }\left(\mathbb{P}-\mathbb{P}_{n}\right) g_{f}>\psi\left(r_{0} ; \frac{\delta}{C_{r_{0}}}\right)\right\}
$$

Then from the surrogate property of $\psi$ and the fact $C_{r_{0}} \geq 2$, we have

$$
\begin{equation*}
\operatorname{Prob}\left(\mathcal{A}_{2}\right) \leq \frac{\delta}{C_{r_{0}}} \leq \frac{\delta}{2} \tag{A.2.3}
\end{equation*}
$$

Combining (A.2.2) and (A.2.3) by an union bound, we have

$$
\operatorname{Prob}\left(\mathcal{A}_{1} \cup \mathcal{A}_{2}\right) \leq \operatorname{Prob}\left(\mathcal{A}_{1}\right)+\operatorname{Prob}\left(\mathcal{A}_{2}\right) \leq \delta
$$

From the above argument, it is straightforward to prove that with probability at least $1-\delta$, for all $f \in \mathcal{F}$,

$$
\left(\mathbb{P}-\mathbb{P}_{n}\right) g_{f} \leq \psi\left(2 T(f) \vee r_{0} ; \frac{\delta}{C_{r_{0}}}\right) .
$$

This completes the proof of Proposition 6.

## A.2.2 Proof of Theorem 1

Let $\mathcal{F}$ be the excess loss class in (1.3.2), and define its member $f$ by $f(z)=\ell(h ; z)-$ $\ell\left(h^{*} ; z\right), \forall z \in \mathcal{Z}$. Clearly, $\mathcal{F}$ is uniformly bounded in $[-2 B, 2 B]$. Let $T(f)=\mathbb{P}\left[f^{2}\right]$. Define $\hat{f}$ by $\hat{f}(z)=\ell\left(\hat{h}_{\text {ERM }} ; z\right)-\ell\left(h^{*} ; z\right), \forall z \in \mathcal{Z}$.

For a fixed $r_{0} \in\left(0,4 B^{2}\right)$, Denote $C_{r_{0}}=2 \log _{2} \frac{8 B^{2}}{r_{0}}$. From now to the end of this proof, we will prove the generalization error bound on the event

$$
\begin{equation*}
\mathcal{A}=\left\{\text { for all } f \in \mathcal{F},\left(\mathbb{P}-\mathbb{P}_{n}\right) f \leq \psi\left(2 T(f) \vee r_{0} ; \frac{\delta}{C_{r_{0}}}\right)\right\} . \tag{A.2.4}
\end{equation*}
$$

From Proposition 6 we know that

$$
\operatorname{Prob}(\mathcal{A}) \geq 1-\delta
$$

This means that proving the generalization error bound on the event $\mathcal{A}$ suffices to prove the theorem.

Denote $g(z)=\ell(h ; z)-\inf _{\mathcal{H}} \ell(h ; z)$ and $\hat{g}(z)=\ell\left(\hat{h}_{\text {ERM }} ; z\right)-\inf _{\mathcal{H}} \ell(h ; z)$. Let $T(g)=\mathbb{P}\left[g^{2}\right]$. We have

$$
f(z)=g(z)-\left(\ell\left(h^{*} ; z\right)-\inf _{\mathcal{H}} \ell(h ; z)\right), \quad \forall z,
$$

which implies that

$$
\begin{aligned}
\mathbb{P}\left[f^{2}\right] \leq & 2 \mathbb{P}\left[g^{2}\right]+2 \mathbb{P}\left[\left(\ell\left(h^{*} ; z\right)-\inf _{\mathcal{H}} \ell(h ; z)\right)^{2}\right] \\
& \leq 2 \mathbb{P}\left[g^{2}\right]+4 B \mathcal{L}^{*} \leq 4 \mathbb{P}\left[g^{2}\right] \vee 8 B \mathcal{L}^{*}
\end{aligned}
$$

Therefore, we have

$$
\begin{equation*}
T(\hat{f}) \leq 4 T(\hat{g}) \vee 8 B \mathcal{L}^{*} \tag{A.2.5}
\end{equation*}
$$

From the property of ERM, we have $\mathbb{P}_{n} \hat{f} \leq 0$, which implies that

$$
\begin{equation*}
\mathcal{E}\left(\hat{h}_{\mathrm{ERM}}\right) \leq\left(\mathbb{P}-\mathbb{P}_{n}\right) \hat{f} \leq \psi\left(2 T(\hat{f}) \vee r_{0} ; \frac{\delta}{C_{r_{0}}}\right) \tag{A.2.6}
\end{equation*}
$$

From (A.2.5) and (A.2.6) we have

$$
\begin{equation*}
\mathbb{P} \hat{g}-\mathcal{L}^{*}=\mathcal{E}\left(\hat{h}_{\mathrm{ERM}}\right) \leq \psi\left(8 T(\hat{g}) \vee 16 B \mathcal{L}^{*} \vee r_{0} ; \frac{\delta}{C_{r_{0}}}\right) . \tag{A.2.7}
\end{equation*}
$$

Since $\hat{g}(z) \in[0,2 B]$ for all $z$, we have $T(\hat{g}) \leq 2 B \mathbb{P} \hat{g}$. From this fact and (A.2.7) we obtain

$$
\begin{aligned}
T(\hat{g}) & \leq 2 B \mathbb{P} \hat{g} \\
& \leq 2 B\left(\mathcal{L}^{*}+\psi\left(8 T(\hat{g}) \vee 16 B \mathcal{L}^{*} \vee r_{0} ; \frac{\delta}{C_{r_{0}}}\right)\right) \\
& =2 B \mathcal{L}^{*}+2 B \psi\left(8 T(\hat{g}) \vee 16 B \mathcal{L}^{*} \vee r_{0} ; \frac{\delta}{C_{r_{0}}}\right) .
\end{aligned}
$$

Whether $B \mathcal{L}^{*} \leq 2 B \psi\left(8 T(\hat{g}) \vee 16 B \mathcal{L}^{*} \vee r_{0} ; \frac{\delta}{C_{r_{0}}}\right)$ or $B \mathcal{L}^{*}>2 B \psi\left(8 T(\hat{g}) \vee 16 B \mathcal{L}^{*} \vee r_{0} ; \frac{\delta}{C_{r_{0}}}\right)$, the above inequality always implies that

$$
\begin{align*}
T(\hat{g}) & \leq 3 B \mathcal{L}^{*} \vee 6 B \psi\left(8 T(\hat{g}) \vee 16 B \mathcal{L}^{*} \vee r_{0} ; \frac{\delta}{C_{r_{0}}}\right) \\
& \leq 3 B \mathcal{L}^{*} \vee 6 B \psi\left(8 T(\hat{g}) ; \frac{\delta}{C_{r_{0}}}\right) \vee 6 B \psi\left(16 B \mathcal{L}^{*} \vee r_{0} ; \frac{\delta}{C_{r_{0}}}\right) \tag{A.2.8}
\end{align*}
$$

Let $r^{*}$ be the fixed point of $6 B \psi\left(8 r ; \frac{\delta}{C_{n}}\right)$. From the definition of fixed points whether $2 B \mathcal{L}^{*} \vee \frac{r_{0}}{8} \leq$ $r^{*}$ or $2 B \mathcal{L}^{*} \vee \frac{r_{0}}{8}>r^{*}$, we always have

$$
6 B \psi\left(16 B \mathcal{L}^{*} \vee r_{0} ; \frac{\delta}{C_{r_{0}}}\right) \leq r^{*} \vee 2 B \mathcal{L}^{*} \vee \frac{r_{0}}{8}
$$

Combining the above inequality with (A.2.8), we have

$$
T(\hat{g}) \leq 3 B \mathcal{L}^{*} \vee 6 B \psi\left(8 T(\hat{g}) ; \frac{\delta}{C_{r_{0}}}\right) \vee r^{*} \vee \frac{r_{0}}{8}
$$

From the above inequality and again the definition of fixed points, it is straightforward to prove that

$$
T(\hat{g}) \leq 3 B \mathcal{L}^{*} \vee r^{*} \vee \frac{r_{0}}{8}
$$

Combining the above inequality with (A.2.5), we have

$$
T(\hat{f}) \leq 12 B \mathcal{L}^{*} \vee 4 r^{*} \vee \frac{r_{0}}{2}
$$

From the above inequality and (A.2.6) we have

$$
\begin{equation*}
\mathcal{E}\left(\hat{h}_{\mathrm{ERM}}\right) \leq\left(\mathbb{P}-\mathbb{P}_{n}\right) \hat{f} \leq \psi\left(24 B \mathcal{L}^{*} \vee 8 r^{*} \vee r_{0} ; \frac{\delta}{C_{r_{0}}}\right) \tag{A.2.9}
\end{equation*}
$$

which implies that

$$
\mathcal{E}\left(\hat{h}_{\mathrm{ERM}}\right) \leq \psi\left(24 B \mathcal{L}^{*} ; \frac{\delta}{C_{r_{0}}}\right) \vee \psi\left(8 r^{*} \vee r_{0} ; \frac{\delta}{C_{r_{0}}}\right) .
$$

Recall that $r^{*}$ is the fixed point of $6 B \psi\left(8 r ; \frac{\delta}{C_{r_{0}}}\right)$. Since $r^{*} \vee \frac{r_{0}}{8} \geq r^{*}$, from the definition of fixed points we have

$$
6 B \psi\left(8 r^{*} \vee 2 r_{0} ; \frac{\delta}{C_{r_{0}}}\right) \leq r^{*} \vee \frac{r_{0}}{8}
$$

So we finally obtain

$$
\mathcal{E}\left(\hat{h}_{\mathrm{ERM}}\right) \leq \psi\left(24 B \mathcal{L}^{*} ; \frac{\delta}{C_{r_{0}}}\right) \vee \frac{r^{*}}{6 B} \vee \frac{r_{0}}{48 B} .
$$

Recall that the generalization error bound holds true on the event $\mathcal{A}$ defined in (A.2.4), whose measure is at least $1-\delta$. This completes the proof.

## A.2.3 Estimating loss-dependent rates from data

In the remarks following Theorem 1, we comment that fully data-dependent loss-dependent bounds can be derived using the empirical "effective $\operatorname{loss},{ }^{,} \mathbb{P}_{n}\left[\ell\left(\hat{h}_{\mathrm{ERM}} ; z\right)-\inf _{\mathcal{H}} \ell(h ; z)\right]$ to estimate the unknown parameter $\mathcal{L}^{*}$. Here we present the full details and some discussion of this approach.

Theorem 22 (estimate of the loss-dependent rate from data) Recall the term $\mathcal{L}^{*}$ is $\mathbb{P}\left[\ell\left(h^{*} ; z\right)\right.$ $\left.\inf _{\mathcal{H}} \ell\left(h^{*} ; z\right)\right]$ and denote $\widehat{\mathcal{L}^{*}}=\mathbb{P}_{n}\left[\ell\left(\hat{h}_{\mathrm{ERM}} ; z\right)-\inf _{\mathcal{H}} \ell(h ; z)\right]$. Under the conditions of Theorem 1, setting $C_{n}=2 \log _{2} n+6$, then for any fixed $\delta \in\left(0, \frac{1}{2}\right)$, with probability at least $1-2 \delta$, we have

$$
\begin{equation*}
\mathcal{E}\left(\hat{h}_{\mathrm{ERM}}\right) \leq \psi\left(c B \widehat{\mathcal{L}^{*}} ; \frac{\delta}{C_{n}}\right) \vee \frac{c r^{*}}{B} \vee \frac{c B \log \frac{2}{\delta}}{n} \tag{A.2.10}
\end{equation*}
$$

and

$$
\begin{equation*}
\mathcal{L}^{*} \leq c_{1}\left(\widehat{\mathcal{L}^{*}} \vee \frac{r^{*}}{B} \vee \frac{B \log \frac{2}{\delta}}{n}\right) \leq c_{2}\left(\mathcal{L}^{*} \vee \frac{r^{*}}{B} \vee \frac{B \log \frac{2}{\delta}}{n}\right), \tag{A.2.11}
\end{equation*}
$$

where $c, c_{1}, c_{2}$ are absolute constants.

Remarks. 1) The $B \log \frac{2}{\delta} / n$ terms (A.2.10) and (A.2.11) are negligible, because $r^{*}$ is at least of order $B^{2} \log \frac{1}{\delta} / n$ for most practical applications. This order is unavoidable in traditional "local Rademacher complexity" analysis and two-sided concentration inequalities.
2) The generalization error bound (A.2.10) shows that without knowledge of $\mathcal{L}^{*}$, one can estimate the order of our loss-dependent rate by using $\widehat{\mathcal{L}^{*}}=\mathbb{P}_{n}\left[\ell\left(\hat{h}_{\text {ERM }} ; z\right)-\inf _{\mathcal{H}} \ell(h ; z)\right]$ as a proxy. Despite replacing $\mathcal{L}^{*}$ by $\widehat{\mathcal{L}^{*}}$, other quantities in the bound remain unchanged in order.
3) The inequality (A.2.11) shows that the estimation of $\mathcal{L}^{*}$ is tight.

Proof of Theorem 22: from the definitions, we know that $\mathcal{L}^{*}=\mathbb{P}\left[\ell\left(h^{*} ; z\right)-\inf _{\mathcal{H}} \ell\left(h^{*} ; z\right)\right]$, $\widehat{\mathcal{L}^{*}}=\mathbb{P}_{n}\left[\ell\left(\hat{h}_{\mathrm{ERM}} ; z\right)-\inf _{\mathcal{H}} \ell(h ; z)\right]$ and $\mathbb{P} \ell\left(h^{*} ; z\right) \leq \mathbb{P} \ell\left(\hat{h}_{\mathrm{ERM}} ; z\right)$. As a result, we have

$$
\begin{align*}
& \mathcal{L}^{*}-\widehat{\mathcal{L}^{*}}=\mathbb{P} \ell\left(h^{*} ; z\right)-\mathbb{P}_{n} \ell\left(\hat{h}_{\mathrm{ERM}} ; z\right)-\left(\mathbb{P}-\mathbb{P}_{n}\right)\left[\inf _{\mathcal{H}} \ell(h ; z)\right] \\
&\left.\leq\left(\mathbb{P}-\mathbb{P}_{n}\right) \ell\left(\hat{h}_{\mathrm{ERM}} ; z\right)-\left(\mathbb{P}-\mathbb{P}_{n}\right) \inf _{\mathcal{H}} \ell(h ; z)\right] \\
&=\left(\mathbb{P}-\mathbb{P}_{n}\right) \hat{f}+\left(\mathbb{P}-\mathbb{P}_{n}\right)\left[\ell\left(h^{*} ; z\right)-\inf _{\mathcal{H}} \ell(h ; z)\right], \tag{A.2.12}
\end{align*}
$$

where $\hat{f}$ is defined by $\hat{f}(z)=\ell\left(\hat{h}_{\mathrm{ERM}} ; z\right)-\ell\left(h^{*} ; z\right), \forall z \in \mathcal{Z}$.

We take $r_{0}=\frac{B^{2}}{n}$ in Theorem 1, and denote $C_{n}:=C_{r_{0}}=2 \log _{2} n+6$. From (A.2.9) in the proof of Theorem 1, on the event $\mathcal{A}$ defined in (A.2.4) (whose measure is at least $1-\delta$ ),

$$
\begin{equation*}
\mathcal{E}\left(\hat{h}_{\mathrm{ERM}}\right) \leq\left(\mathbb{P}-\mathbb{P}_{n}\right) \hat{f} \leq \psi\left(24 B \mathcal{L}^{*} \vee 8 r^{*} \vee \frac{B^{2}}{n} ; \frac{\delta}{C_{n}}\right) \tag{A.2.13}
\end{equation*}
$$

where $\hat{f}$ is defined by $\hat{f}(z)=\ell\left(\hat{h}_{\mathrm{ERM}} ; z\right)-\ell\left(h^{*} ; z\right), \forall z \in \mathcal{Z}$.
Since $3 B \mathcal{L}^{*} \vee r^{*} \vee \frac{B^{2}}{4 n} \geq r^{*}$, from the definition of fixed points we have

$$
\begin{align*}
(\mathbb{P}- & \left.\mathbb{P}_{n}\right) \hat{f} \leq \psi\left(8\left(3 B \mathcal{L}^{*} \vee r^{*} \vee \frac{B^{2}}{8 n}\right) ; \frac{\delta}{C_{n}}\right) \\
& \leq \frac{3 B \mathcal{L}^{*} \vee r^{*} \vee \frac{B^{2}}{8 n}}{6 B} \leq \frac{\mathcal{L}^{*}}{2}+\frac{r^{*}}{6 B}+\frac{B}{48 n} . \tag{A.2.14}
\end{align*}
$$

This result holds together with the result of Theorem 1 on the event $\mathcal{A}$.
The random variable $\ell\left(h^{*} ; z\right)-\inf _{\mathcal{H}} \ell(h ; z)$ is uniformly bounded by [0,2B]. From Bernstein's inequality and the fact $\operatorname{Var}\left[\ell\left(h^{*} ; z\right)-\inf _{\mathcal{H}} \ell(h ; z)\right] \leq 2 B \mathcal{L}^{*}$, with probability at least $1-\delta$,

$$
\begin{equation*}
\left|\left(\mathbb{P}-\mathbb{P}_{n}\right)\left[\ell\left(h^{*} ; z\right)-\inf _{\mathcal{H}} \ell(h ; z)\right]\right| \leq \sqrt{\frac{4 B \mathcal{L}^{*} \log \frac{2}{\delta}}{n}}+\frac{2 B \log \frac{2}{\delta}}{n} \leq \frac{\mathcal{L}^{*}}{4}+\frac{3 B \log \frac{2}{\delta}}{n} . \tag{A.2.15}
\end{equation*}
$$

Consider the event

$$
\mathcal{A}_{3}=\mathcal{A} \cup\{\text { inequality (A.2.15) holds true }\}
$$

whose measure is at least $1-2 \delta$. On the event $\mathcal{A}_{3}$, from inequalities (A.2.12) (A.2.14) (A.2.15), it is straightforward to show that

$$
\mathcal{L}^{*}-\widehat{\mathcal{L}^{*}} \leq \frac{3}{4} \mathcal{L}^{*}+\frac{r^{*}}{6 B}+\frac{4 B \log \frac{2}{\delta}}{n}
$$

which implies

$$
\begin{equation*}
\mathcal{L}^{*} \leq 4 \widehat{\mathcal{L}^{*}}+\frac{2 r^{*}}{3 B}+\frac{16 B \log \frac{2}{\delta}}{n} \tag{A.2.16}
\end{equation*}
$$

From this result and (A.2.13), it is straightforward to show that

$$
\mathcal{E}\left(\hat{h}_{\mathrm{ERM}}\right) \leq \psi\left(c B \widehat{\mathcal{L}^{*}} ; \frac{\delta}{C_{n}}\right) \vee \frac{c r^{*}}{n} \vee \frac{c B \log \frac{2}{\delta}}{n},
$$

where $c$ is an absolute constant.
We also have

$$
\begin{aligned}
& \widehat{\mathcal{L}^{*}}-\mathcal{L}^{*}=\mathbb{P}_{n} \ell\left(\hat{h}_{\mathrm{ERM}}\right)-\mathbb{P} \ell\left(h^{*} ; z\right)-\left(\mathbb{P}_{n}-\mathbb{P}\right)\left[\inf _{\mathcal{H}} \ell(h ; z)\right] \\
& \leq\left(\mathbb{P}_{n}-\mathbb{P}\right) \ell\left(h^{*} ; z\right)-\left(\mathbb{P}_{n}-\mathbb{P}\right)\left[\inf _{\mathcal{H}} \ell(h ; z)\right] \\
&=\left(\mathbb{P}_{n}-\mathbb{P}\right)\left[\ell\left(h^{*} ; z\right)-\inf _{\mathcal{H}} \ell(h ; z)\right] .
\end{aligned}
$$

From this result and (A.2.15), on the event $\mathcal{A}_{3}$,

$$
\begin{equation*}
\widehat{\mathcal{L}^{*}} \leq \frac{5}{4} \mathcal{L}^{*}+\frac{3 B \log \frac{2}{\delta}}{n} \tag{A.2.17}
\end{equation*}
$$

Combine (A.2.16) and (A.2.17) we obtain

$$
\mathcal{L}^{*} \leq c_{1}\left(\widehat{\mathcal{L}^{*}} \vee \frac{r^{*}}{B} \vee \frac{B \log \frac{2}{\delta}}{n}\right) \leq c_{2}\left(\mathcal{L}^{*} \vee \frac{r^{*}}{B} \vee \frac{B \log \frac{2}{\delta}}{n}\right)
$$

where $c_{1}$ and $c_{2}$ are absolute constants. This completes the proof.

## A.2.4 Proof of Theorem 2

The main goal of this subsection is to prove Theorem 2. We first prove Theorem 23 (the bound (1.3.6) in the main paper), a guarantee for the second-stage moment penalized estimator $\hat{h}_{\mathrm{MP}}$. In order to prove Theorem 2, we then combine Theorem 23 with a guarantee for the first-stage empirical risk minimization (ERM) estimator.

## Analysis for the second-stage moment-penalized estimator

Theorem 23 (variance-dependent rate of the second-stage estimator) Given arbitrary preliminary estimate $\widehat{\mathcal{L}_{0}^{*}} \in[-B, B]$, the generalization error of the moment-penalized estimator $\hat{h}_{\mathrm{MP}}$ in Strategy 2 is bounded by

$$
\mathcal{E}\left(\hat{h}_{\mathrm{MP}}\right) \leq 2 \psi\left(c_{0}\left[\mathcal{V}^{*} \vee\left(\widehat{\mathcal{L}_{0}^{*}}-\mathcal{L}_{0}^{*}\right)^{2} \vee r^{*}\right] ; \frac{\delta}{C_{n}}\right),
$$

with probability at least $1-\delta$, where $c_{0}$ is an absolute constant and $r^{*}$ is the fixed point of $16 B \psi\left(r ; \frac{\delta}{C_{n}}\right)$.

Proof of Theorem 23: the proof of Theorem 23 consist of four parts.

Part I: use $\psi$ to upper bound localized empirical processes. Let $\mathcal{F}$ be the excess loss class in (1.3.2), and define its member $f$ is defined by $f(z)=\ell(h ; z)-\ell\left(h^{*} ; z\right), \forall z \in \mathcal{Z}$. We have the following lemma.

Lemma 14 (bound on localized empirical processes) Given a fixed $\delta_{1} \in(0,1)$, let $r_{1}^{*}\left(\delta_{1}\right)$ be the fixed point of $16 B \psi\left(r ; \delta_{1}\right)$ where $\psi$ is defined in Strategy 2. Then with probability at least $1-\delta_{1}$, for all $r>0$,

$$
\begin{equation*}
\sup _{\mathbb{P}\left[f^{2}\right] \leq r}\left(\mathbb{P}-\mathbb{P}_{n}\right) f \leq \psi\left(r \vee r_{1}^{*}\left(\delta_{1}\right) ; \delta_{1}\right) . \tag{A.2.18}
\end{equation*}
$$

Proof of Lemma 14: clearly, $\mathcal{F}$ is uniformly bounded in $[-2 B, 2 B]$. When $\mathbb{P}\left[f^{2}\right] \leq r$, we have $\mathbb{P}\left[f^{4}\right] \leq 4 B^{2} r$. From Lemma 16 (the two-sided version of its second inequality), with probability
at least $1-\frac{\delta_{1}}{2}$,

$$
\begin{aligned}
& \sup _{\mathbb{P}\left[f^{2}\right] \leq r}\left|\left(\mathbb{P}-\mathbb{P}_{n}\right) f^{2}\right| \\
\leq & 4 \Re_{n}\left\{f^{2}: \mathbb{P}\left[f^{2}\right] \leq r\right\}+2 B \sqrt{\frac{2 r \log \frac{8}{\delta_{1}}}{n}}+\frac{18 B^{2} \log \frac{8}{\delta_{1}}}{n} \\
\leq & 16 B \Re_{n}\left\{f: \mathbb{P}\left[f^{2}\right] \leq r\right\}+2 B \sqrt{\frac{2 r \log \frac{8}{\delta_{1}}}{n}}+\frac{18 B^{2} \log \frac{8}{\delta_{1}}}{n}
\end{aligned}
$$

where the last inequality follows from the Lipchitz contraction property of Rademahcer complexity (see, e.g., Theorem 7 in [138]), and the fact that for all $f_{1}, f_{2} \in \mathcal{F},\left|f_{1}^{2}(z)-f_{2}^{2}(z)\right| \leq 4 B \mid f_{1}(z)-$ $f_{2}(z) \mid$. We conclude that with probability at least $1-\frac{\delta_{1}}{2}$,

$$
\begin{equation*}
\sup _{\mathbb{P}\left[f^{2}\right] \leq r}\left|\left(\mathbb{P}-\mathbb{P}_{n}\right) f^{2}\right| \leq \varphi_{\delta_{1}}(r), \tag{A.2.19}
\end{equation*}
$$

where $\varphi_{\delta_{1}}(r):=16 B \Re_{n}\left\{f: \mathbb{P}\left[f^{2}\right] \leq r\right\}+2 B \sqrt{\frac{2 r \log \frac{8}{\delta_{1}}}{n}}+\frac{18 B^{2} \log \frac{8}{\delta_{1}}}{n}$.
Denote $r_{2}^{*}\left(\delta_{1}\right)$ the fixed point of $4 \varphi_{\delta_{1}}(r)$ (the fixed point must exist as $4 \varphi_{\delta_{1}}(r)$ is a nondecreasing, non-negative and bounded function). From (A.2.19) and the fact that $r_{2}^{*}\left(\delta_{1}\right)$ is the fixed point of $4 \varphi_{\delta_{1}}(r)$, if $r>r_{2}^{*}\left(\delta_{1}\right)$, then with probability at least $1-\frac{\delta_{1}}{2}$,

$$
\begin{equation*}
\sup _{\mathbb{P}\left[f^{2}\right] \leq r}\left|\left(\mathbb{P}-\mathbb{P}_{n}\right) f^{2}\right| \leq \frac{r}{4} \tag{A.2.20}
\end{equation*}
$$

(A.2.20) implies that with probability at least $1-\frac{\delta_{1}}{2}$, for all $r>r_{2}^{*}\left(\delta_{1}\right), \mathbb{P}\left[f^{2}\right] \leq r$ implies that

$$
\begin{equation*}
\mathbb{P}_{n}\left[f^{2}\right] \leq \frac{5}{4} r \leq 2 r . \tag{A.2.21}
\end{equation*}
$$

Again from the two-sided version of the second inequality in Lemma 16, we know that with
probability at least $1-\frac{\delta_{1}}{2}$,

$$
\sup _{\mathbb{P}\left[f^{2}\right] \leq r}\left|\left(\mathbb{P}-\mathbb{P}_{n}\right) f\right| \leq 4 \mathfrak{R}_{n}\left\{f: \mathbb{P}\left[f^{2}\right] \leq r\right\}+\sqrt{\frac{2 r \log \frac{8}{\delta_{1}}}{n}}+\frac{9 B \log \frac{8}{\delta_{1}}}{n}
$$

Combining the above inequality and (A.2.21) using a union bound, we know that with probability at least $1-\frac{\delta_{1}}{2}-\frac{\delta_{1}}{2}=1-\delta_{1}$, if $r>r_{2}^{*}\left(\delta_{1}\right)$, then

$$
\begin{array}{r}
\sup _{\mathbb{P}\left[f^{2}\right] \leq r}\left(\mathbb{P}-\mathbb{P}_{n}\right) f \leq 4 \Re_{n}\left\{f: \mathbb{P}\left[f^{2}\right] \leq r\right\}+\sqrt{\frac{2 r \log \frac{8}{\delta_{1}}}{n}}+\frac{9 B \log \frac{8}{\delta_{1}}}{n} \\
\leq 4 \Re_{n}\left\{f: \mathbb{P}_{n}\left[f^{2}\right] \leq 2 r\right\}+\sqrt{\frac{2 r \log \frac{8}{\delta_{1}}}{n}}+\frac{9 B \log \frac{8}{\delta_{1}}}{n} \tag{A.2.22}
\end{array}
$$

Recall that the $\psi$ function satisfies that $\forall r>0$,

$$
4 \Re_{n}\left\{f: \mathbb{P}_{n}\left[f^{2}\right] \leq 2 r\right\}+\sqrt{\frac{2 r \log \frac{8}{\delta_{1}}}{n}}+\frac{9 B \log \frac{8}{\delta_{1}}}{n} \leq \psi\left(r ; \delta_{1}\right) .
$$

From this fact and (A.2.22), we see that with probability at least $1-\delta_{1}$, for all $r>0$,

$$
\begin{equation*}
\sup _{\mathbb{P}\left[f^{2}\right] \leq r}\left(\mathbb{P}-\mathbb{P}_{n}\right) f \leq \psi\left(r \vee r_{2}^{*}\left(\delta_{1}\right) ; \delta_{1}\right) \tag{A.2.23}
\end{equation*}
$$

From (A.2.23), in order to prove the result (A.2.18) in Lemma 14, we only need to prove that

$$
\begin{equation*}
r_{2}^{*}\left(\delta_{1}\right) \leq r_{1}^{*}\left(\delta_{1}\right) \tag{A.2.24}
\end{equation*}
$$

Assume this is not true, i.e. $r_{2}^{*}\left(\delta_{1}\right)>r_{1}^{*}\left(\delta_{1}\right)$. Since $r_{1}^{*}\left(\delta_{1}\right)$ is the fixed point of $16 B \psi\left(r ; \delta_{1}\right)$, from the definition of fixed points we have

$$
r_{2}^{*}\left(\delta_{1}\right)>16 B \psi\left(r_{2}^{*}\left(\delta_{1}\right) ; \delta_{1}\right)
$$

From the definitions of $\psi$ and $\varphi_{\delta_{1}}$, for all $r>r_{1}^{*}\left(\delta_{1}\right)$,

$$
4 \varphi_{\delta_{1}}(r) \leq 16 B \psi\left(r ; \delta_{1}\right) .
$$

From the above two inequalities and $r_{2}^{*}\left(\delta_{1}\right)>r_{1}^{*}\left(\delta_{1}\right)$, we have

$$
\begin{equation*}
r_{2}^{*}\left(\delta_{1}\right)>16 B \psi\left(r_{2}^{*}\left(\delta_{1}\right) ; \delta_{1}\right) \geq 4 \varphi_{\delta_{1}}\left(r_{2}^{*}\left(\delta_{1}\right)\right) \tag{A.2.25}
\end{equation*}
$$

From the fact that $r_{2}^{*}\left(\delta_{1}\right)$ is the fixed point of $4 \varphi_{\delta_{1}}$, we have

$$
\begin{equation*}
4 \varphi_{\delta_{1}}\left(r_{2}^{*}\left(\delta_{1}\right)\right)=r_{2}^{*}\left(\delta_{1}\right) \tag{A.2.26}
\end{equation*}
$$

The above two inequalities (A.2.25) and (A.2.26) result in a contradiction. So the assumption $r_{2}^{*}\left(\delta_{1}\right)>r_{1}^{*}\left(\delta_{1}\right)$ is false. Therefore $r_{2}^{*}\left(\delta_{1}\right) \leq r_{1}^{*}\left(\delta_{1}\right)$, and this completes the proof of Lemma 14.

## Part II: a "uniform localized convergence" argument with data-dependent measurement.

Based on Lemma 14, we will modify the proof of Proposition 1 to obtain a "uniform localized convergence" argument with the data-dependent "measurement" functional $\mathbb{P}_{n}\left[f^{2}\right]$.

## Lemma 15 (a "uniform localized convergence" argument with the data-dependent "measurement" functio

Given a fixed $\delta_{1} \in(0,1)$, let $r_{1}^{*}\left(\delta_{1}\right)$ be the fixed point of $16 B \psi\left(r ; \delta_{1}\right)$ where $\psi$ is defined in Strategy 2. Then with probability at least $1-\left(\log _{2} \frac{8 B^{2} v 2 r_{1}^{*}\left(\delta_{1}\right)}{r_{1}^{*}\left(\delta_{1}\right)}+\frac{1}{2}\right) \delta_{1}$, for all $f \in \mathcal{F}$ either $\mathbb{P}\left[f^{2}\right] \leq r_{1}^{*}\left(\delta_{1}\right)$, or

$$
\begin{equation*}
\left(\mathbb{P}-\mathbb{P}_{n}\right) f \leq \psi\left(4 \mathbb{P}_{n}\left[f^{2}\right] ; \delta_{1}\right) \tag{A.2.27}
\end{equation*}
$$

Proof of Lemma 15: from the definition of $\psi$ and the fact that $r_{1}^{*}\left(\delta_{1}\right)$ is the fixed point of $16 B \psi\left(r ; \delta_{1}\right)$, we know that $r_{1}^{*}\left(\delta_{1}\right) \geq \frac{144 B^{2} \log \frac{8}{\delta_{1}}}{n}>0$. Take $r_{0}=r_{1}^{*}\left(\delta_{1}\right)$.

Take $R=4 B^{2} \vee r_{0}$ to be a uniform upper bound for $\mathbb{P} f^{2}$, and take $r_{k}=2^{k} r_{0}, k=1, \cdots,\left\lceil\log _{2} \frac{R}{r_{0}}\right\rceil$. Note that $\left\lceil\log _{2} \frac{R}{r_{0}}\right\rceil \leq \log _{2} \frac{2 R}{r_{0}}$. We use the union bound to establish that $\sup _{\mathbb{P}\left[f^{2}\right] \leq r}\left(\mathbb{P}-\mathbb{P}_{n}\right) f \leq$
$\psi\left(r ; \delta_{1}\right)$ holds for all $\left\{r_{k}\right\}$ simultaneously: with probability at least $1-\log _{2} \frac{2 R}{r_{0}} \delta_{1}$,

$$
\sup _{\mathbb{P}\left[f^{2}\right] \leq r_{k}}\left(\mathbb{P}-\mathbb{P}_{n}\right) f \leq \psi\left(r_{k} ; \delta_{1}\right), \quad k=1, \cdots,\left\lceil\log _{2} \frac{R}{r_{0}}\right\rceil .
$$

For any fixed $f \in \mathcal{F}$, if $\mathbb{P}\left[f^{2}\right] \leq r_{0}$ is false, let $k$ be the non-negative integer such that $2^{k} r_{0}<\mathbb{P}\left[g(h ; z)^{2}\right] \leq 2^{k+1} r_{0}$. We further have that $r_{k+1}=2^{k+1} r_{0} \leq 2 \mathbb{P}\left[f^{2}\right]$. Therefore, with probability at least $1-\log _{2} \frac{2 R}{r_{0}} \delta_{1}$,

$$
\begin{align*}
\mathbb{P} f & \leq \mathbb{P}_{n} f+\sup _{\tilde{f} \in \mathcal{F}: \mathbb{P}\left[f^{2}\right] \leq r_{k+1}}\left(\mathbb{P}-\mathbb{P}_{n}\right) \tilde{f} \\
& \leq \mathbb{P}_{n} f+\psi\left(r_{k+1} ; \delta_{1}\right) \tag{A.2.28}
\end{align*}
$$

By (A.2.19) we know that with probability at least $1-\frac{\delta_{1}}{2}$,

$$
\sup _{\mathbb{P}\left[f^{2}\right] \leq r}\left(\mathbb{P}\left[f^{2}\right]-\mathbb{P}_{n}\left[f^{2}\right]\right) \leq \frac{r}{4}
$$

for all $r>r_{0}$ (here we have used the fact $r_{0}=r_{1}^{*}\left(\delta_{1}\right) \geq r_{2}^{*}\left(\delta_{1}\right)$, which is the result (A.2.24) in the proof of Lemma 14). From the union bound, with probability at least $1-\left(\log _{2} \frac{2 R}{r_{0}}+\frac{1}{2}\right) \delta_{1}$, the condition $r_{k+1} \geq \mathbb{P}\left[f^{2}\right]>r_{k}$ will imply

$$
\mathbb{P}_{n}\left[f^{2}\right] \geq \mathbb{P}\left[f^{2}\right]-\frac{1}{4} r_{k+1} \geq \frac{1}{4} r_{k+1}
$$

so

$$
r_{k+1} \leq 4 \mathbb{P}_{n}\left[f^{2}\right]
$$

Combining this result with (A.2.28), we have that for all $f$ such that $T(f)>r_{0}$, with probability
at least $1-\left(\log _{2} \frac{2 R}{r_{0}}+\frac{1}{2}\right) \delta_{1}$,

$$
\begin{aligned}
\mathbb{P} f & \leq \mathbb{P}_{n} f+\psi\left(r_{k+1} ; \delta_{1}\right) \\
& \leq \mathbb{P}_{n} f+\psi\left(4 \mathbb{P}_{n}\left[f^{2}\right] ; \delta_{1}\right) .
\end{aligned}
$$

We conclude that with probability at least $1-\left(\log _{2} \frac{2 R}{r_{0}}+\frac{1}{2}\right) \delta_{1}$, for all $f \in \mathcal{F}$, either $\mathbb{P}\left[f^{2}\right] \leq r_{1}^{*}\left(\delta_{1}\right)$, or

$$
\left(\mathbb{P}-\mathbb{P}_{n}\right) f \leq \psi\left(4 \mathbb{P}_{n}\left[f^{2}\right] ; \delta_{1}\right)
$$

This completes the proof of Lemma 15.

Part III: specify the moment-penalized estimator and its error bound.
We define the event

$$
\mathcal{A}_{1}=\left\{\text { there exists } f \in \mathcal{F} \text { such that } \mathbb{P}\left[f^{2}\right] \geq r_{0} \text { and }\left(\mathbb{P}-\mathbb{P}_{n}\right) f>\psi\left(4 \mathbb{P}_{n}\left[f^{2}\right] ; \delta_{1}\right)\right\} .
$$

Lemma 15 has proven that

$$
\begin{equation*}
\operatorname{Prob}\left(\mathcal{A}_{1}\right) \leq\left(\log _{2} \frac{8 B^{2} \vee 2 r_{1}^{*}\left(\delta_{1}\right)}{r_{1}^{*}\left(\delta_{1}\right)}+\frac{1}{2}\right) \delta_{1} \tag{A.2.29}
\end{equation*}
$$

We denote the event

$$
\mathcal{A}_{2}=\left\{\text { there exists } f \in \mathcal{F} \text { such that } \mathbb{P}\left[f^{2}\right] \leq r_{0} \text { and }\left(\mathbb{P}-\mathbb{P}_{n}\right) f>\psi\left(r_{0} ; \delta_{1}\right)\right\} .
$$

Due to the surrogate property of $\psi$, we have

$$
\begin{equation*}
\operatorname{Prob}\left(\mathcal{A}_{2}\right) \leq \delta_{1} \tag{A.2.30}
\end{equation*}
$$

## Denote the event

$$
\mathcal{A}=\left\{\text { for all } f \in \mathcal{F},\left(\mathbb{P}-\mathbb{P}_{n}\right) f \leq \psi\left(4 \mathbb{P}_{n}\left[f^{2}\right] \vee r_{1}^{*}\left(\delta_{1}\right) ; \delta_{1}\right)\right\} .
$$

From (A.2.29) and (A.2.30), it is straightforward to prove that

$$
\begin{align*}
\operatorname{Prob}(\mathcal{A}) & \geq 1-\operatorname{Prob}\left(\mathcal{A}_{1}\right)-\operatorname{Prob}\left(\mathcal{A}_{2}\right) \\
& \geq 1-\left(\log _{2} \frac{8 B^{2} \vee 2 r_{1}^{*}\left(\delta_{1}\right)}{r_{1}^{*}\left(\delta_{1}\right)}+\frac{1}{2}\right) \delta_{1}-\delta_{1} \\
& \geq 1-\left(\log _{2} \frac{8 B^{2} \vee 2 r_{1}^{*}\left(\delta_{1}\right)}{r_{1}^{*}\left(\delta_{1}\right)}+\frac{3}{2}\right) \delta_{1} . \tag{A.2.31}
\end{align*}
$$

Since $f$ is the excess loss. we can equivalently write the event $\mathcal{A}$ as

$$
\begin{equation*}
\mathcal{A}=\left\{\text { for all } h \in \mathcal{H}, \mathcal{E}(h) \leq \mathbb{P}_{n}\left[\ell(h ; z)-\ell\left(h^{*} z\right)\right]+\psi\left(4 \mathbb{P}_{n}\left[\left(\ell(h ; z)-\ell\left(h^{*} z\right)\right)^{2}\right] \vee r_{1}^{*}\left(\delta_{1}\right) ; \delta_{1}\right)\right\} . \tag{A.2.32}
\end{equation*}
$$

Denote $w(h ; z)=\ell(h ; z)-\widehat{\mathcal{L}_{0}^{*}}$, we have

$$
\begin{array}{r}
4 \mathbb{P}_{n}\left[\left(\ell(h ; z)-\ell\left(h^{*} z\right)\right)^{2}\right] \leq 8 \mathbb{P}_{n}\left[w(h ; z)^{2}\right]+8 \mathbb{P}_{n}\left[w\left(h^{*} ; z\right)^{2}\right] \\
\leq 16 \mathbb{P}_{n}\left[w(h ; z)^{2}\right] \vee 16 \mathbb{P}_{n}\left[w\left(h^{*} ; z\right)^{2}\right] .
\end{array}
$$

From the above conclusion and (A.2.32), we obtain that on the event $\mathcal{A}$,

$$
\begin{array}{r}
\mathcal{E}(h)+\mathbb{P}_{n} \ell\left(h^{*} ; z\right) \leq \mathbb{P}_{n} \ell(h ; z)+\psi\left(4 \mathbb{P}_{n}\left[\left(\ell(h ; z)-\ell\left(h^{*} ; z\right)\right)^{2}\right] \vee r_{1}^{*}\left(\delta_{1}\right) ; \delta_{1}\right) \\
\leq \mathbb{P}_{n}(h ; z)+\psi\left(16 \mathbb{P}_{n}\left[w(h ; z)^{2}\right] \vee 16 \mathbb{P}_{n}\left[w\left(h^{*} ; z\right)^{2}\right] \vee r_{1}^{*}\left(\delta_{1}\right) ; \delta_{1}\right) \\
\leq \mathbb{P}_{n}(h ; z)+\psi\left(16 \mathbb{P}_{n}\left[w(h ; z)^{2}\right] \delta_{1}\right)+\psi\left(16 \mathbb{P}_{n}\left[w\left(h^{*} ; z\right)^{2}\right] \vee r_{1}^{*}\left(\delta_{1}\right) ; \delta_{1}\right) . \tag{A.2.33}
\end{array}
$$

We specify the moment-penalized estimator to be

$$
\hat{h}_{\mathrm{MP}}=\arg \min _{\mathcal{H}}\left\{\mathbb{P}_{n} \ell(h ; z)+\psi\left(16 \mathbb{P}_{n}\left[\left(\ell(h ; z)-\widehat{\mathcal{L}_{0}^{*}}\right)^{2}\right] ; \delta_{1}\right)\right\} .
$$

Then we have

$$
\begin{equation*}
\mathbb{P}_{n} \ell\left(\hat{h}_{\mathrm{MP}} ; z\right)+\psi\left(16 \mathbb{P}_{n}\left[w\left(\hat{h}_{\mathrm{MP}} ; z\right)^{2}\right] ; \delta_{1}\right) \leq \mathbb{P}_{n} \ell\left(h^{*} ; z\right)+\psi\left(16 \mathbb{P}_{n}\left[w\left(h^{*} ; z\right)^{2}\right] ; \delta_{1}\right) \tag{A.2.34}
\end{equation*}
$$

Therefore, on the event $\mathcal{A}$,

$$
\begin{array}{r}
\mathcal{E}\left(\hat{h}_{\mathrm{MP}}\right) \leq \mathbb{P}_{n} \ell\left(\hat{h}_{\mathrm{MP}} ; z\right)+\psi\left(16 \mathbb{P}_{n}\left[w\left(\hat{h}_{\mathrm{MP}} ; z\right)^{2}\right] ; \delta_{1}\right)+\psi\left(16 \mathbb{P}_{n}\left[w\left(h^{*} ; z\right)^{2}\right] \vee r_{1}^{*}\left(\delta_{1}\right) ; \delta_{1}\right)-\mathbb{P}_{n} \ell\left(h^{*} ; z\right) \\
=\underset{\mathcal{H}}{\arg \min }\left\{\mathbb{P}_{n} \ell(h ; z)+\psi\left(16 \mathbb{P}_{n}[w(h ; z)] ; \delta_{1}\right)\right\}-\mathbb{P}_{n} \ell\left(h^{*} ; z\right)+\psi\left(16 \mathbb{P}_{n}\left[w\left(h^{*} ; z\right)^{2}\right] \vee r_{1}^{*}\left(\delta_{1}\right) ; \delta_{1}\right) \\
\leq \psi\left(16 \mathbb{P}_{n}\left[w\left(h^{*} ; z\right)^{2}\right] ; \delta_{1}\right)+\psi\left(16 \mathbb{P}_{n}\left[w\left(h^{*} ; z\right)^{2}\right] \vee r_{1}^{*}\left(\delta_{1}\right) ; \delta_{1}\right) \\
\leq 2 \psi\left(16 \mathbb{P}_{n}\left[w\left(h^{*} ; z\right)^{2}\right] \vee r_{1}^{*}\left(\delta_{1}\right) ; \delta_{1}\right), \tag{A.2.35}
\end{array}
$$

where the first inequality is due to (A.2.33) and the second inequality is due to (A.2.34).
From Bernstein's inequality at the single element $h^{*}$, for any fixed $\delta_{2} \in(0,1)$, with probability at least $1-\delta_{2}$,

$$
\begin{array}{r}
\mathbb{P}_{n}\left[w\left(h^{*} ; z\right)^{2}\right] \leq \mathbb{P}\left[w\left(h^{*} ; z\right)^{2}\right]+2 B \sqrt{\frac{2 \mathbb{P}\left[w\left(h^{*} ; z\right)^{2}\right] \log \frac{2}{\delta_{2}}}{n}}+\frac{4 B^{2} \log \frac{2}{\delta_{2}}}{n} \\
\leq 2 \mathbb{P}\left[w\left(h^{*} ; z\right)^{2}\right]+\frac{6 B^{2} \log \frac{2}{\delta_{2}}}{n} \tag{A.2.36}
\end{array}
$$

From (A.2.31) (A.2.35) (A.2.36), with probability at least

$$
\operatorname{Prob}(\mathcal{A})-\delta_{2} \geq 1-\left(\log _{2} \frac{8 B^{2} \vee 2 r_{1}^{*}\left(\delta_{1}\right)}{r_{1}^{*}\left(\delta_{1}\right)}+\frac{3}{2}\right) \delta_{1}-\delta_{2}
$$

we have

$$
\begin{align*}
\mathcal{E}\left(\hat{h}_{\mathrm{MP}}\right) & \leq 2 \psi\left(16 \mathbb{P}_{n}\left[w\left(h^{*} ; z\right)\right] \vee r_{1}^{*}\left(\delta_{1}\right) \vee \frac{B^{2}}{n} ; \delta_{1}\right) \\
& \leq 2 \psi\left(\left(32 \mathbb{P}\left[w\left(h^{*} ; z\right)^{2}\right]+\frac{96 B^{2} \log \frac{2}{\delta_{2}}}{n}\right) \vee r_{1}^{*}\left(\delta_{1}\right) \vee \frac{B^{2}}{n} ; \delta_{1}\right), \tag{A.2.37}
\end{align*}
$$

where the first inequality is due to (A.2.35) and the second inequality is due to (A.2.36).

## Part IV: final steps.

From the definition of $\psi$ and the fact that $r_{1}^{*}\left(\delta_{1}\right)$ is the fixed point of $16 B \psi\left(r ; \delta_{1}\right)$, we know that

$$
\begin{equation*}
r_{1}^{*}\left(\delta_{1}\right) \geq \frac{144 B^{2} \log \frac{8}{\delta_{1}}}{n} \tag{A.2.38}
\end{equation*}
$$

Denote $C_{n}:=2 \log _{2} n+5$ and take

$$
\delta_{1}=\frac{\delta}{C_{n}}
$$

then we have

$$
\begin{aligned}
2 \log _{2} \frac{8 B^{2} \vee 2 r_{1}^{*}\left(\delta_{1}\right)}{r_{1}^{*}\left(\delta_{1}\right)}+3 \leq \max & \left\{2 \log _{2} \frac{8 n}{144 \log 8}, 2+3\right\} \\
& \leq \max \left\{2 \log _{2} n, 5\right\} \leq C_{n}
\end{aligned}
$$

so

$$
\begin{equation*}
\left(\log _{2} \frac{8 B^{2} \vee 2 r_{1}^{*}\left(\delta_{1}\right)}{r_{1}^{*}\left(\delta_{1}\right)}+\frac{3}{2}\right) \delta_{1} \leq \frac{\delta}{2} \tag{A.2.39}
\end{equation*}
$$

Set $r^{*}=r_{1}^{*}\left(\delta_{1}\right)$ and take $\delta_{2}=\frac{\delta}{2}$. From (A.2.37), we obtain that with probability at least $1-\delta$,
the generalization error of $\hat{h}_{\mathrm{MP}}$ is upper bounded by

$$
\begin{equation*}
\mathcal{E}\left(\hat{h}_{\mathrm{MP}}\right) \leq 2 \psi\left(c\left[\mathbb{P}\left[w\left(h^{*} ; z\right)^{2}\right] \vee r^{*} \vee \frac{B^{2} \log \frac{4}{\delta}}{n}\right] ; \frac{\delta}{C_{n}}\right), \tag{A.2.40}
\end{equation*}
$$

where $c$ is an absolute constant. From (A.2.38) we have $r_{1}^{*}\left(\delta_{1}\right) \geq \frac{144 B^{2} \log \frac{8 C_{n}}{\delta}}{n} \geq \frac{B^{2} \log \frac{4}{\delta}}{n}$. Combine this fact with the inequality (A.2.40), we obtain that

$$
\begin{align*}
\mathcal{E}\left(\hat{h}_{\mathrm{MP}}\right) & \leq 2 \psi\left(c\left[\mathbb{P}\left[\left(\ell\left(h^{*} ; z\right)-\widehat{\mathcal{L}_{0}^{*}}\right)^{2}\right] \vee r^{*}\right] ; \frac{\delta}{C_{n}}\right) \\
& \leq 2 \psi\left(c_{0}\left[\mathcal{V}^{*} \vee r^{*} \vee\left(\widehat{\mathcal{L}_{0}^{*}}-\mathcal{L}_{0}^{*}\right)^{2}\right] ; \frac{\delta}{C_{n}}\right) . \tag{A.2.41}
\end{align*}
$$

where $c_{0}$ is an absolute constant. This completes the proof of Theorem 23.

## Analysis of the first-stage ERM estimator

After proving Theorem 23, the remaining part needed to prove Theorem 2 is to bound $\left(\widehat{\mathcal{L}_{0}^{*}}-\right.$ $\left.\mathcal{L}_{0}^{*}\right)^{2}$-the error of the first-stage ERM estimator.

The remaining steps in the proof of Theorem 2: We will give a guarantee on the first-stage ERM estimator, and combine this guarantee with Theorem 23 to prove Theorem 2. Recall that $\mathbb{P}_{S^{\prime}}$ is the empirical distribution of the "auxiliary" data set. Denote $\hat{h}_{\mathrm{ERM}} \in \arg \min _{\mathcal{H}} \mathbb{P}_{S^{\prime}} \ell(h ; z)$.

From Part I in the proof of Theorem $23, \forall \delta \in\left(0, \frac{1}{2}\right)$, with probability at least $1-\delta$,

$$
\sup _{\mathcal{F}}\left|\left(\mathbb{P}-\mathbb{P}_{n}\right) f\right| \leq \psi\left(4 B^{2} ; \delta\right) \leq \psi\left(4 B^{2} ; \frac{\delta}{C_{n}}\right) .
$$

Since $\psi$ is sub-root with respect to its first argument, we have

$$
\frac{\psi\left(4 B^{2} ; \frac{\delta}{C_{n}}\right)}{\sqrt{4 B^{2}}} \leq \frac{\psi\left(r^{*} ; \frac{\delta}{C_{n}}\right)}{\sqrt{r^{*}}}=\frac{\sqrt{r^{*}}}{16 B},
$$

where $r^{*}$ is the fixed point of $16 B \psi\left(r ; \frac{\delta}{C_{n}}\right)$. So we have proved that $\psi\left(4 B^{2} ; \frac{\delta}{C_{n}}\right) \leq \frac{\sqrt{r^{*}}}{8}$. Therefore,

$$
\sup _{\mathcal{F}}\left|\left(\mathbb{P}-\mathbb{P}_{n}\right) f\right| \leq \frac{\sqrt{r^{*}}}{8} .
$$

Because $\hat{h}_{\mathrm{ERM}} \in \arg \min _{\mathcal{H}} \mathbb{P}_{S^{\prime}} \ell(h ; z)$ and $\mathbb{P}_{S^{\prime}} \ell\left(\hat{h}_{\mathrm{ERM}} ; z\right)=\widehat{\mathcal{L}_{0}^{*}}$, we have

$$
\begin{array}{r}
\widehat{\mathcal{L}_{0}^{*}}-\mathcal{L}_{0}^{*}=\left(\mathbb{P}_{S^{\prime}} \ell\left(\hat{h}_{\mathrm{ERM}} ; z\right)-\mathbb{P}_{S^{\prime}} \ell\left(h^{*} ; z\right)\right)+\left(\mathbb{P}_{S^{\prime}} \ell\left(h^{*} ; z\right)-\mathbb{P} \ell\left(h^{*} ; z\right)\right) \\
\leq \mathbb{P}_{S^{\prime}} \ell\left(h^{*} ; z\right)-\mathbb{P} \ell\left(h^{*} ; z\right) \leq \sup _{\mathcal{F}}\left|\left(\mathbb{P}-\mathbb{P}_{n}\right) f\right|
\end{array}
$$

and

$$
\begin{array}{r}
\left.\widehat{\mathcal{L}_{0}^{*}}-\mathcal{L}_{0}^{*}=\left(\mathbb{P}_{S^{\prime}} \ell\left(\hat{h}_{\mathrm{ERM}} ; z\right)\right)-\mathbb{P} \ell\left(\hat{h}_{\mathrm{ERM}} ; z\right)\right)+\left(\mathbb{P} \ell\left(\hat{h}_{\mathrm{ERM}} ; z\right)-\mathbb{P} \ell\left(h^{*} ; z\right)\right) \\
\left.\geq \mathbb{P}_{S^{\prime}} \ell\left(\hat{h}_{\mathrm{ERM}} ; z\right)\right)-\mathbb{P} \ell\left(\hat{h}_{\mathrm{ERM}} ; z\right) \geq-\sup _{\mathcal{F}}\left|\left(\mathbb{P}-\mathbb{P}_{n}\right) f\right|
\end{array}
$$

Hence we have

$$
\left(\widehat{\mathcal{L}_{0}^{*}}-\mathcal{L}_{0}^{*}\right)^{2} \leq\left(\sup _{\mathcal{F}}\left|\left(\mathbb{P}-\mathbb{P}_{n}\right) f\right|\right)^{2} \leq \frac{r^{*}}{64}
$$

Combine this result with (A.2.41), we have with probability $1-2 \delta$,

$$
\begin{aligned}
\mathcal{E}\left(\hat{h}_{\mathrm{MP}}\right) & \leq 2 \psi\left(c_{1}\left(\mathcal{V}^{*} \vee r^{*}\right) ; \frac{\delta}{C_{n}}\right) \\
& \leq 2\left(\psi\left(c_{1} \mathcal{V}^{*} ; \frac{\delta}{C_{n}}\right) \vee \psi\left(c_{1} r^{*} ; \frac{\delta}{C_{n}}\right)\right) \\
& \leq 2 \psi\left(c_{1} \mathcal{V}^{*} ; \frac{\delta}{C_{n}}\right) \vee \frac{c_{1} r^{*}}{8 B}
\end{aligned}
$$

where $c_{1}=\max \left\{c_{0}, 16\right\}$ is an absolute constant, and the last inequality follows from the fact that $\frac{c_{1} r^{*}}{16}>r^{*}$ and the definition of fixed points. This completes the proof of Theorem 2.

## A.2.5 Estimating variance-dependent rates from data

In the remark following Theorem 2, we comment that fully data-dependent variance-dependent bounds can be derived by employing an empirical estimate to the unknown parameter $\mathcal{V}^{*}$. Here we present the full details and some discussion of this approach.

Theorem 24 (estimate of the variance-dependent rate from data) Consider the empirical centered second moment

$$
\left.\widehat{\mathcal{V}^{*}}:=\mathbb{P}_{n}\left[\ell\left(\hat{h}_{\mathrm{NMP}} ; z\right)-\widehat{\mathcal{L}_{0}^{*}}\right)^{2}\right],
$$

where $\widehat{\mathcal{L}_{0}^{*}} \in[-B, B]$ is the preliminary estimate of $\mathcal{L}^{*}$ obtained in the first-stage, $\psi$ is defined in Strategy 2, and

$$
\hat{h}_{\mathrm{NMP}} \in \underset{\mathcal{H}}{\arg \min } \mathbb{P}_{n} \ell(h ; z)-2 \psi\left(16 \mathbb{P}_{n}\left[\left(\ell(h ; z)-\widehat{\mathcal{L}}_{0}^{*}\right)^{2}\right]\right) .
$$

For any fixed $\delta \in(0,1)$, by performing the moment-penalized estimator in Strategy 2, with probability at least $1-\frac{\delta}{2}$,

$$
\begin{equation*}
\mathcal{E}\left(\hat{h}_{\mathrm{MP}}\right) \leq 4 \psi\left(16 \widehat{\mathcal{V}^{*}} ; \frac{\delta}{C_{n}}\right) \vee \frac{r^{*}}{8 B} \tag{A.2.42}
\end{equation*}
$$

where $r^{*}$ is the fixed point of $16 B \psi\left(r ; \frac{\delta}{C_{n}}\right)$.

Remarks. 1) The subscript "NMP" within $\hat{h}_{\mathrm{NMP}}$ means "negative moment penalization." Note that $\hat{h}_{\mathrm{NMP}}$ may not have good generalization performance, it is only used to compute $\widehat{\mathcal{V}^{*}}$ so that we can evaluate the estimator $\hat{h}_{\mathrm{MP}}$ proposed in Strategy 2.
2) While the fully data-dependent generalization error bound (A.2.42) provides a way to evaluate the moment-penalized estimator in Strategy 2 from training data, it seems that $\widehat{V^{*}}$ and $\mathcal{V}^{*}$ are not necessarily of the same order. Therefore, (A.2.42) may not be as tight as the original variancedependent rate in Theorem 2. One should view (A.2.42) as a relaxation of the original variance-
dependent rate in Theorem 2.
3) We also comment that the "sub-root" assumption in Theorem 2 is not needed here as we do not discuss the precision of $\widehat{\mathcal{L}_{0}^{*}}$. It is easy to combine Theorem 24 with the guarantee on $\widehat{\mathcal{L}_{0}^{*}}$ proved in Appendix A.2.4.

Proof of Theorem 24: define $\hat{f}_{\mathrm{NMP}}$ by $\hat{f}_{\mathrm{NMP}}(z)=\ell\left(\hat{h}_{\mathrm{NMP}} ; z\right)-\ell\left(h^{*} ; z\right), \forall z \in \mathcal{Z}$, and $w(h ; z)=$ $\ell(h ; z)-\widehat{\mathcal{L}_{0}^{*}}$. From the the results (A.2.31) (A.2.35) (A.2.39) in the proof of Theorem 23, we have with probability at least $1-\frac{\delta}{2}$,

$$
\begin{equation*}
\left(\mathbb{P}-\mathbb{P}_{n}\right) f \leq \psi\left(4 \mathbb{P}_{n}\left[f^{2}\right] \vee r^{*} ; \frac{\delta}{C_{n}}\right), \quad \forall f \in \mathcal{F} \tag{A.2.43}
\end{equation*}
$$

and

$$
\begin{equation*}
\mathcal{E}\left(\hat{h}_{\mathrm{MP}}\right) \leq 2 \psi\left(16 \mathbb{P}_{n}\left[w\left(h^{*} ; z\right)^{2}\right] \vee r^{*} ; \frac{\delta}{C_{n}}\right) . \tag{A.2.44}
\end{equation*}
$$

From the definition of $\hat{h}_{\mathrm{NMP}}$,

$$
\begin{equation*}
\mathbb{P}_{n} \ell\left(\hat{h}_{\mathrm{NMP}} ; z\right)-2 \psi\left(16 \mathbb{P}_{n}\left[w\left(\hat{h}_{\mathrm{NMP}} ; z\right)^{2}\right] ; \frac{\delta}{C_{n}}\right) \leq \mathbb{P}_{n} \ell\left(h^{*} ; z\right)-2 \psi\left(16 \mathbb{P}_{n}\left[w\left(h^{*} ; z\right)^{2}\right] ; \frac{\delta}{C_{n}}\right) \tag{A.2.45}
\end{equation*}
$$

Therefore, with probability at least $1-\frac{\delta}{2}$, we have

$$
\begin{align*}
& 2 \psi\left(16 \mathbb{P}_{n}\left[w\left(h^{*} ; z\right)^{2}\right] ; \frac{\delta}{C_{n}}\right) \\
& \leq 2 \psi\left(16 \mathbb{P}_{n}\left[w\left(\hat{h}_{\mathrm{NMP}} ; z\right)^{2}\right] ; \frac{\delta}{C_{n}}\right)+\mathbb{P}_{n} \ell\left(h^{*} ; z\right)-\mathbb{P}_{n} \ell\left(\hat{h}_{\mathrm{NMP}} ; z\right) \\
& =2 \psi\left(16 \mathbb{P}_{n}\left[w\left(\hat{h}_{\mathrm{NMP}} ; z\right)^{2}\right] ; \frac{\delta}{C_{n}}\right)+\mathbb{P}\left[\ell\left(h^{*} ; z\right)-\ell\left(\hat{h}_{\mathrm{NMP}} ; z\right)\right]+\left(\mathbb{P}_{n}-\mathbb{P}\right)\left[\ell\left(h^{*} ; z\right)-\ell\left(\hat{h}_{\mathrm{NMP}} ; z\right)\right] \\
& \leq 2 \psi\left(16 \mathbb{P}_{n}\left[w\left(\hat{h}_{\mathrm{NMP}} ; z\right)^{2}\right] ; \frac{\delta}{C_{n}}\right)+\left(\mathbb{P}-\mathbb{P}_{n}\right) \hat{\mathrm{NMP}}_{\mathrm{NMP}} \\
& \leq 2 \psi\left(16 \mathbb{P}_{n}\left[w\left(\hat{h}_{\mathrm{NMP}} ; z\right)^{2}\right] ; \frac{\delta}{C_{n}}\right)+\psi\left(4 \mathbb{P}_{n}\left[\hat{f}_{\mathrm{NMP}}^{2}\right] ; \frac{\delta}{C_{n}}\right) \tag{A.2.46}
\end{align*}
$$

where the first inequality is due to (A.2.45), the second inequality is due to the fact that $h^{*}$ minimizes the population risk; and the last inequality is due to (A.2.43).

Note that

$$
\begin{array}{r}
4 \mathbb{P}_{n}\left[\hat{f}_{\mathrm{NMP}}^{2}\right] \leq 8 \mathbb{P}_{n}\left[w\left(\hat{h}_{\mathrm{NMP}} ; z\right)^{2}\right]+8 \mathbb{P}_{n}\left[w\left(h^{*} ; z\right)^{2}\right] \\
\leq 16 \mathbb{P}_{n}\left[w\left(\hat{h}_{\mathrm{NMP}} ; z\right)^{2}\right] \vee 16 \mathbb{P}_{n}\left[w\left(h^{*} ; z\right)^{2}\right] .
\end{array}
$$

From the above inequality and (A.2.46), with probability at least $1-\frac{\delta}{2}$, we have

$$
\begin{align*}
& 2 \psi\left(16 \mathbb{P}_{n}\left[w\left(h^{*} ; z\right)^{2}\right] ; \frac{\delta}{C_{n}}\right) \\
& \leq 2 \psi\left(16 \mathbb{P}_{n}\left[w\left(\hat{h}_{\mathrm{NMP}} ; z\right)^{2}\right] ; \frac{\delta}{C_{n}}\right)+\psi\left(16 \mathbb{P}_{n}\left[w\left(\hat{h}_{\mathrm{NMP}} ; z\right)^{2}\right] ; \frac{\delta}{C_{n}}\right) \vee \psi\left(16 \mathbb{P}_{n}\left[w\left(h^{*} ; z\right)^{2}\right] ; \frac{\delta}{C_{n}}\right) . \tag{A.2.47}
\end{align*}
$$

Whether $\mathbb{P}_{n}\left[w\left(h^{*} ; z\right)^{2}\right] \leq 16 \mathbb{P}_{n}\left[w\left(\hat{h}_{\mathrm{NMP}} ; z\right)^{2}\right]$ or $\mathbb{P}_{n}\left[w\left(h^{*} ; z\right)^{2}\right]>16 \mathbb{P}_{n}\left[w\left(\hat{h}_{\mathrm{NMP}} ; z\right)^{2}\right]$, the inequality (A.2.47) always implies

$$
\begin{equation*}
\psi\left(16 \mathbb{P}_{n}\left[w\left(h^{*} ; z\right)^{2}\right] ; \frac{\delta}{C_{n}}\right) \leq 2 \psi\left(16 \mathbb{P}_{n}\left[w\left(\hat{h}_{\mathrm{NMP}} ; z\right)^{2}\right] ; \frac{\delta}{C_{n}}\right)=2 \psi\left(16 \widehat{\mathcal{V}^{*}} ; \frac{\delta}{C_{n}}\right) . \tag{A.2.48}
\end{equation*}
$$

(Note that $\widehat{\mathcal{V}^{*}}:=\mathbb{P}_{n}\left[w\left(\hat{h}_{\mathrm{NMP}} ; z\right)^{2}\right]$.) We conclude that with probability at least $1-\frac{\delta}{2}$,

$$
\begin{aligned}
\mathcal{E}\left(\hat{h}_{\mathrm{MP}}\right) & \leq 2 \psi\left(16 \mathbb{P}_{n}\left[w\left(h^{*} ; z\right)^{2}\right] \vee r^{*} ; \frac{\delta}{C_{n}}\right) \\
& =2 \psi\left(16 \mathbb{P}_{n}\left[w\left(h^{*} ; z\right)^{2}\right] ; \frac{\delta}{C_{n}}\right) \vee 2 \psi\left(r^{*} ; \frac{\delta}{C_{n}}\right) \\
& \leq 4 \psi\left(16 \widehat{\mathcal{V}^{*}} ; \frac{\delta}{C_{n}}\right) \vee \frac{r^{*}}{8 B},
\end{aligned}
$$

where the first inequality is due to (A.2.44) and the last inequality is due to (A.2.48). This completes the proof.

## A.2.6 Auxiliary lemmata

Lemma 16 (Talagrand's concentration inequality for empirical processes, [1]) Let $\mathcal{F}$ be a class of functions that map $\mathcal{Z}$ into $\left[B_{1}, B_{2}\right]$. Assume that there is some $r>0$ such that for every $f \in \mathcal{F}$, $\operatorname{Var}\left[f\left(z_{i}\right)\right] \leq r$. Then, for every $\delta \in(0,1)$, with probability at least $1-\delta$,

$$
\sup _{f \in \mathcal{F}}\left(\mathbb{P}-\mathbb{P}_{n}\right) f \leq 3 \Re \mathcal{F}+\sqrt{\frac{2 r \log \frac{1}{\delta}}{n}}+\left(B_{2}-B_{1}\right) \frac{\log \frac{1}{\delta}}{n},
$$

and with probability at least $1-\delta$,

$$
\sup _{f \in \mathcal{F}}\left(\mathbb{P}-\mathbb{P}_{n}\right) f \leq 4 \Re_{n} \mathcal{F}+\sqrt{\frac{2 r \log \frac{2}{\delta}}{n}}+\frac{9}{2}\left(B_{2}-B_{1}\right) \frac{\log \frac{2}{\delta}}{n} .
$$

Moreover, the same results hold for the quantity sup $f_{f \in \mathcal{F}}\left(\mathbb{P}_{n}-\mathbb{P}\right) f$.
Lemma 17 (Bernstein's inequality, [139]) Let $X_{1}, \cdots, X_{n}$ be real-valued, independent, meanzero random variables and suppose that for some constants $\sigma, B>0$,

$$
\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left|X_{i}\right|^{k} \leq \frac{k!}{2} \sigma^{2} B^{k-2}, \quad k=2,3, \cdots
$$

Then, $\forall \delta \in(0,1)$, with probability at least $1-\delta$

$$
\begin{equation*}
\left|\frac{1}{n} \sum_{i=1}^{n} X_{i}\right| \leq \sqrt{\frac{2 \sigma^{2} \log \frac{2}{\delta}}{n}}+\frac{B \log \frac{2}{\delta}}{n} . \tag{A.2.49}
\end{equation*}
$$

## A.3 Proofs for Section 1.5, Section 1.6 and Section 1.7

## A.3.1 Proof of Lemma 2

Fix $u \in \mathcal{B}^{d}(0,1)$ and $\theta_{1}, \theta_{2} \in \Theta$, then we have

$$
u^{T}\left(\nabla \ell\left(\theta_{1} ; z\right)-\nabla \ell\left(\theta_{2} ; z\right)\right)^{T}=\int_{0}^{1} u^{T}\left[\nabla^{2} \ell\left(\theta_{2}+v\left(\theta_{1}-\theta_{2}\right) ; z\right)\right]\left(\theta_{1}-\theta_{2}\right) d v
$$

By Jensen's inequality,

$$
\begin{aligned}
\exp \left(\frac{u^{T}\left(\nabla \ell\left(\theta_{1} ; z\right)-\nabla \ell\left(\theta_{2} ; z\right)\right)}{\beta\left\|\theta_{1}-\theta_{2}\right\|}\right) & =\exp \left(\int_{0}^{1} u^{T}\left[\nabla^{2} \ell\left(\theta_{2}+v\left(\theta_{1}-\theta_{2}\right) ; z\right)\right] \frac{\left(\theta_{1}-\theta_{2}\right)}{\left\|\theta_{1}-\theta_{2}\right\|} d v\right) \\
& \leq \int_{0}^{1} \exp \left(u^{T}\left[\nabla^{2} \ell\left(\theta_{2}+v\left(\theta_{1}-\theta_{2}\right) ; z\right)\right] \frac{\left(\theta_{1}-\theta_{2}\right)}{\left\|\theta_{1}-\theta_{2}\right\|}\right) d v
\end{aligned}
$$

It is then straightforward to prove the lemma by taking expectation with respect to $z$ in the above inequality and using the condition (1.5.2).

## A.3.2 Proof of Proposition 2

Take $V=\left\{v \in \mathbb{R}^{d}:\|v\| \leq \max \left\{\Delta_{M}, \frac{1}{n}\right\}\right\}$. We will first prove a "uniform localized convergence" argument over all $\theta \in \Theta$ and $v \in V$.

## Proposition 7 (directional "uniform localized convergence" of gradient) Under Assumption 1,

 $\forall \delta \in(0,1)$, with probability at least $1-\delta$, for all $\theta \in \Theta, v \in V$, either $\left\|\theta-\theta^{*}\right\|^{2}+\|v\|^{2} \leq \frac{2}{n^{2}}$, or$$
\begin{aligned}
& \left(\mathbb{P}-\mathbb{P}_{n}\right)\left[\left(\nabla \ell(\theta ; z)-\nabla \ell\left(\theta^{*} ; z\right)\right)^{T} v\right] \\
& \leq c_{1} \beta \max \left\{\left\|\theta-\theta^{*}\right\|^{2}+\|v\|^{2}, \frac{2}{n^{2}}\right\}\left(\sqrt{\frac{d+\log \frac{2 \log _{2}\left(2 n^{2} \Delta_{M}^{2}+2\right)}{\delta}}{n}}+\frac{d+\log \frac{\left.2 \log _{2}\left(2 n^{2} \Delta_{M}^{2}+2\right)\right\}}{\delta}}{n}\right),
\end{aligned}
$$

where $c_{1}$ is an absolute constant.

Proof of Proposition 7: for $(\theta, v) \in \Theta \times V$, let $g_{(\theta, v)}=\left(\nabla \ell(\theta ; z)-\nabla \ell\left(\theta^{*} ; z\right)\right)^{T} v$. For $\left(\theta_{1}, v_{1}\right)$ and $\left(\theta_{2}, v_{2}\right) \in \Theta \times v$, define the norm on the product space $\Theta \times V$ as

$$
\begin{equation*}
\left\|\left(\theta_{1}, v_{1}\right)-\left(\theta_{2}, v_{2}\right)\right\|_{\mathrm{pro}}=\sqrt{\left\|\theta_{1}-\theta_{2}\right\|^{2}+\left\|v_{1}-v_{2}\right\|^{2}} \tag{A.3.1}
\end{equation*}
$$

Denote $\mathcal{B}(\sqrt{r}):=\left\{(\theta, v) \in \Theta \times V:\left\|\theta-\theta^{*}\right\|^{2}+\|v\|^{2} \leq r\right\}$. Given $\left(\theta_{1}, v_{1}\right),\left(\theta_{2}, v_{2}\right) \in \mathcal{B}(\sqrt{r})$, we perform the following re-arrangement and decomposition steps:

$$
\begin{align*}
& g_{\left(\theta_{1}, v_{1}\right)}(z)-g_{\left(\theta_{2}, v_{2}\right)}(z) \\
& =\left(\nabla \ell\left(\theta_{1} ; z\right)-\nabla \ell\left(\theta^{*} ; z\right)\right)^{T} v_{1}-\left(\nabla \ell\left(\theta_{2} ; z\right)-\nabla \ell\left(\theta^{*} ; z\right)\right)^{T} v_{2} \\
& =\left(\nabla \ell\left(\theta_{1} ; z\right)-\nabla \ell\left(\theta^{*} ; z\right)\right)^{T}\left(v_{1}-v_{2}\right)+\left(\nabla \ell\left(\theta_{1} ; z\right)-\nabla \ell\left(\theta^{*} ; z\right)\right)^{T} v_{2}+\left(\nabla \ell\left(\theta^{*} ; z\right)-\nabla \ell\left(\theta_{2} ; z\right)\right)^{T} v_{2} \\
& =\left(\nabla \ell\left(\theta_{1} ; z\right)-\nabla \ell\left(\theta^{*} ; z\right)\right)^{T}\left(v_{1}-v_{2}\right)+\left(\nabla \ell\left(\theta_{1} ; z\right)-\nabla \ell\left(\theta_{2} ; z\right)\right)^{T} v_{2} \tag{A.3.2}
\end{align*}
$$

When $\left(\theta_{1}, v_{1}\right),\left(\theta_{2}, v_{2}\right) \in \mathcal{B}(\sqrt{r})$, we have

$$
\left\|\theta_{1}-\theta^{*}\right\|\left\|v_{1}-v_{2}\right\| \leq \sqrt{r}\left\|v_{1}-v_{2}\right\| \leq \sqrt{r}\left\|\left(\theta_{1}, v_{1}\right)-\left(\theta_{2}, v_{2}\right)\right\|_{\mathrm{pro}},
$$

so from Assumption 1 , $\left(\nabla \ell\left(\theta_{1} ; z\right)-\nabla \ell\left(\theta^{*} ; z\right)\right)^{T}\left(v_{1}-v_{2}\right)$ is $\beta \sqrt{r}\left\|\left(\theta_{1}, v_{1}\right)-\left(\theta_{2}, v_{2}\right)\right\|_{\text {pro }}-$ subexponential. Similarly, we can prove $\left(\nabla \ell\left(\theta_{1} ; z\right)-\nabla \ell\left(\theta_{2} ; z\right)\right)^{T} v_{2}$ to be $\beta \sqrt{r}\left\|\left(\theta_{1}, v_{1}\right)-\left(\theta_{2}, v_{2}\right)\right\|_{\text {pro }}-$ subexponential. From the decomposition (A.3.2) and Jensen's inequality, for all $\left(\theta_{1}, v_{1}\right),\left(\theta_{2}, v_{2}\right) \in$ $\mathcal{B}(\sqrt{r})$, we have

$$
\begin{aligned}
& \exp \left(\frac{g_{\left(\theta_{1}, v_{1}\right)}(z)-g_{\left(\theta_{2}, v_{2}\right)}(z)}{2 \beta \sqrt{r}\left\|\left(\theta_{1}, v_{1}\right)-\left(\theta_{2}, v_{2}\right)\right\|_{\mathrm{pro}}}\right) \\
& \leq \frac{1}{2} \exp \left(\frac{\left(\nabla \ell\left(\theta_{1} ; z\right)-\nabla \ell\left(\theta^{*} ; z\right)\right)^{T}\left(v_{1}-v_{2}\right)}{\beta \sqrt{r}\left\|\left(\theta_{1}, v_{1}\right)-\left(\theta_{2}, v_{2}\right)\right\|_{\mathrm{pro}}}\right)+\frac{1}{2} \exp \left(\frac{\left(\nabla \ell\left(\theta_{1} ; z\right)-\nabla \ell\left(\theta_{2} ; z\right)\right)^{T} v_{2}}{\beta \sqrt{r}\left\|\left(\theta_{1}, v_{1}\right)-\left(\theta_{2}, v_{2}\right)\right\|_{\mathrm{pro}}}\right) .
\end{aligned}
$$

By taking expectation with respect to $z$ in the above inequality, we prove that $g_{\left(\theta_{1}, v_{1}\right)}(z)-g_{\left(\theta_{2}, v_{2}\right)}(z)$ is a $2 \beta \sqrt{r}\left\|\left(\theta_{1}, v_{1}\right)-\left(\theta_{2}, v_{2}\right)\right\|_{\text {pro }}-$ sub-exponential random variable, i.e.,

$$
\left\|g_{\left(\theta_{1}, v_{1}\right)}(z)-g_{\left(\theta_{2}, v_{2}\right)}(z)\right\|_{\text {Orlicz }_{1}} \leq 2 \beta \sqrt{r}\left\|\left(\theta_{1}, v_{1}\right)-\left(\theta_{2}, v_{2}\right)\right\|_{\text {pro }} .
$$

From Bernstein inequality for sub-exponential variables (Lemma 21), for any fixed $u \geq 0$ and

$$
\begin{aligned}
& \left(\theta_{1}, v_{1}\right),\left(\theta_{2}, v_{2}\right) \in \Theta \times V \\
& \operatorname{Prob}\left\{\left\lvert\,\left(\mathbb{P}-\mathbb{P}_{n}\right)\left[g_{\left(\theta_{1}, v_{1}\right)}(z)-g_{\left(\theta_{2}, v_{2}\right)}(z)\right] \geq 2 \beta \sqrt{r}\left\|\left(\theta_{1}, v_{1}\right)-\left(\theta_{2}, v_{2}\right)\right\|_{\text {pro }} \sqrt{\frac{2}{n}} \sqrt{u}\right.\right. \\
& \\
& \left.+\frac{2 \beta \sqrt{r}\left\|\left(\theta_{1}, v_{1}\right)-\left(\theta_{2}, v_{2}\right)\right\|_{\text {pro }}}{n} u\right\} \leq 2 \exp (-u) .
\end{aligned}
$$

The above inequality implies that the empirical process $\left(\mathbb{P}-\mathbb{P}_{n}\right) g_{(\theta, v)}$ has a mixed sub-Gaussian-sub-exponential increments with respect to the metrics $\left(\frac{2 \beta \sqrt{r}}{n}\|\cdot\|_{\text {pro }}, \frac{2 \sqrt{2} \beta \sqrt{r}}{\sqrt{n}}\|\cdot\|_{\text {pro }}\right.$ ) (see Definition 15).

From Lemma 24, there exists an absolute constants $C$ such that $\forall \delta \in(0,1)$, with probability at least $1-\delta$,

$$
\begin{array}{r}
\sup _{\left\|\theta-\theta^{*}\right\|^{2}+\|v\|^{2} \leq r}\left(\mathbb{P}-\mathbb{P}_{n}\right) g_{(\theta, v)} \leq C\left(\gamma_{2}\left(\mathcal{B}(\sqrt{r}), \frac{2 \sqrt{2} \beta \sqrt{r}}{\sqrt{n}}\|\cdot\|_{\text {pro }}\right)+\right. \\
\gamma_{1}\left(\mathcal{B}(\sqrt{r}), \frac{2 \beta \sqrt{r}}{n}\|\cdot\|_{\text {pro }}\right) \\
\left.+\beta r \sqrt{\frac{\log \frac{1}{\delta}}{n}}+\beta r \frac{\log \frac{1}{\delta}}{n}\right) .
\end{array}
$$

Using Dudley's integral (Lemma 23) to bound the $\gamma_{1}$ functional and the $\gamma_{2}$ functional, we obtain that there exist absolute constant $c_{1}$ such that $\forall \delta \in(0,1)$, with probability at least $1-\delta$,

$$
\begin{equation*}
\sup _{\left\|\left(\theta, \theta^{\prime}\right)-\left(\theta^{*}, \theta^{*}\right)\right\|^{2} \leq r}\left|\left(\mathbb{P}-\mathbb{P}_{n}\right) g_{(\theta, v)}\right| \leq c_{1} \beta r\left(\sqrt{\frac{d+\log \frac{1}{\delta}}{n}}+\frac{d+\log \frac{1}{\delta}}{n}\right) . \tag{A.3.3}
\end{equation*}
$$

We set

$$
\psi(r ; \delta)=c_{1} \beta r\left(\sqrt{\frac{d+\log \frac{1}{\delta}}{n}}+\frac{d+\log \frac{1}{\delta}}{n}\right)
$$

Denote $R=2\left(\Delta_{M}^{2}+\frac{1}{n^{2}}\right)$ and $r_{0}=\frac{2}{n^{2}}$. Since $V$ is a $d$-dimensional ball centered at the origin with radius $\max \left\{\Delta_{M}, \frac{1}{n}\right\}$, we know that $\left\|\theta-\theta^{*}\right\|^{2}+\|v\|^{2} \leq 2 \Delta_{M}^{2}+\frac{1}{n^{2}} \leq R$. We apply Proposition 6 and
obtain: for any fixed $\delta \in(0,1)$, with probability at least $1-\delta$, for all $\theta \in \Theta$ and $v \in V$,

$$
\begin{aligned}
& \left(\mathbb{P}-\mathbb{P}_{n}\right)\left[\left(\nabla \ell(\theta ; z)-\nabla \ell\left(\theta^{*} ; z\right)\right)^{T} v\right]=\left(\mathbb{P}-\mathbb{P}_{n}\right) g_{(\theta, v)} \\
& \leq \psi\left(\max \left\{\left\|\theta-\theta^{*}\right\|^{2}+\|v\|^{2}, \frac{2}{n^{2}}\right\} ; \frac{\delta}{2 \log _{2}\left(2 R / \frac{2}{n^{2}}\right)}\right) \\
& =c_{1} \beta \max \left\{\left\|\theta-\theta^{*}\right\|^{2}+\|v\|^{2}, \frac{2}{n^{2}}\right\}\left(\sqrt{\frac{d+\log \frac{2 \log _{2}\left(n^{2} R\right)}{\delta}}{n}}+\frac{d+\log \frac{2 \log _{2}\left(n^{2} R\right)}{\delta}}{n}\right) .
\end{aligned}
$$

This completes the proof of Proposition 7.

Proof of Proposition 2: in order to uniformly bound $\left\|\left(\mathbb{P}-\mathbb{P}_{n}\right)\left(\nabla \ell(\theta ; z)-\nabla \ell\left(\theta^{*} ; z\right)\right)\right\|$ for all $\theta \in \Theta$, we take

$$
v=\max \left\{\left\|\theta-\theta^{*}\right\|, \frac{1}{n}\right\} \cdot \frac{\left(\mathbb{P}-\mathbb{P}_{n}\right)\left(\nabla \ell(\theta ; z)-\nabla \ell\left(\theta^{*} ; z\right)\right)}{\left\|\left(\mathbb{P}-\mathbb{P}_{n}\right)\left(\nabla \ell(\theta ; z)-\nabla \ell\left(\theta^{*} ; z\right)\right)\right\|}
$$

in Proposition 7. Clearly $\|v\|=\max \left\{\left\|\theta-\theta^{*}\right\|, \frac{1}{n}\right\}$. From Proposition 2, we can prove that there exists an absolute constant $c$ such that $\forall \delta \in(0,1)$, with probability at least $1-\delta$, for all $\theta \in \Theta$,

$$
\begin{aligned}
& \left\|\left(\mathbb{P}-\mathbb{P}_{n}\right)\left(\nabla \ell(\theta ; z)-\nabla \ell\left(\theta^{*} ; z\right)\right)\right\| \\
& \leq c \beta \max \left\{\left\|\theta-\theta^{*}\right\|, \frac{1}{n}\right\}\left(\sqrt{\frac{d+\log \frac{2 \log _{2}\left(2 n^{2} \Delta_{M}^{2}+2\right)}{\delta}}{n}}+\frac{d+\log \frac{2 \log _{2}\left(2 n^{2} \Delta_{M}^{2}+2\right)}{n}}{n}\right) \\
& \leq c \beta \max \left\{\left\|\theta-\theta^{*}\right\|, \frac{1}{n}\right\}\left(\sqrt{\frac{d+\log \frac{4 \log _{2}\left(2 n \Delta_{M}+2\right)}{\delta}}{n}}+\frac{d+\log \frac{4 \log _{2}\left(2 n \Delta_{M}+2\right)}{\delta}}{n}\right) .
\end{aligned}
$$

This completes the proof of Proposition 2.

## A.3.3 Proof of Theorem 3

We first prove a proposition on the uniform localized convergence of gradients under Assumption 1 and Assumption 2.

Proposition 8 (uniform localized convergence of gradients) Let Assumption 1, Assumption 2 hold along with the optimality condition $\mathbb{P} \nabla \ell\left(\theta^{*} ; z\right)=0$. Given $\delta \in(0,1)$, denote

$$
\begin{aligned}
& \text { term I }:=\sqrt{\frac{2 \mathbb{P}\left[\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{2}\right] \log \frac{4}{\delta}}{n}}+\frac{G_{*} \log \frac{4}{\delta}}{n}, \\
& \text { term II }:=\sqrt{\frac{d+\log \frac{8 \log _{2}\left(2 n \Delta_{M}+2\right)}{\delta}}{n}}+\frac{d+\log \frac{\log _{2}\left(2 n \Delta_{M}+2\right)}{\delta}}{n} .
\end{aligned}
$$

Then with probability at least $1-\delta$, we have the following:

$$
\begin{equation*}
\left\|\left(\mathbb{P}_{n}-\mathbb{P}\right) \nabla \ell\left(\theta^{*} ; z\right)\right\| \leq \text { term } \mathrm{I} \tag{A.3.4}
\end{equation*}
$$

and

$$
\begin{equation*}
\left\|\left(\mathbb{P}_{n}-\mathbb{P}\right) \nabla \ell(\theta ; z)\right\| \leq \operatorname{term} \mathrm{I}+c_{0} \beta \max \left\{\frac{1}{n},\left\|\theta-\theta^{*}\right\|\right\} \cdot \operatorname{term} \mathrm{II}, \quad \forall \theta \in \Theta \tag{A.3.5}
\end{equation*}
$$

where $c_{0}$ is an absolute constant.

Proof of Proposition 8: from Proposition 2, there exists an absolute constant $c_{0}$ such that $\forall \delta_{1}>$ 0 , with probability at least $1-\frac{\delta}{2}$, for all $\theta \in \Theta$,

$$
\begin{align*}
& \left\|\left(\mathbb{P}_{n}-\mathbb{P}\right) \nabla \ell(\theta ; z)\right\| \\
& \leq\left\|\left(\mathbb{P}_{n}-\mathbb{P}\right) \nabla \ell\left(\theta^{*} ; z\right)\right\|+c_{0} \beta \max \left\{\left\|\theta-\theta^{*}\right\|, \frac{1}{n}\right\} \cdot \text { term II. } \tag{A.3.6}
\end{align*}
$$

From Bernstein's inequality for vectors (Lemma 22), we have with probability at least $1-\frac{\delta}{2}$,

$$
\begin{equation*}
\left\|\mathbb{P} \nabla \ell\left(\theta^{*} ; z\right)-\mathbb{P}_{n} \nabla \ell\left(\theta^{*} ; z\right)\right\| \leq \sqrt{\frac{2 \mathbb{P}\left[\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{2}\right] \log \frac{4}{\delta}}{n}}+\frac{G_{*} \log \frac{4}{\delta}}{n}=\text { term I } \tag{A.3.7}
\end{equation*}
$$

Combining (A.3.6) and (A.3.7) by a union bound, we complete the proof of Proposition 8.

We first present the following lemma.

Lemma 18 (relationship between curvature conditions) For a function $F$, consider the following conditions:

1. Strong convexity (SC): for all $\theta_{1}, \theta_{2} \in \Theta$ we have

$$
F\left(\theta_{1}\right) \geq F\left(\theta_{2}\right)+\nabla F\left(\theta_{2}\right)^{T}\left(\theta_{1}-\theta_{2}\right)+\frac{\mu}{2}\left\|\theta_{1}-\theta_{2}\right\|^{2}
$$

2. Polyak-Lojasiewisz (PL): for all $\theta \in \Theta$ we have

$$
F(\theta)-F\left(\theta^{*}\right) \leq \frac{1}{2 \mu}\|\nabla F(\theta)\|^{2}
$$

3. Error Bound (EB): for all $\theta \in \Theta$ we have

$$
\|\nabla F(\theta)\| \geq \mu\left\|\theta-\theta^{*}\right\| .
$$

4. Quadratic Growth ( $Q G$ ): for all $\theta \in \Theta$ we have

$$
F(\theta)-F\left(\theta^{*}\right) \geq \frac{\mu}{2}\left\|\theta-\theta^{*}\right\|^{2}
$$

Then, the following hold:

$$
(S C) \Longrightarrow(P L) \Longrightarrow(E B) \Longrightarrow(Q G)
$$

Proof of Lemma 18 can be adapted from [[32], Appendix A]. Note that some parameters in the
original statements in [32] have typos though the proof ideas are correct. In Lemma 18 we fix those typos on the parameters. As argued in [32], (PL) and the equivalent (QG) (under the smoothness condition and change of parameters) are the most general conditions that allow linear convergence to a global minimizer.

We now prove Theorem 3.

Proof of Theorem 3: we prove the results on the event

$$
\mathcal{A}:=\{\text { the results (A.3.4) (A.3.5) in Proposition } 8 \text { hold true }\}
$$

whose measure is at least $1-\delta$. We keep the notations "term I" and "term II" used in Proposition 8 , which are defined by

$$
\begin{array}{r}
\text { term I }:=\sqrt{\frac{2 \mathbb{P}\left[\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{2}\right] \log \frac{4}{\delta}}{n}}+\frac{G_{*} \log \frac{4}{\delta}}{n}, \\
\text { term II }:=\sqrt{\frac{d+\log \frac{8 \log _{2}\left(2 n \Delta_{M}+2\right)}{\delta}}{n}+\frac{d+\log \frac{\log _{2}\left(2 n \Delta_{M}+2\right)}{\delta}}{n}} .
\end{array}
$$

The PL condition (Assumption 3) implies that $\mathbb{P} \nabla \ell\left(\theta^{*} ; z\right)=0$. From the result (A.3.4) in Proposition 8,

$$
\left\|\mathbb{P}_{n} \nabla \ell\left(\theta^{*} ; z\right)\right\|=\left\|\left(\mathbb{P}_{n}-\mathbb{P}\right) \nabla \ell\left(\theta^{*} ; z\right)\right\| \leq \text { term } \mathrm{I} .
$$

So we know that the equation

$$
\begin{equation*}
\left\|\mathbb{P}_{n} \nabla \ell(\theta ; z)\right\| \leq \text { term I. } \tag{A.3.8}
\end{equation*}
$$

must have a solution within $\Theta$.

The result (A.3.5) implies that for all $\theta \in \Theta$ such that $\left\|\theta-\theta^{*}\right\| \leq \frac{1}{n}$,

$$
\|\mathbb{P} \nabla \ell(\theta ; z)\| \leq\left\|\mathbb{P}_{n} \nabla \ell(\theta ; z)\right\|+\text { term I }+c_{0} \beta\left\|\theta-\theta^{*}\right\| \cdot \text { term II. }
$$

Since the PL condition implies (see Lemma 18)

$$
\|\mathbb{P} \nabla \ell(\theta ; z)\| \geq \mu\left\|\theta-\theta^{*}\right\|,
$$

for all $\theta \in \Theta$ such that $\left\|\theta-\theta^{*}\right\| \leq \frac{1}{n}$, we have

$$
\mu\left\|\theta-\theta^{*}\right\| \leq\|\mathbb{P} \nabla \ell(\theta ; z)\| \leq\left\|\mathbb{P}_{n} \nabla \ell(\theta ; z)\right\|+\text { term I }+c_{0} \beta\left\|\theta-\theta^{*}\right\| \cdot \text { term II, }
$$

where $c$ is an absolute constant. Therefore, for all $\theta \in \Theta$, we must have

$$
\begin{equation*}
\mu\left\|\theta-\theta^{*}\right\| \leq\|\mathbb{P} \nabla \ell(\theta ; z)\| \leq\left\|\mathbb{P}_{n} \nabla \ell(\theta ; z)\right\|+\text { term } \mathrm{I}+c_{0} \beta\left\|\theta-\theta^{*}\right\| \cdot \text { term II }+\frac{\mu}{n} . \tag{A.3.9}
\end{equation*}
$$

Let $\hat{\theta} \in \Theta$ be an arbitrary solution that satisfies (A.3.8). From (A.3.9), we obtain the inequalities for $\left\|\hat{\theta}-\theta^{*}\right\|$ :

$$
\begin{equation*}
\mu\left\|\hat{\theta}-\theta^{*}\right\| \leq\|\mathbb{P} \nabla \ell(\hat{\theta} ; z)\| \leq 2 \cdot \operatorname{term} \mathrm{I}+c_{0} \beta \cdot \operatorname{term} \mathrm{II} \cdot\left\|\theta-\theta^{*}\right\|+\frac{\mu}{n} . \tag{A.3.10}
\end{equation*}
$$

Let $c=\max \left\{4 c_{0}^{2}, 1\right\}$. When

$$
n \geq \frac{c \beta^{2}\left(d+\log \frac{4 \log \left(2 n \Delta_{M}+1\right)}{\delta}\right)}{\mu^{2}}
$$

we have $c_{0} \beta \cdot$ term II $\leq \frac{\mu}{2}$ so that from (A.3.10),

$$
\left\|\hat{\theta}-\theta^{*}\right\| \leq \frac{2}{\mu}\left(2 \cdot \operatorname{term} \mathrm{I}+\frac{\mu}{n}\right)
$$

and the event $\mathcal{A}$. Plugging in " $c_{0} \beta \cdot$ term II $\leq \frac{\mu}{2}$ " and " $\left\|\hat{\theta}-\theta^{*}\right\| \leq \frac{2}{\mu}\left(2 \cdot\right.$ term $\left.\mathrm{I}+\frac{\mu}{n}\right)$ " into the second inequality within (A.3.10), we further have

$$
\begin{aligned}
\|\mathbb{P} \nabla \ell(\hat{\theta} ; z)\| \leq 2 \cdot \operatorname{term} \mathrm{I} & +\frac{\mu}{n}+\frac{\mu}{2}\left\|\hat{\theta}-\theta^{*}\right\| \\
& \leq 4 \cdot \operatorname{term} \mathrm{I}+\frac{2 \mu}{n} .
\end{aligned}
$$

Lastly, since the PL condition implies (see Lemma 18)

$$
\mathbb{P} \ell(\hat{\theta} ; z)-\mathbb{P} \ell\left(h^{*} ; z\right) \leq \frac{\|\mathbb{P} \ell(\hat{\theta} ; z)\|^{2}}{2 \mu}
$$

by plugging in " $\|\mathbb{P} \nabla \ell(\hat{\theta} ; z)\| \leq 4 \cdot$ term $\mathrm{I}+\frac{2 \mu}{n}$ " we have

$$
\begin{aligned}
\mathbb{P} \ell(\hat{\theta} ; z)-\mathbb{P} \ell\left(h^{*} ; z\right) & \leq \frac{\|\mathbb{P} \ell(\hat{\theta} ; z)\|^{2}}{2 \mu} \\
& \leq \frac{16}{\mu}(\operatorname{term~I})^{2}+\frac{4 \mu}{n^{2}} \\
& \leq \frac{64 \mathbb{P}\left[\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{2}\right] \log \frac{4}{\delta}}{\mu n}+\frac{32 G_{*}^{2} \log ^{2} \frac{4}{\delta}+4 \mu^{2}}{\mu n^{2}} .
\end{aligned}
$$

This completes the proof of Theorem 3.

## A.3.4 Proof of Theorem 4

We first prove a simple proposition, which studies how the accumulation of sample approximation errors at every step influences the convergence of the algorithm.

Proposition 9 (localized statistical error of a linearly convergent iterative algorithm) Consider a function $F$ (for which we call the "Lyapunov function") and a parameter $\gamma \in(0,1)$. Assume an
algorithm satisfies for all $t=0,1, \ldots$

$$
\begin{array}{r}
F\left(\theta^{t+1}\right) \leq(1-\gamma) F\left(\theta^{t}\right)+\varepsilon^{t}(n) \\
\varepsilon^{t}(n) \leq \alpha(n) F\left(\theta^{t}\right)+\varepsilon^{*}(n) \\
\text { and } \quad \theta^{t} \in \Theta
\end{array}
$$

When the sample size $n$ is large enough such that $\alpha(n) \leq \frac{\gamma}{2}$, we have

$$
F\left(\theta^{t}\right) \leq\left(1-\frac{\gamma}{2}\right)^{t} F\left(\theta^{0}\right)+\frac{2}{\gamma} \varepsilon^{*}(n), \quad t=0,1, \cdots
$$

Proof of Proposition 9: we have

$$
\begin{aligned}
F\left(\theta^{t+1}\right) \leq(1 & -\gamma+\alpha(n)) F\left(\theta^{t}\right)+\varepsilon^{*}(n) \\
& \leq\left(1-\frac{\gamma}{2}\right) F\left(\theta^{*}\right)+\varepsilon^{*}(n)
\end{aligned}
$$

Then by induction we have

$$
F\left(\theta^{t}\right) \leq\left(1-\frac{\gamma}{2}\right)^{t} F\left(\theta^{0}\right)+\frac{2}{\gamma} \varepsilon^{*}(n), \quad t=0,1, \cdots
$$

This completes the proof of Proposition 9.

We now prove Theorem 4.

Proof of Theorem 4: Assumption 1 implies that the population risk is $\beta$-smooth. Consider the gradient descent algorithm (1.5.8) with fixed step size $\frac{1}{\beta}$. We have for all $t=0,1, \cdots$,

$$
\theta^{t+1}=\theta^{t}-\frac{1}{\beta} \mathbb{P}_{n} \nabla \ell\left(\theta^{t} ; z\right) .
$$

So we have

$$
\begin{array}{r}
\mathbb{P} \ell\left(\theta^{t+1} ; z\right)-\mathbb{P} \ell\left(\theta^{t} ; z\right) \leq\left(\mathbb{P} \nabla \ell\left(\theta^{t} ; z\right)\right)^{T}\left(\theta^{t+1}-\theta^{t}\right)+\frac{\beta}{2}\left\|\theta^{t+1}-\theta^{t}\right\|^{2} \\
=-\frac{1}{\beta}\left(\mathbb{P} \nabla \ell\left(\theta^{t} ; z\right)\right)^{T}\left(\mathbb{P}_{n} \nabla \ell\left(\theta^{t} ; z\right)\right)+\frac{1}{2 \beta}\left\|\mathbb{P}_{n} \nabla \ell\left(\theta^{t} ; z\right)\right\|^{2} \\
=-\frac{1}{\beta}\left\|\mathbb{P} \nabla \ell\left(\theta^{t} ; z\right)\right\|^{2}-\frac{1}{\beta}\left(\mathbb{P} \nabla \ell\left(\theta^{t} ; z\right)\right)^{T}\left(\mathbb{P}_{n} \nabla \ell\left(\theta^{t} ; z\right)-\mathbb{P} \nabla \ell\left(\theta^{t} ; z\right)\right) \\
+\frac{1}{2 \beta}\left\|\mathbb{P} \nabla \ell\left(\theta^{t} ; z\right)+\left(\mathbb{P}_{n} \nabla \ell\left(\theta^{t} ; z\right)-\mathbb{P} \nabla \ell\left(\theta^{t} ; z\right)\right)\right\|^{2} \\
=-\frac{1}{2 \beta}\left\|\mathbb{P} \nabla \ell\left(\theta^{t} ; z\right)\right\|^{2}+\frac{1}{2 \beta}\left\|\mathbb{P}_{n} \nabla \ell\left(\theta^{t} ; z\right)-\mathbb{P} \nabla \ell\left(\theta^{t} ; z\right)\right\|^{2} \\
\leq-\frac{\mu}{\beta}\left(\mathbb{P} \ell\left(\theta^{t} ; z\right)-\mathbb{P} \ell\left(\theta^{*} ; z\right)\right)+\frac{1}{2 \beta}\left\|\mathbb{P}_{n} \nabla \ell\left(\theta^{t} ; z\right)-\mathbb{P} \nabla \ell\left(\theta^{t} ; z\right)\right\|^{2} .
\end{array}
$$

Rearranging the above inequality, and subtracting $\mathbb{P} \ell\left(\theta^{*} ; z\right)$ from both sides, we obtain

$$
\begin{equation*}
\mathbb{P} \ell\left(\theta^{t+1} ; z\right)-\mathbb{P} \ell\left(\theta^{*} ; z\right) \leq\left(1-\frac{\mu}{\beta}\right)\left(\mathbb{P} \ell\left(\theta^{t} ; z\right)-\mathbb{P} \ell\left(\theta^{*} ; z\right)\right)+\frac{1}{2 \beta}\left\|\mathbb{P}_{n} \nabla \ell\left(\theta^{t} ; z\right)-\mathbb{P} \nabla \ell\left(\theta^{t} ; z\right)\right\|^{2} \tag{A.3.11}
\end{equation*}
$$

Applying Proposition 8, we continue the proof on the event

$$
\mathcal{A}:=\{\text { the results (A.3.4) (A.3.5) in Proposition } 8 \text { hold true }\}
$$

whose measure is at least $1-\delta$. We keep the notations "term I" and "term II" used in Proposition 8 , which are defined by

$$
\begin{array}{r}
\text { term I }:=\sqrt{\frac{2 \mathbb{P}\left[\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{2}\right] \log \frac{4}{\delta}}{n}}+\frac{G_{*} \log \frac{4}{\delta}}{n}, \\
\text { term II }:=\sqrt{\frac{d+\log \frac{8 \log _{2}\left(2 n \Delta_{M}+2\right)}{\delta}}{n}}+\frac{d+\log \frac{\log _{2}\left(2 n \Delta_{M}+2\right)}{\delta}}{n} .
\end{array}
$$

The result (A.3.5) in Proposition 8 implies that $\forall \theta \in \Theta$,

$$
\begin{array}{r}
\left\|\mathbb{P}_{n} \nabla \ell(\theta ; z)-\mathbb{P} \nabla \ell(\theta ; z)\right\| \leq \operatorname{term} \mathrm{I}+c_{0} \beta \max \left\{\left\|\theta-\theta^{*}\right\|, \frac{1}{n}\right\} \cdot \text { term II } \\
\leq\left(\operatorname{term~I}+\frac{c_{0} \beta}{n} \cdot \operatorname{term} \mathrm{II}\right)+c_{0} \beta \cdot \operatorname{term~II} \cdot\left\|\theta-\theta^{*}\right\|
\end{array}
$$

where $c_{0}$ is an absolute constant. Since the PL condition implies (see Lemma 18) that

$$
\mathbb{P} \ell(\theta ; z)-\mathbb{P} \ell\left(\theta^{*} ; z\right) \geq \frac{\mu}{2}\left\|\theta-\theta^{*}\right\|^{2}, \quad \forall \theta \in \Theta
$$

we have

$$
\begin{align*}
\| \mathbb{P}_{n} \nabla \ell(\theta ; z)- & \mathbb{P} \nabla \ell(\theta ; z) \|^{2} \leq 2\left(\text { term I }+\frac{c_{0} \beta}{n} \cdot \text { term II }\right)^{2} \\
+ & \frac{4 c_{0}^{2} \beta^{2}}{\mu}\left(\mathbb{P} \ell(\theta ; z)-\mathbb{P} \ell\left(\theta^{*} ; z\right)\right)(\text { term II })^{2} . \tag{A.3.12}
\end{align*}
$$

Combining (A.3.11) and (A.3.12), we have that for all $t=0,1, \ldots$,

$$
\begin{gathered}
\mathcal{E}\left(\theta^{t+1}\right) \leq\left(1-\frac{\mu}{\beta}\right) \mathcal{E}\left(\theta^{t}\right)+\varepsilon^{t}(n), \\
\varepsilon^{t}(n) \leq \alpha(n) \mathcal{E}\left(\theta^{t}\right)+\varepsilon^{*}(n)
\end{gathered}
$$

where

$$
\begin{array}{r}
\varepsilon^{t}(n)=\frac{1}{2 \beta}\left\|\mathbb{P}_{n} \nabla \ell\left(\theta^{t} ; z\right)-\mathbb{P} \nabla \ell\left(\theta^{t} ; z\right)\right\|^{2} \\
\alpha(n)=\frac{2 c_{0}^{2} \beta}{\mu}(\text { term II })^{2} \\
\varepsilon^{*}(n)=\frac{1}{\beta}\left(\text { term I }+\frac{c_{0} \beta}{n} \cdot \text { term II }\right)^{2}
\end{array}
$$

Consider the following two conditions on the sample size (note that they will be satisfied as
long as $n$ is large enough):

$$
\begin{array}{r}
\alpha(n) \leq \frac{\mu}{2 \beta}, \\
\varepsilon^{*}(n) \leq \frac{\mu^{2}}{4 \beta} \Delta_{m}^{2} . \tag{A.3.14}
\end{array}
$$

Now consider the condition

$$
n \geq \frac{c \beta^{2}}{\mu^{2}}\left(d+\log \frac{8 \log _{2}\left(2 n \Delta_{M}+2\right)}{\delta}\right)
$$

where $c=\max \left\{16 c_{0}^{2}, 1\right\}$ is an absolute constant. Then (A.3.13) holds. Since we also require the sample size $n$ to be large enough such that the "statistical error" term in (1.5.9) is smaller than $\frac{\mu}{2} \Delta_{m}^{2}$, the condition (A.3.14) is also true because

$$
\frac{\mu}{2} \Delta_{m}^{2} \geq \frac{16 \mathbb{P}\left[\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{2}\right] \log \frac{4}{\delta}}{\mu n}+\frac{8 G_{*}^{2} \log ^{2} \frac{4}{\delta}+\mu^{2}}{\mu n^{2}} \geq \frac{2}{\mu}\left(\text { term } \mathrm{II}^{2}+\frac{\mu}{2 n}\right)^{2} \geq \frac{2 \beta}{\mu} \cdot \varepsilon^{*}(n) .
$$

Therefore, both condition (A.3.13) and condition (A.3.14) hold true under Theorem 4's requirement on the sample size.

Since both (A.3.13) and (A.3.14) are true, we can use induction to prove that with probability at least $1-\delta$, for all $t=0,1, \ldots$,

$$
\mathcal{E}\left(\theta^{t}\right) \leq \frac{\mu}{2} \Delta_{m}^{2}
$$

Therefore, for all $t=0,1, \ldots$,

$$
\theta^{t} \in \mathcal{B}^{d}\left(\theta^{*}, \Delta_{m}\right) \subseteq \Theta .
$$

We choose the "Lyapunov function" in Proposition 9 to be the excess risk function $\mathcal{E}(\theta)$. Applying Proposition 9, we obtain that: when the sample size $n$ is large enough such that the condi-
tions (A.3.13) and (A.3.14) hold ture, we have

$$
\begin{aligned}
\mathbb{P} \ell\left(\theta^{t} ; z\right)-\mathbb{P} \ell\left(\theta^{*} ; z\right) & \leq\left(1-\frac{\mu}{2 \beta}\right)^{t} \mathcal{E}\left(\theta^{0}\right)+\frac{2 \beta}{\mu} \cdot \varepsilon^{*}(n) \\
& \leq\left(1-\frac{\mu}{2 \beta}\right)^{t} \mathcal{E}\left(\theta^{0}\right)+\frac{2}{\mu}\left(\operatorname{term} \mathrm{I}+\frac{\mu}{2 n}\right)^{2}, \\
& \leq\left(1-\frac{\mu}{2 \beta}\right)^{t} \mathcal{E}\left(\theta^{0}\right)+\frac{16 \mathbb{P}\left[\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{2}\right] \log \frac{8}{\delta}}{\mu n}+\frac{8 G_{*}^{2} \log ^{2} \frac{4}{\delta}+\mu^{2}}{\mu n^{2}} .
\end{aligned}
$$

This completes the proof of Theorem 4.

## A.3.5 Proof of Corollary 5

We first verify Assumption 1. We have

$$
\left.\nabla^{2} \ell(\theta ; z)=2\left(\eta^{\prime}\left(\theta^{T} x\right)^{2}+\left(\eta\left(\theta^{T} x\right)-y\right) \eta^{\prime \prime}\left(\theta^{T} x\right)\right]\right) x x^{T}
$$

Since $x$ is $\tau$-sub-Gaussian, $x x^{T}$ is a $\tau^{2}$-sub-exponential. From the fact

$$
\left|2\left(\eta^{\prime}\left(\theta^{T} x\right)^{2}+\left(\eta\left(\theta^{T} x\right)-y\right) \eta^{\prime \prime}\left(\theta^{T} x\right)\right)\right| \leq C_{\eta}\left(B+C_{\eta}\right),
$$

Assumption 1 holds with $\beta=2 C_{\eta}\left(C_{\eta}+\sqrt{B}\right) \tau^{2}$.
We then verify Assumption 2. We know

$$
\nabla \ell\left(\theta^{*} ; z\right)=2\left(\eta\left(x^{T} \theta^{*}\right)-y\right) \eta^{\prime}\left(x^{T} \theta^{*}\right) x
$$

So we have for all $z$,

$$
\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\| \leq \sqrt{d}\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|_{\infty} \leq 2 C_{\eta} \sqrt{B d} .
$$

So Assumption 2 holds with $G_{*}=2 C_{\eta} \sqrt{B d}$.
Lastly, by the inequality (16) in Lemma 5 from [20], Assumption 3 holds with $\mu=\frac{2 c_{\eta}^{3} \tau^{2} \gamma}{C_{\eta}}$. This
completes the proof.

## A.3.6 Proof of Theorem 6

Before proving Theorem 6, we refer to Theorem 1 in [21] for the following result on the population-based first-order EM update.

Lemma 19 (linear convergence of population-based first-order EM) Under Assumption 5, Assumption 6 and the condition that $\mathbb{P} \ell(\theta ; z)$ is $\beta$-smooth, the following update,

$$
\theta^{+}=\theta-\frac{2}{\beta+\mu_{1}} \mathbb{P} \nabla \ell_{\theta}(\theta ; z)
$$

satisfies that

$$
\left\|\theta^{+}-\theta^{*}\right\| \leq\left(1-\frac{2 \mu_{1}-\mu_{2}}{\beta+\mu_{1}}\right)\left\|\theta-\theta^{*}\right\| .
$$

We now prove Theorem 6.

Proof of Theorem 6: Assumption 1 implies that $\mathbb{P} \ell(\theta ; z)$ is $\beta$-smooth, so Lemma 19 holds under the assumptions of Theorem 6. Now we turn to analyze the sample-based first-order EM. Consider the update of sample-based first-order EM,

$$
\theta^{t+1}=\theta^{t}-\frac{2}{\beta+\mu_{1}} \mathbb{P}_{n} \nabla \ell_{\theta^{t}}\left(\theta^{t} ; z\right), \quad t=0,1, \ldots
$$

Fix $t \geq 0$. We have

$$
\begin{align*}
\left\|\theta^{t+1}-\theta^{*}\right\| & \leq\left\|\theta^{t}-\frac{2}{\beta+\mu_{1}} \mathbb{P} \nabla \ell_{\theta^{t}}\left(\theta^{t} ; z\right)\right\|+\frac{2}{\beta+\mu_{1}}\left\|\left(\mathbb{P}-\mathbb{P}_{n}\right) \nabla \ell_{\theta^{t}}\left(\theta^{t} ; z\right)\right\| \\
& \leq\left(1-\frac{2 \mu_{1}-\mu_{2}}{\beta+\mu_{1}}\right)\left\|\theta^{t}-\theta^{*}\right\|+\frac{2}{\beta+\mu_{1}}\left\|\left(\mathbb{P}-\mathbb{P}_{n}\right) \nabla \ell\left(\theta^{t} ; z\right)\right\| . \tag{A.3.15}
\end{align*}
$$

Applying Proposition 8, we continue the proof on the event

$$
\mathcal{A}:=\{\text { the results (A.3.4) (A.3.5) in Proposition } 8 \text { hold true }\}
$$

whose measure is at least $1-\delta$. We keep the notations "term I" and "term II" used in Proposition 8 , which are defined by

$$
\begin{array}{r}
\text { term I }:=\sqrt{\frac{2 \mathbb{P}\left[\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{2}\right] \log \frac{4}{\delta}}{n}}+\frac{G_{*} \log \frac{4}{\delta}}{n}, \\
\text { term II }:=\sqrt{\frac{d+\log \frac{8 \log _{2}\left(2 n \Delta_{M}+2\right)}{\delta}}{n}}+\frac{d+\log \frac{\log _{2}\left(2 n \Delta_{M}+2\right)}{\delta}}{n} .
\end{array}
$$

Note that we have the optimality condition $\nabla \ell\left(\theta^{*} ; z\right)=0$, because the true parameter $\theta^{*}$ is assumed to minimizes the population risk over $\mathbb{R}^{d}$ in the problem setting. The result (A.3.5) in Proposition 8 implies that $\forall \theta \in \Theta$,

$$
\begin{array}{r}
\left\|\mathbb{P}_{n} \nabla \ell(\theta ; z)-\mathbb{P} \nabla \ell(\theta ; z)\right\| \leq \operatorname{term} \mathrm{I}+c_{0} \beta \max \left\{\left\|\theta-\theta^{*}\right\|, \frac{1}{n}\right\} \cdot \text { term II } \\
\leq\left(\operatorname{term~I}+\frac{c_{0} \beta}{n} \cdot \operatorname{term~II}\right)+c_{0} \beta \cdot \operatorname{term~II} \cdot\left\|\theta-\theta^{*}\right\| \tag{A.3.17}
\end{array}
$$

where $c_{0}$ is an absolute constant. Therefore, we have that for all $t=0,1, \ldots$,

$$
\begin{aligned}
\mathcal{E}\left(\theta^{t+1}\right) & \leq\left(1-\frac{2 \mu_{1}-\mu_{2}}{\beta+\mu_{1}}\right) \mathcal{E}\left(\theta^{t}\right)+\varepsilon^{t}(n), \\
\varepsilon^{t}(n) & \leq \alpha(n) \mathcal{E}\left(\theta^{t}\right)+\varepsilon^{*}(n),
\end{aligned}
$$

where

$$
\begin{array}{r}
\varepsilon^{t}(n)=\frac{2}{\beta+\mu_{1}}\left\|\left(\mathbb{P}-\mathbb{P}_{n}\right) \nabla \ell\left(\theta^{t} ; z\right)\right\|, \\
\alpha(n)=\frac{2 c_{0} \beta}{\beta+\mu_{1}} \cdot \text { term II, } \\
\varepsilon^{*}(n)=\frac{2}{\beta+\mu_{1}}\left(\text { term I }+\frac{c_{0} \beta}{n} \cdot \text { term II }\right) .
\end{array}
$$

Consider the following two conditions on the sample size (note that they will be satisfied as long as $n$ is large enough):

$$
\begin{gather*}
\alpha(n) \leq \frac{2 \mu_{1}-\mu_{2}}{2\left(\beta+\mu_{1}\right)}  \tag{A.3.18}\\
\varepsilon^{*}(n) \leq \frac{2 \mu_{1}-\mu_{2}}{2\left(\beta+\mu_{1}\right)} \Delta_{m} \tag{A.3.19}
\end{gather*}
$$

When the sample size $n$ is large enough so that both (A.3.18) and (A.3.19) are true, we can use induction to prove that with probability at least $1-\delta$, for all $t=0,1, \ldots$,

$$
\left\|\theta^{t}-\theta^{*}\right\| \leq \Delta_{m}^{2}
$$

Therefore, for all $t=0,1, \ldots$,

$$
\theta^{t} \in \mathcal{B}^{d}\left(\theta^{*}, \Delta_{m}\right) \subseteq \Theta
$$

We choose the "Lyapunov function" in Proposition 9 to be $\left\|\theta-\theta^{*}\right\|$. Applying Proposition 9, we obtain: when the sample size $n$ is large enough such that the conditions (A.3.18) and (A.3.19) hold ture, we have

$$
\begin{align*}
\| \theta^{t}- & \theta^{*}\left\|\leq\left(1-\frac{2 \mu_{1}-\mu_{2}}{2\left(\beta+\mu_{1}\right)}\right)^{t}\right\| \theta^{0}-\theta^{*} \|+\frac{2\left(\beta+\mu_{1}\right)}{2 \mu_{1}-\mu_{2}} \cdot \varepsilon^{*}(n) \\
& \leq\left(1-\frac{2 \mu_{1}-\mu_{2}}{2\left(\beta+\mu_{1}\right)}\right)^{t}\left\|\theta^{0}-\theta^{*}\right\|+\frac{4}{2 \mu_{1}-\mu_{2}} \cdot \operatorname{term} \mathrm{I}+\frac{2}{n} \tag{A.3.20}
\end{align*}
$$

When the sample size is large enough such that
$n \geq \frac{c \beta^{2}}{\left(2 \mu_{1}-\mu_{2}\right)^{2}}\left(d+\log \frac{8 \log _{2}\left(2 n \Delta_{M}+2\right)}{\delta}\right) \quad$ and $\quad$ term $\mathrm{I}+\frac{2 \mu_{1}-\mu_{2}}{2 n} \leq \frac{\left(2 \mu_{1}-\mu_{2}\right) \Delta_{m}}{4}$,
where $c=\max \left\{64 c_{0}^{2}, 1\right\}$ is an absolute constant, we have

$$
\text { term } \mathrm{I} \leq \frac{2 \mu_{1}-\mu_{2}}{4}\left(\Delta_{m}-\frac{2}{n}\right) \quad \text { and } \quad \text { term II } \leq \frac{2 \mu_{1}-\mu_{2}}{4 c_{0} \beta}
$$

which further guarantee that both the condition (A.3.18) and the condition (A.3.19) are true. We conclude that when the sample size condition (A.3.21) is true, we have the bound (A.3.20).

Now we use the fact $\mu_{1} \leq 2 \mu_{1}-\mu_{2} \leq 2 \mu_{1}$ to simplify the sample size condition (A.3.21) and the bound (A.3.20). It is straightforward to verify that the sample size condition (A.3.21) will be satisfied when

$$
\begin{equation*}
n \geq \max \left\{\frac{c \beta^{2}}{\mu_{1}^{2}}\left(d+\log \frac{8 \log _{2}\left(2 n \Delta_{M}+2\right)}{\delta}\right), \quad \frac{128 \mathbb{P}\left[\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{2}\right] \log \frac{4}{\delta}}{\mu_{1} \Delta_{M}}, \quad \frac{8 G_{*} \log \frac{4}{\delta}+8 \mu_{1}}{\mu_{1} \Delta_{M}}\right\} \tag{A.3.22}
\end{equation*}
$$

and the bound (A.3.20) implies

$$
\begin{equation*}
\left\|\theta^{t}-\theta^{*}\right\| \leq \frac{4}{\mu_{1}}\left(\sqrt{\frac{2 \mathbb{P}\left[\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{2}\right] \log \frac{4}{\delta}}{n}}+\frac{G_{*} \log \frac{4}{\delta}+\mu_{1}}{n}\right)+\left(1-\frac{2 \mu_{1}-\mu_{2}}{2\left(\beta+\mu_{1}\right)}\right)^{t}\left\|\theta^{0}-\theta^{*}\right\| \tag{A.3.23}
\end{equation*}
$$

Since we always have

$$
\mathcal{E}\left(\theta^{t}\right) \leq \frac{\beta}{2}\left\|\theta^{t}-\theta^{*}\right\|^{2}
$$

the bound (A.3.23) will imply

$$
\begin{equation*}
\mathcal{E}\left(\theta^{t}\right) \leq \frac{16 \beta}{\mu_{1}^{2}}\left(\sqrt{\frac{2 \mathbb{P}\left[\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{2}\right] \log \frac{4}{\delta}}{n}}+\frac{G_{*} \log \frac{4}{\delta}+\mu_{1}}{n}\right)^{2}+\left(1-\frac{2 \mu_{1}-\mu_{2}}{2\left(\beta+\mu_{1}\right)}\right)^{2 t} \beta\left\|\theta^{0}-\theta^{*}\right\|^{2} \tag{A.3.24}
\end{equation*}
$$

Clearly, the sample size condition (A.3.22) and the bounds (A.3.23) (A.3.24) are identical to those presented in the statement of the theorem. This completes the proof.

## A.3.7 Proof of Corollary 7

For both examples, verification of Assumptions 1,2 and 5 is trivial. The parameters can be specified as $\beta=1, G_{*}=\sigma \sqrt{d}$, and $\mu_{1}=1$.

As for verification of Assumption 6, we refer to the following results that are direct consequence of [21].

Lemma 20 (verification of Assumption 6) (a) Lemma 2 in [21]: Consider Example 4 under the SNR condition (1.7.6), where $\eta$ is a sufficiently large constant such that $\eta>\frac{4 \sqrt{3}}{3}$ and $c_{1}\left(1+\frac{1}{\eta^{2}}+\right.$ $\left.\eta^{2}\right) e^{-c_{2} \eta^{2}}<1$. Then Assumption 6 holds with $\mu_{2}=c_{1}\left(1+\frac{1}{\eta^{2}}+\eta^{2}\right) e^{-c_{2} \eta^{2}}$. Here $c_{1}$ and $c_{2}$ denote the same absolute constants as in the proof of Lemma 2 in [21]. Clearly, we can verify Assumption 6 for all $\eta$ larger than a certain absolute constant.
(b) Lemma 3 in [21]: Consider Example 5 under the SNR condition (1.7.6), where $\eta$ is a sufficiently large constant such that

$$
\begin{array}{r}
c \eta^{1-\frac{c_{\tau}^{2}}{2}}+c_{\tau}^{2} \frac{\log \eta}{\eta}+\frac{2}{\eta}
\end{array} \leq \frac{1}{8}, ~ 子 \sqrt{\frac{\left\|\theta^{*}\right\|}{8 \eta}+\left(4+\frac{2}{31}\right) \frac{C_{\tau}^{2} \log \eta}{\eta}+3 \eta^{1-\frac{C_{\tau}^{2}}{2}}} \leq \frac{1}{8}
$$

hold true for some sufficiently large constants $c_{\tau}, C_{\tau}$ and an absolute constant $c$. Then Assumption 6 holds with $\mu_{2}=\frac{1}{4}$. Here $c, c_{\tau}, C_{\tau}$ denote the same quantity as in the proof of Lemma 3 in [21].

## Clearly, we can verify Assumption 6 for all $\eta$ larger than a certain absolute constant.

To prove the generalization error bound in this corollary, we need to upper bound the problemdependent parameter $\mathbb{P}\left[\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{2}\right]$ for the two examples.

Bounding $\mathbb{P}\left[\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{2}\right]$ for Example 4: we define the function $g: \mathbb{R}^{d} \rightarrow(0,2)$ as

$$
g(u)=\frac{2 e^{-\frac{\left\|2 \theta^{*}-u\right\|^{2}}{2 \sigma^{2}}}}{e^{-\frac{\|u\|^{2}}{2 \sigma^{2}}}+e^{-\frac{\left\|2 \theta^{*}-u\right\|^{2}}{2 \sigma^{2}}}}=\frac{2}{e^{\frac{2\left\|\theta^{*}\right\|^{2}-2 u^{T} \theta^{*}}{\sigma^{2}}}+1} .
$$

In Example 4, when conditioned on $w=1$ (i.e., when $z$ is drawn from $N\left(\theta^{*}, \sigma^{2} I_{d \times d}\right)$ ), the random gradient $\nabla \ell\left(\theta^{*} ; z\right)$ at $\theta^{*}$ can be shown to be equal to

$$
\left(\nabla \ell\left(\theta^{*} ; z\right) \mid w=1\right)=u \underbrace{\left.\frac{e^{-\frac{\|u\|^{2}}{2 \sigma^{2}}}-e^{-\frac{\left\|2 \theta^{*}-u\right\|^{2}}{2 \sigma^{2}}}}{e^{-\frac{\|u\|^{2}}{2 \sigma^{2}}}+e^{-\frac{\left\|2 \theta^{*}-u\right\|^{2}}{2 \sigma^{2}}}}\right)}_{1-g(u)}+\theta^{*} \underbrace{\left(\frac{2 e^{-\frac{\left\|2 \theta^{*}-u\right\|^{2}}{2 \sigma^{2}}}}{e^{-\frac{\|u\|^{2}}{2 \sigma^{2}}}+e^{-\frac{\left\|2 \theta^{*}-u\right\|^{2}}{2 \sigma^{2}}}}\right)}_{g(u)},
$$

where $u=\theta^{*}-z$ is a random vector drawn from $N\left(0, \sigma^{2} I_{d \times d}\right)$. And when conditioned on $w=-1$ (i.e., when $z$ is drawn from $N\left(0, \sigma^{2} I_{d \times d}\right)$ ), $\nabla \ell\left(\theta^{*} ; z\right)$ can be shown to be equal to

$$
\left(\nabla \ell\left(\theta^{*} ; z\right) \mid w=-1\right)=v \underbrace{\left(\frac{e^{-\frac{\|v\|^{2}}{2 \sigma^{2}}}-e^{-\frac{\left\|2 \theta^{*}-v\right\|^{2}}{2 \sigma^{2}}}}{e^{-\frac{\|v\|^{2}}{2 \sigma^{2}}}+e^{-\frac{\left\|2 \theta^{*}-v\right\|^{2}}{2 \sigma^{2}}}}\right)}_{1-g(v)}+\theta^{*} \underbrace{\left(\frac{2 e^{-\frac{\left\|2 \theta^{*}-v\right\|^{2}}{2 \sigma^{2}}}}{e^{-\frac{\|v\|^{2}}{2 \sigma^{2}}}+e^{-\frac{\left\|2 \theta^{*}-v\right\|^{2}}{2 \sigma^{2}}}}\right)}_{g(v)},
$$

where $v=\theta^{*}+z$ is a random vector drawn from $N\left(0, \sigma^{2} I_{d \times d}\right)$.
Therefore, we have

$$
\begin{align*}
& \mathbb{P}\left[\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{2}\right] \\
& =\frac{1}{2} \mathbb{E}\left[\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{2} \mid w=1\right]+\frac{1}{2} \mathbb{E}\left[\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{2} \mid w=-1\right] \\
& =\frac{1}{2} \mathbb{E}_{u}\left[\left\|u \cdot(1-g(u))+\theta^{*} \cdot g(u)\right\|^{2}\right]+\frac{1}{2} \mathbb{E}_{v}\left[\left\|v \cdot(1-g(v))+\theta^{*} \cdot g(v)\right\|^{2}\right] \\
& =\mathbb{E}_{u}\left[\left\|u \cdot(1-g(u))+\theta^{*} \cdot g(u)\right\|^{2}\right], \tag{A.3.25}
\end{align*}
$$

where the notation $\mathbb{E}_{u}$ means taking expectation with respect to $u \sim N\left(0, \sigma^{2} I_{d \times d}\right)$, and the notation $\mathbb{E}_{v}$ means taking expectation with respect to $v \sim N\left(0, \sigma^{2} I_{d \times d}\right)$.

Since $0<g(u)<2$, we have $|1-g(u)| \leq 1$. Thus

$$
\begin{align*}
& \left\|u \cdot(1-g(u))+\theta^{*} \cdot g(u)\right\|^{2} \\
& \leq 2\|u\|^{2} \cdot|1-g(u)|^{2}+2\left\|\theta^{*}\right\|^{2} \cdot|g(u)|^{2} \\
& =2\|u\|^{2}+2\left\|\theta^{*}\right\|^{2} \cdot g(u)^{2} . \tag{A.3.26}
\end{align*}
$$

From (A.3.25) and (A.3.26), we have

$$
\begin{array}{r}
\mathbb{P}\left[\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{2}\right] \leq 2 \mathbb{E}_{u}\left[\|u\|^{2}\right]+2\left\|\theta^{*}\right\|^{2} \mathbb{E}_{u}\left[g(u)^{2}\right] \\
=2 \sigma^{2} d+2\left\|\theta^{*}\right\|^{2} \mathbb{E}_{u}\left[g(u)^{2}\right] . \tag{A.3.27}
\end{array}
$$

Now we know that $u^{T} \theta^{*}$ is a $\left\|\theta^{*}\right\| \sigma$-sub-Gaussian vector with mean 0 . From Markov's inequality,

$$
\begin{align*}
& \operatorname{Prob}\left(\left|u^{T} \theta^{*}\right|>\frac{1}{2}\left\|\theta^{*}\right\|^{2}\right) \leq 2 \exp \left(-\frac{\frac{1}{4}\left\|\theta^{*}\right\|^{4}}{\left\|\theta^{*}\right\|^{2} \sigma^{2}}\right) \\
&=\frac{2}{\exp \left(\frac{\left\|\theta^{*}\right\|^{2}}{4 \sigma^{2}}\right)} \leq \frac{8 \sigma^{2}}{\left\|\theta^{*}\right\|^{2}} \tag{A.3.28}
\end{align*}
$$

When $\left|u^{T} \theta^{*}\right| \leq \frac{1}{2}\left\|\theta^{*}\right\|^{2}$, we have

$$
g(u)=\frac{2}{e^{\frac{\left\|\theta^{*}\right\|^{2}-u^{T} \theta^{*}}{\sigma^{2}}}+1} \leq \frac{2}{e^{\frac{\left\|\theta^{*}\right\|^{2}}{2 \sigma^{2}}}} \leq \frac{4 \sigma^{2}}{\left\|\theta^{*}\right\|^{2}} .
$$

Since $0<g(u)<2$, when $\left|u^{T} \theta^{*}\right| \leq \frac{1}{2}\left\|\theta^{*}\right\|^{2}$, we have

$$
\begin{equation*}
g(u)^{2} \leq \frac{8 \sigma^{2}}{\left\|\theta^{*}\right\|^{2}} \tag{A.3.29}
\end{equation*}
$$

As a result,

$$
\begin{aligned}
& \mathbb{E}_{u} {\left[g(u)^{2}\right] } \\
& \leq \operatorname{Prob}\left(\left|u^{T} \theta^{*}\right|>\frac{1}{2}\left\|\theta^{*}\right\|^{2}\right) \mathbb{E}\left[g(u)^{2}| | u^{T} \theta^{*} \left\lvert\,>\frac{1}{2}\left\|\theta^{*}\right\|^{2}\right.\right] \\
&+\operatorname{Prob}\left(\left|u^{T} \theta^{*}\right| \leq \frac{1}{2}\left\|\theta^{*}\right\|^{2}\right) \mathbb{E}\left[g(u)^{2}| | u^{T} \theta^{*} \left\lvert\, \leq \frac{1}{2}\left\|\theta^{*}\right\|^{2}\right.\right] \\
& \leq 4 \cdot \operatorname{Prob}\left(\left|u^{T} \theta^{*}\right|>\frac{1}{2}\left\|\theta^{*}\right\|^{2}\right)+\frac{8 \sigma^{2}}{\left\|\theta^{*}\right\|^{2}} \\
& \leq \frac{40 \sigma^{2}}{\left\|\theta^{*}\right\|^{2}}
\end{aligned}
$$

where the second inequality is due to the fact $0<g(u)<2$ and (A.3.29), and the last inequality is due to (A.3.28). Combining the above result with (A.3.27), we have

$$
\begin{equation*}
\mathbb{P}\left[\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{2}\right] \leq(2 d+40) \sigma^{2} \tag{A.3.30}
\end{equation*}
$$

Bounding $\mathbb{P}\left[\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{2}\right]$ for Example 5: we define the function $g: \mathbb{R} \times \mathbb{R}^{d} \rightarrow(0,2)$ as

$$
g(u, x)=\frac{2 e^{-\frac{\left(2 x^{T} \theta^{*}-u\right)^{2}}{2 \sigma^{2}}}}{e^{-\frac{u^{2}}{2 \sigma^{2}}}+e^{-\frac{\left(2 x^{T} \theta^{*}-u\right)^{2}}{2 \sigma^{2}}}}=\frac{2}{e^{\frac{2\left(x^{T} \theta^{*}\right)^{2}-2 u\left(x^{T} \theta^{*}\right)}{\sigma^{2}}}+1}
$$

In Example 5, we have
where $u=x^{T} \theta^{*}-y$ is a random vector drawn from $N\left(0, \sigma^{2}\right)$. And we have
where $v=x^{T} \theta^{*}+y$ is a random vector drawn from $N\left(0, \sigma^{2}\right)$.
Therefore, we have

$$
\begin{align*}
& \mathbb{P}\left[\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{2}\right] \\
& =\frac{1}{2} \mathbb{E}\left[\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{2} \mid w=1\right]+\frac{1}{2} \mathbb{E}\left[\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{2} \mid w=-1\right] \\
& =\frac{1}{2} \mathbb{E}_{x}\left[\|x\|^{2} \mathbb{E}_{u}\left[\left(u \cdot(1-g(u, x))+\theta^{*} \cdot g(u, x)\right)^{2} \mid x\right]\right]+\frac{1}{2} \mathbb{E}_{x}\left[\|x\|^{2} \mathbb{E}_{v}\left[\left(v \cdot(1-g(v, x))+\theta^{*} \cdot g(v, x)\right)^{2}\right]\right] \\
& =\mathbb{E}_{x}\left[\|x\|^{2} \mathbb{E}_{u}\left[\left(u \cdot(1-g(u, x))+\theta^{*} \cdot g(u, x)\right)^{2} \mid x\right]\right] \tag{A.3.31}
\end{align*}
$$

where the notation $\mathbb{E}_{u}$ means taking expectation with respect to $u \sim N\left(0, \sigma^{2}\right)$, the notation $\mathbb{E}_{v}$ means taking expectation with respect to $v \sim N\left(0, \sigma^{2}\right)$, and the notation $\mathbb{E}_{x}$ means taking expectation with respect to $x \sim N\left(0, I_{d \times d}\right)$.

Similar to the last part (i.e., the proof of (A.3.30)), we can prove

$$
\mathbb{E}_{u}\left[\left(u \cdot(1-g(u, x))+\theta^{*} \cdot g(u, x)\right)^{2} \mid x\right] \leq 42 \sigma^{2}, \quad \forall x .
$$

Combine this result with (A.3.31), we obtain

$$
\mathbb{P}\left[\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{2}\right] \leq 42 \sigma^{2} \mathbb{E}_{x}\left[\|x\|^{2}\right]=42 \sigma^{2} d
$$

This gives an upper bound on $\mathbb{P}\left[\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{2}\right]$ in Example 5.
Given that we have upper bounded $\mathbb{P}\left[\left\|\nabla \ell\left(\theta^{*} ; z\right)\right\|^{2}\right]$ by $42 \sigma^{2} d$ in both Example 4 and Example 5 , it is straightforward to prove the generalization error bound in Corollary 7.
A.3.8 Auxiliary definitions and lemmata

Definition 13 (Orlicz norms, sub-Gaussian, sub-exponential) For every $\alpha \in(0,+\infty)$ we define the Orlicz- $\alpha$ norm of a random $u$ :

$$
\|u\|_{\text {Orlicz }_{\alpha}}=\inf \left\{K>0: \mathbb{E} \exp \left((|u| / K)^{\alpha}\right) \leq 2\right\} .
$$

A random variable/vector $X \in \mathbb{R}^{d}$ is $K$-sub-Gaussian if $\forall \lambda \in \mathbb{R}^{d}$, we have

$$
\left\|\lambda^{T} X\right\|_{\text {Orlicz }_{2}} \leq K\|\lambda\|_{2} .
$$

A random variable/vector $X \in \mathbb{R}^{d}$ is $K$-sub-exponential if $\forall \lambda \in \mathbb{R}^{d}$, we have

$$
\left\|\lambda^{T} X\right\|_{\text {Orlicz }_{1}} \leq K\|\lambda\|_{2}
$$

Lemma 21 (Bernstein's inequality for sub-exponential random variables) If $X_{1}, \cdots, X_{m}$ are subexponential random variables, then Bernstein's inequality (the inequality (A.2.49) in Lemma 17 holds with

$$
\sigma^{2}=\frac{1}{n} \sum_{i=1}^{n}\left\|X_{i}\right\|_{\text {Orlicz }_{1}}^{2}, \quad B=\max _{1 \leq i \leq n}\left\|X_{i}\right\|_{\text {Orlicz }_{1}} .
$$

Lemma 22 (vector Bernstein's inequality, $[140,141]$ ) Let $\left\{X_{i}\right\}_{i=1}^{n}$ be a sequence of i.i.d. random variables taking values in a real separable Hilbert space. Assume that $\mathbb{E}\left[X_{i}\right]=\mu, \mathbb{E}\left[\| X_{i}-\right.$ $\left.\mu \|^{2}\right]=\sigma^{2}, \forall 1 \leq i \leq n$. We say that vector Bernstein's condition with parameter $B$ holds if for all $1 \leq i \leq n$,

$$
\mathbb{E}\left[\left\|X_{i}-\mu\right\|^{k}\right] \leq \frac{1}{2} k!\sigma^{2} B^{k-2}, \quad \forall 2 \leq k \leq n .
$$

If this condition holds, then for all $\delta \in(0,1)$, with probability at least $1-\delta$ we have

$$
\left\|\frac{1}{n} \sum_{i=1}^{n} X_{i}-\mu\right\| \leq \sqrt{\frac{2 \sigma^{2} \log \frac{2}{\delta}}{n}}+\frac{B \log \frac{2}{\delta}}{n} .
$$

The following definitions and lemmata provide some background on generic chaining.

Definition 14 (Orlicz- $\alpha$ processes) Let $\left\{X_{f}\right\}_{f \in \mathcal{F}}$ be a sequence of random variables. $\left\{X_{f}\right\}_{f \in \mathcal{F}}$ is called an Orlicz- $\alpha$ process for a metric metr $(\cdot, \cdot)$ on $\mathcal{F}$ if

$$
\left\|X_{f_{1}}-X_{f_{2}}\right\|_{\text {Orlicz }_{\alpha}} \leq \operatorname{metr}\left(f_{1}, f_{2}\right), \forall f_{1}, f_{2} \in \mathcal{F} .
$$

In particular, Orlicz-2 process is called "process with sub-Gaussian increments" and Orlicz-1 process is called "process with sub-exponential increments".

Definition 15 (mixed sub-Gaussian-sub-exponential increments, [139]) We say a process $\left(X_{\theta}\right)_{\theta \in \Theta}$ has mixed sub-Gaussian-sub-exponential increments with respect to the pair $\left(\operatorname{metr}_{1}, \operatorname{metr}_{2}\right)$ if for all $\theta_{1}, \theta_{2} \in \Theta$,

$$
\operatorname{Prob}\left(\left\|X_{\theta_{1}}-X_{\theta_{2}}\right\| \geq \sqrt{u} \cdot \operatorname{metr}_{2}\left(\theta_{1}, \theta_{2}\right)+u \cdot \operatorname{met}_{1}\left(\theta_{1}, \theta_{2}\right)\right) \leq 2 e^{-u}, \forall u \geq 0 .
$$

Definition 16 (Talagrand's $\gamma_{\alpha}$-functional) A sequence $F=\left(\mathcal{F}_{n}\right)_{n \geq 0}$ of subsets of $\mathcal{F}$ is called admissible if $\left|\mathcal{F}_{0}\right|=1$ and $\left|\mathcal{F}_{n}\right| \leq 2^{2^{n}}$ for all $n \geq 1$. For any $0<\alpha<\infty$, the $\gamma_{\alpha}-$ functional of $(\mathcal{F}$, metr $)$ is defined by

$$
\gamma_{\alpha}(F, d)=\inf _{F} \sup _{f \in \mathcal{F}} \sum_{n=0}^{\infty} 2^{\frac{n}{\alpha}} \operatorname{metr}\left(f, \mathcal{F}_{n}\right),
$$

where the infimum is taken over all admissible sequences and we write $\operatorname{met} r\left(f, \mathcal{F}_{n}\right)=\inf _{s \in \mathcal{F}_{n}} \operatorname{metr}(f, s)$.

Lemma 23 (Dudley's integral bound for $\gamma_{\alpha}$ functional, [142]) There exist a constant $C_{\alpha}$ depend-
ing only on $\alpha$ such that

$$
\gamma_{\alpha}(\mathcal{F}, \text { met } r) \leq C_{\alpha} \int_{0}^{+\infty}(\log N(\varepsilon, \mathcal{F}, \text { metr }))^{\frac{1}{\alpha}} d \varepsilon
$$

Lemma 24 (generic chaining for a process with mixed tail increments, [139]) If $\left(X_{f}\right)_{f \in \mathcal{F}}$ has mixed sub-Gaussian-sub-exponential increments with respect to the pair (metr$r_{1}$, met $r_{2}$ ), then there are absolute constants $c, C>0$ such that $\forall \delta \in(0,1)$,

$$
\begin{array}{r}
\sup _{\theta \in \Theta}\left\|X_{f}-X_{f_{0}}\right\| \leq C\left(\gamma_{2}\left(\mathcal{F}, \text { met } \text { r }_{2}\right)+\gamma_{1}\left(\mathcal{F}, \text { met } \text { r }_{1}\right)\right)+ \\
c\left(\sqrt{\log \frac{1}{\delta}} \sup _{f_{1}, f_{2} \in \mathcal{F}}\left[\operatorname{metr}_{2}\left(f_{1}, f_{2}\right)\right]+\log \frac{1}{\delta} \sup _{f_{1}, f_{2} \in \mathcal{F}}\left[\operatorname{metr} r_{1}\left(f_{1}, f_{2}\right)\right]\right),
\end{array}
$$

with probability at least $1-\delta$.

## A. 4 Proofs for Section A. 1

## A.4.1 Proof of Theorem 20

The proof consists of five parts. Among them, the main purpose of Part I and Part IV is to localized the strong convexity parameter. When there is no need to localized the strong convexity parameter (e.g., when one uses the square cost), the proof can be simplified-Part I and Part IV will be quite straightforward, and all the "upper-side" truncation analysis related to $\frac{2\|\xi\|_{L_{2}}}{\sqrt{c_{\kappa}}}, \frac{4\|\xi\|_{L_{2}}^{2}}{c_{\kappa}}$ or $\frac{4\|\xi\|_{L_{2}}^{2}}{\kappa^{2} c_{\kappa}}$ will be unnecessary.

## Part I: analysis of the concentrated functions.

Denote $T(h)=\left\|h-h^{*}\right\|_{L_{2}}^{2}$ and

$$
v_{h}=\min \left\{\kappa\left\|h-h^{*}\right\|_{L_{2}}, \frac{2\|\xi\|_{L_{2}}}{\sqrt{c_{\kappa}}}\right\} .
$$

For every $h \in \mathcal{H}$, define

$$
\begin{aligned}
f_{h}(x, y) & =\frac{2}{\alpha\left(4\|\xi\|_{L_{2}} / \sqrt{c_{\kappa}}\right.} \partial_{1} \ell_{\mathrm{SV}}\left(h^{*}(x), y\right)\left(h(x)-h^{*}(x)\right), \\
g_{h}(x, y) & =\min \left\{\left(h(x)-h^{*}(x)\right)^{2}, v_{h}^{2}\right\} \cdot \mathbb{1}\left\{|\xi| \leq \frac{2\|\xi\|_{L_{2}}}{\sqrt{c_{\kappa}}}\right\} .
\end{aligned}
$$

One can view $g_{h}$ as a truncated version of the quadratic form $\left(h(x)-h^{*}(x)\right)^{2}$. Later we will use concentration to control $\left(\mathbb{P}-\mathbb{P}_{n}\right)\left(f_{h}+g_{h}\right)$ uniformly.

From Lemma 25 (for which we defer to the end of Section A.4.1), we can show

$$
\begin{array}{r}
\ell_{\mathrm{sv}}(h(x), y)-\ell_{\mathrm{sv}}\left(h^{*}(x), y\right)-\partial_{1} \ell_{\mathrm{sv}}\left(h^{*}(x), y\right)\left(h(x)-h^{*}(x)\right) \geq \frac{\alpha\left(2 v_{h}\right)}{2} \min \left\{\left(h(x)-h^{*}(x)\right)^{2}, v_{h}^{2}\right\} \\
\geq \frac{\alpha\left(4\|\xi\|_{L_{2}} / \sqrt{c_{k}}\right)}{2} g_{h}(x, y)
\end{array}
$$

The above inequality implies that

$$
\begin{equation*}
\mathbb{P}_{n}\left(f_{h}+g_{h}\right) \leq \frac{2}{\alpha\left(4\|\xi\|_{L_{2}} / \sqrt{c_{k}}\right)} \mathbb{P}_{n}\left[\ell_{\mathrm{sv}}(h(x), y)-\ell_{\mathrm{sv}}\left(h^{*}(x), y\right)\right] . \tag{A.4.1}
\end{equation*}
$$

Recall that $\xi=h^{*}(x)-y$. By Markov's inequality,

$$
\begin{equation*}
\operatorname{Prob}\left(|\xi| \geq \frac{2\|\xi\|_{L_{2}}}{\sqrt{c_{K}}}\right) \leq \frac{c_{K}}{4} \tag{A.4.2}
\end{equation*}
$$

From the definition of $g_{h}$ and $v_{h}$, it is straightforward to show that

$$
\begin{aligned}
& \mathbb{P} g_{h}=\mathbb{P}\left[\min \left\{\left(h(x)-h^{*}(x)\right)^{2}, v_{h}^{2}\right\} \cdot \mathbb{1}\left\{|\xi| \leq \frac{2\|\xi\|_{L_{2}}}{\sqrt{c_{\kappa}}}\right\}\right] \\
& \geq \mathbb{P}\left[\min \left\{\left(h(x)-h^{*}(x)\right)^{2}, v_{h}^{2}\right\} \cdot \mathbb{1}\left\{\left|h(x)-h^{*}(x)\right| \geq \kappa\left\|h-h^{*}\right\|_{L_{2}}\right\} \cdot \mathbb{1}\left\{|\xi| \leq \frac{2\|\xi\|_{L_{2}}}{\sqrt{c_{\kappa}}}\right\}\right] \\
& \geq \mathbb{P}\left[v_{h}^{2} \cdot \mathbb{1}\left\{\left|h(x)-h^{*}(x)\right| \geq \kappa\left\|h-h^{*}\right\|_{L_{2}}\right\} \cdot \mathbb{1}\left\{|\xi| \leq \frac{2\|\xi\|_{L_{2}}}{\sqrt{c_{\kappa}}}\right\}\right] \\
& =v_{h}^{2} \cdot \operatorname{Prob}\left(\left|h(x)-h^{*}(x)\right| \geq \kappa\left\|h-h^{*}\right\|_{L_{2}},|\xi| \leq \frac{2\|\xi\|_{L_{2}}}{\sqrt{c_{\kappa}}}\right) \\
& \geq v_{h}^{2} \cdot\left(\operatorname{Prob}\left(\left|h(x)-h^{*}(x)\right| \geq \kappa\left\|h-h^{*}\right\|_{L_{2}}\right)-\operatorname{Prob}\left(\| \xi \left\lvert\,>\frac{2\|\xi\|_{L_{2}}}{\sqrt{c_{K}}}\right.\right)\right) \\
& \geq \frac{3 c_{k}}{4} v_{h}^{2}
\end{aligned}
$$

where the first inequality is due to $1 \geq \mathbb{1}\left\{\left|h(x)-h^{*}(x)\right| \geq \kappa\left\|h-h^{*}\right\|_{L_{2}}\right\}$; the second inequality is due to the definition of $v_{h}$; and the last inequality is due to Assumption 11 and (A.4.2). From Assumption 10, we have

$$
\begin{equation*}
\mathbb{P}\left(f_{h}+g_{h}\right) \geq \mathbb{P} g_{h} \geq \frac{3 c_{\kappa}}{4} v_{h}^{2} \tag{A.4.3}
\end{equation*}
$$

Let us summarize the results from this part. We use the empirical average of the excess loss to upper bound $\mathbb{P}_{n}\left(f_{h}+g_{h}\right)$ in (A.4.1), and use the (truncated) quadratic form to lower bound $\mathbb{P}\left(f_{h}+g_{h}\right)$ in (A.4.3). The next steps are to prove concentration of $f_{h}$ and $g_{h}$ and establish a "uniform localized convergence" argument.

## Part II: bound the localized empirical process.

Given $r>0$, we want to bound the localized empirical process

$$
\sup _{\frac{r}{\lambda} \leq T(h) \leq r}\left(\mathbb{P}-\mathbb{P}_{n}\right)\left(f_{h}+g_{h}\right)
$$

where $\lambda>1$ is a fixed value that we will specify later. From the definition of $\varphi_{\text {noise }}(r ; \delta)$ in

Assumption 12 , for any $\delta \in(0,1)$, with probability $1-\frac{\delta}{2}$, we have

$$
\begin{equation*}
\sup _{\frac{r}{\lambda} \leq T(h) \leq r}\left(\mathbb{P}-\mathbb{P}_{n}\right)\left(f_{h}+g_{h}\right) \leq \sup _{\frac{r}{\lambda} \leq T(h) \leq r}\left(\mathbb{P}-\mathbb{P}_{n}\right) g_{h}+\frac{2}{\alpha\left(4\|\xi\|_{L_{2}} / \sqrt{c_{\kappa}}\right)} \varphi_{\text {noise }}\left(r ; \frac{\delta}{2}\right) \tag{A.4.4}
\end{equation*}
$$

Given $r>0$, denote the hypothesis class $\mathcal{H}\left(\frac{r}{\lambda}, r\right)=\left\{h \in \mathcal{H}: \frac{r}{\lambda} \leq T(h) \leq r\right\}$, and define the function $g_{h, r}$ as

$$
g_{h, r}(x, y)=\min \left\{\left(h(x)-h^{*}(x)\right)^{2}, \kappa^{2} r, \frac{4\|\xi\|_{L_{2}}^{2}}{c_{\kappa}}\right\} \cdot \mathbb{1}\left\{|\xi| \leq \frac{2\|\xi\|_{L_{2}}}{\sqrt{c_{\kappa}}}\right\} .
$$

Recall that $g_{h}$ is defined by

$$
g_{h}(x, y)=\min \left\{\left(h(x)-h^{*}(x)\right)^{2}, \kappa^{2} T(h), \frac{4\|\xi\|_{L_{2}}^{2}}{c_{\kappa}}\right\} \cdot \mathbb{1}\left\{|\xi| \leq \frac{2\|\xi\|_{L_{2}}}{\sqrt{c_{\kappa}}}\right\} .
$$

For every $h \in \mathcal{H}\left(\frac{r}{\lambda}, r\right)$ and any $(x, y) \in \mathcal{X} \times \mathcal{Y}$, we have

$$
\begin{equation*}
g_{h, \frac{r}{\lambda}}(x, y) \leq g_{h}(x, y) \leq g_{h, r}(x, y), \tag{A.4.5}
\end{equation*}
$$

and

$$
\begin{align*}
& g_{h, r}(x, y)-g_{h, \frac{r}{\lambda}}(x, y) \\
& \leq \min \left\{\left(h(x)-h^{*}(x)\right)^{2}, \kappa^{2} r, \frac{4\|\xi\|_{L_{2}}^{2}}{c_{k}}\right\}-\min \left\{\left(h(x)-h^{*}(x)\right)^{2}, \frac{\kappa^{2} r}{\lambda}, \frac{4\|\xi\|_{L_{2}}^{2}}{c_{\kappa}}\right\} \\
& \leq \min \left\{\kappa^{2} r, \frac{4\|\xi\|_{L_{2}}^{2}}{c_{k}}\right\}-\min \left\{\frac{\kappa^{2} r}{\lambda}, \frac{4\|\xi\|_{L_{2}}^{2}}{c_{\kappa}}\right\} \\
& \leq\left(1-\frac{1}{\lambda}\right) \min \left\{\kappa^{2} r, \frac{4\|\xi\|_{L_{2}}^{2}}{c_{\kappa}}\right\} . \tag{A.4.6}
\end{align*}
$$

From (A.4.5) and (A.4.6), for every $h \in \mathcal{H}\left(\frac{r}{\lambda}, r\right)$ and any $(x, y) \in \mathcal{X} \times \mathcal{Y}$,

$$
-\left(1-\frac{1}{\lambda}\right) \min \left\{\kappa^{2} r, \frac{4\|\xi\|_{L_{2}}^{2}}{c_{\kappa}}\right\} \leq g_{h}(x, y)-g_{h, r}(x, y) \leq 0
$$

which implies

$$
\left(\mathbb{P}-\mathbb{P}_{n}\right) g_{h} \leq\left(\mathbb{P}-\mathbb{P}_{n}\right) g_{h, r}+\left(1-\frac{1}{\lambda}\right) \min \left\{\kappa^{2} r, \frac{4\|\xi\|_{L_{2}}^{2}}{c_{\kappa}}\right\}
$$

As a result, we have

$$
\begin{align*}
\sup _{\frac{r}{K} \leq T(h) \leq r}\left(\mathbb{P}-\mathbb{P}_{n}\right) g_{h} \leq & \sup _{\frac{r}{K} \leq T(h) \leq r}\left(\mathbb{P}-\mathbb{P}_{n}\right) g_{h, r}+\left(1-\frac{1}{\lambda}\right) \min \left\{\kappa^{2} r, \frac{4\|\xi\|_{L_{2}}^{2}}{c_{\kappa}}\right\} \\
& \leq \sup _{T(h) \leq r}\left(\mathbb{P}-\mathbb{P}_{n}\right) g_{h, r}+\left(1-\frac{1}{\lambda}\right) \min \left\{\kappa^{2} r, \frac{4\|\xi\|_{L_{2}}^{2}}{c_{K}}\right\} . \tag{A.4.7}
\end{align*}
$$

We know that $g_{h, r}$ is uniformly bounded by $\left[0, \min \left\{\kappa^{2} r, \frac{4\|\xi\|_{L_{2}}^{2}}{c_{\kappa}}\right\}\right]$. Form the standard bound for global Rademacher complexity [3], $\forall \delta \in(0,1)$, with probability at least $1-\frac{\delta}{2}$,

$$
\begin{equation*}
\sup _{T(h) \leq r}\left(\mathbb{P}-\mathbb{P}_{n}\right) g_{h, r} \leq 2 \mathfrak{R}\left\{g_{h, r}: T(h) \leq r\right\}+\min \left\{\kappa^{2} r, \frac{4\|\xi\|_{L_{2}}^{2}}{c_{\kappa}}\right\} \sqrt{\frac{\log \frac{2}{\delta}}{2 n}} \tag{A.4.8}
\end{equation*}
$$

It is straightforward to verify that for all $h_{1}, h_{2} \in \mathcal{H}$ and $(x, y) \in \mathcal{X} \times \mathcal{Y}$,

$$
\left|g_{h_{1}, r}(x)-g_{h_{2}, r}(x)\right| \leq 2 \kappa \sqrt{r}\left|h_{1}(x)-h_{2}(x)\right| .
$$

From the Lipchitz contraction property of Rademacher complexity (see, e.g., Theorem 7 in [138]), we have

$$
\begin{equation*}
\mathfrak{R}\left\{g_{h, r}\right\} \leq 2 \kappa \sqrt{r} \Re\{h: T(h) \leq r\} \leq 2 \kappa \sqrt{r} \varphi(r), \tag{A.4.9}
\end{equation*}
$$

where $\varphi(r)$ is defined in Assumption 12. Define the $\psi$ function as

$$
\begin{equation*}
\psi(r ; \delta)=4 \kappa \sqrt{r} \varphi(r)+\left(\sqrt{\frac{\log \frac{2}{\delta}}{2 n}}+1-\frac{1}{\lambda}\right) \min \left\{\kappa^{2} r, \frac{4\|\xi\|_{L_{2}}^{2}}{c_{\kappa}}\right\}+\frac{2}{\alpha\left(4\|\xi\|_{L_{2}} / \sqrt{c_{\kappa}}\right)} \varphi_{\text {noise }}\left(r ; \frac{\delta}{2}\right) \tag{A.4.10}
\end{equation*}
$$

Combining the definition (A.4.10) with (A.4.4) (A.4.7) (A.4.8) (A.4.9), for any $\delta \in(0,1)$, with probability at least $1-\delta$, we have

$$
\begin{equation*}
\sup _{\frac{r}{K} \leq T(h) \leq r}\left(\mathbb{P}-\mathbb{P}_{n}\right)\left(f_{h}+g_{h}\right) \leq \psi(r ; \delta) . \tag{A.4.11}
\end{equation*}
$$

## Part III: the "uniform localized convergence" argument.

Applying Proposition 5 , for any $\delta_{1} \in(0,1)$ and $r_{0} \in\left(0,4 \Delta^{2}\right)$, with probability at least $1-\delta_{1}$, for all $h \in \mathcal{H}$, either $T(h) \leq r_{0}$ or

$$
\begin{align*}
& \left(\mathbb{P}-\mathbb{P}_{n}\right)\left(f_{h}+g_{h}\right) \leq \psi\left(\lambda T(h) ; \delta_{1}\left(\log _{K} \frac{4 K \Delta^{2}}{r_{0}}\right)^{-1}\right) \\
& =4 \kappa \sqrt{\lambda T(h)} \varphi(\lambda T(h))+\frac{2}{\alpha\left(4\|\xi\|_{L_{2}} / \sqrt{c_{\kappa}}\right)} \varphi_{\text {noise }}\left(\lambda T(h) ; \frac{\delta_{1}}{2 \log _{\lambda} \frac{4 \lambda \Delta^{2}}{r_{0}}}\right) \\
& \quad+\underbrace{\min \left\{\lambda \kappa^{2} T(h), \frac{4\|\xi\|_{L_{2}}^{2}}{c_{\kappa}}\right\}\left(\sqrt{\frac{\log \frac{2 \log _{\lambda} \frac{4 \lambda \Delta^{2}}{\delta_{0}}}{2 n}}{2 n}}+1-\frac{1}{\lambda}\right)}_{\text {last term in (A.4.12) }} . \tag{A.4.12}
\end{align*}
$$

We specify

$$
\lambda=\frac{8+2 c_{\kappa}}{8+c_{\kappa}}
$$

Then when $n>\frac{32}{c_{K}^{2}} \log \frac{2 \log _{\lambda} \frac{4 \lambda \lambda^{2}}{r_{0}}}{\delta_{1}}$, for all $h \in \mathcal{H}$ we have

$$
\lambda\left(\sqrt{\frac{\log \frac{2 \log _{\frac{2}{} \frac{4 \lambda \lambda^{2}}{r_{0}}}^{\delta_{1}}}{2 n}}{2 n}}+1-\frac{1}{\lambda}\right)<\frac{c_{K}}{4}
$$

which implies when $T(h)>0$,

$$
\begin{equation*}
\text { last term in }(\mathrm{A} .4 .12)<\frac{c_{\kappa}}{4} \min \left\{\kappa^{2} r, \frac{4\|\xi\|_{L_{2}}^{2}}{\lambda c_{\kappa}}\right\} \leq \frac{c_{\kappa}}{4} \min \left\{\kappa^{2} r, \frac{4\|\xi\|_{L_{2}}^{2}}{c_{\kappa}}\right\} . \tag{A.4.13}
\end{equation*}
$$

Denote $C_{r_{0}}=2+\left(\frac{16}{c_{\kappa}}+2\right) \log \frac{4 \Delta^{2}}{r_{0}}$, then

$$
2 \log _{\lambda} \frac{4 \lambda \Delta^{2}}{r_{0}}=2+\frac{\log \frac{4 \Delta^{2}}{r_{0}}}{\log \lambda} \leq 2+\left(\frac{16}{c_{\kappa}}+2\right) \log \frac{4 \Delta^{2}}{r_{0}}=C_{r_{0}} .
$$

For any $\delta \in(0,1)$, taking $\delta_{1}=\frac{2 \log _{\lambda} \frac{44 \Delta^{2}}{r_{0}}}{C_{r_{0}}} \delta$, from (A.4.12) (A.4.13) and the fact $\lambda<2$, we have the following conclusion: when $n>\frac{32}{c_{\kappa}^{2}} \log \frac{C_{r_{0}}}{\delta}$, with probability at least $1-\delta$, for all $h \in \mathcal{H}$, either $T(h) \leq r_{0}$ or

$$
\begin{align*}
& \left(\mathbb{P}-\mathbb{P}_{n}\right)\left(f_{h}+g_{h}\right) \\
& <4 \kappa \sqrt{2 T(h)} \varphi(2 T(h))+\frac{2}{\alpha\left(4\|\xi\|_{L_{2}} / \sqrt{c_{\kappa}}\right)} \varphi_{\text {noise }}\left(2 T(h) ; \frac{\delta}{C_{r_{0}}}\right)+\frac{c_{\kappa}}{4} \min \left\{\kappa^{2} T(h), \frac{4\|\xi\|_{L_{2}}^{2}}{c_{\kappa}}\right\} . \tag{A.4.14}
\end{align*}
$$

Let $\hat{h} \in \arg \min \mathbb{P}_{n} \ell_{\mathrm{sv}}(h(x), y)$ be the empirical risk minimizer. From (A.4.1) and the property of $\hat{h}$, we have

$$
\begin{equation*}
\mathbb{P}_{n}\left(f_{\hat{h}}+g_{\hat{h}}\right) \leq \frac{2}{\alpha\left(4\|\xi\|_{L_{2}} / \sqrt{c_{k}}\right)} \mathbb{P}_{n}\left[\ell_{\mathrm{SV}}(\hat{h}(x)-y)-\ell_{\mathrm{SV}}\left(h^{*}(x)-y\right)\right] \leq 0 \tag{A.4.15}
\end{equation*}
$$

Recall the result (A.4.3) proved in Part I,

$$
\begin{equation*}
\mathbb{P}\left(f_{h}+g_{h}\right) \geq \frac{3 c_{\kappa}}{4} v_{h}^{2}=\frac{3 c_{\kappa}}{4} \min \left\{\kappa^{2} T(h), \frac{4\|\xi\|_{L_{2}}^{2}}{c_{\kappa}}\right\} . \tag{A.4.16}
\end{equation*}
$$

From (A.4.14) (A.4.15) (A.4.16), when $n>\frac{32}{c_{\kappa}^{2}} \log \frac{C_{r_{0}}}{\delta}$, with probability at least $1-\delta$, either

$$
T(\hat{h}) \leq r_{0} \text { or }
$$

$$
\frac{3 c_{\kappa}}{4} \min \left\{\kappa^{2} T(\hat{h}), \frac{4\|\xi\|_{L_{2}}^{2}}{c_{\kappa}}\right\}
$$

$$
<4 \kappa \sqrt{2 T(\hat{h})} \varphi(2 T(\hat{h}))+\frac{2}{\alpha\left(4\|\xi\|_{L_{2}} / \sqrt{c_{\kappa}}\right)} \varphi_{\text {noise }}\left(2 T(\hat{h}) ; \frac{\delta}{C_{r_{0}}}\right)+\frac{c_{\kappa} \kappa^{2}}{4} \min \left\{T(\hat{h}), \frac{4\|\xi\|_{L_{2}}^{2}}{c_{\kappa}}\right\}
$$

i.e.,

$$
\begin{align*}
& \frac{c_{\kappa}}{2} \min \left\{\kappa^{2} T(\hat{h}), \frac{4\|\xi\|_{L_{2}}^{2}}{c_{\kappa}}\right\} \\
& <4 \kappa \sqrt{2 T(\hat{h})} \varphi(2 T(\hat{h}))+\frac{2}{\alpha\left(4\|\xi\|_{L_{2}} / \sqrt{c_{\kappa}}\right)} \varphi_{\text {noise }}\left(2 T(\hat{h}) ; \frac{\delta}{C_{r_{0}}}\right) . \tag{A.4.17}
\end{align*}
$$

In the theorem we have asked $n>\frac{72}{c_{\kappa}^{2}} \log \frac{C_{r_{0}}}{\delta}$. Denote the event

$$
\mathcal{A}:=\left\{\text { either } T(\hat{h}) \leq r_{0} \text { or (A.4.17) is true }\right\} .
$$

Then we have $\operatorname{Prob}(\mathcal{A}) \geq 1-\delta$.

## Part IV: preliminary localization.

We first prove a preliminary localization result $T(\hat{h}) \leq \max \left\{\frac{4\|\xi\|_{L_{2}}^{2}}{\kappa^{2} c_{\kappa}}, r_{0}\right\}$ on the event $\mathcal{A}$. The essential purpose of this step is to localize the strong convexity parameter. If $T(\hat{h}) \in\left(\max \left\{\frac{4\|\xi\|_{L_{2}}^{2}}{\kappa^{2} c_{\kappa}}, r_{0}\right\}, 4 \Delta^{2}\right]$ is true, then on the event $\mathcal{A}$ one have

$$
\begin{equation*}
\text { RHS of }(\mathrm{A} .4 .17)>2\|\xi\|_{L_{2}}^{2} . \tag{A.4.18}
\end{equation*}
$$

In the theorem we ask $n>\max \left\{\bar{N}_{\delta, r_{0}}, \frac{72}{c_{\kappa}^{2}} \log \frac{C_{r_{0}}}{\delta}\right\}$. According to Assumption 12, this implies that

$$
\begin{align*}
\varphi_{\text {noise }}\left(8 \Delta^{2} ; \frac{\delta}{C_{r_{0}}}\right) \leq & \frac{\alpha\left(4\|\xi\|_{L_{2}} / \sqrt{c_{\kappa}}\right)\|\xi\|_{L_{2}}^{2}}{2},  \tag{A.4.19}\\
& \varphi\left(8 \Delta^{2}\right) \leq \frac{\sqrt{2 c_{\kappa}}\|\xi\|_{L_{2}}^{2}}{16 \Delta}, \tag{A.4.20}
\end{align*}
$$

which further imply

$$
\begin{equation*}
\text { RHS of }(\text { A.4.17 }) \leq\|\xi\|_{L_{2}}^{2}+\|\xi\|^{2}=2\|\xi\|_{L_{2}}^{2} . \tag{A.4.21}
\end{equation*}
$$

(A.4.18) and (A.4.21) result in a contradiction. Therefore, $T(\hat{h})$ must be bounded by $\max \left\{\frac{4\|\xi\|_{L_{2}}^{2}}{\kappa^{2} c_{\kappa}}, r_{0}\right\}$. Then on the event $\mathcal{A}$, either $T(\hat{h}) \leq r_{0}$ or

$$
\begin{equation*}
\frac{c_{\kappa} \kappa^{2}}{2} T(\hat{h})<\text { RHS of (A.4.17) } \tag{A.4.22}
\end{equation*}
$$

## Part V: final steps.

Let $r_{\text {noise }}^{*}$ be the fixed point of

$$
\frac{4}{c_{\kappa} \kappa^{2} \cdot \alpha\left(4\|\xi\|_{L_{2}} / \sqrt{c_{\kappa}}\right)} \varphi_{\text {noise }}\left(2 r ; \frac{\delta}{C_{r_{0}}}\right),
$$

and $r_{\text {ver }}^{*}$ be the fixed point of

$$
\frac{8}{c_{\kappa} K} \sqrt{2 r} \varphi(2 r) .
$$

From the definition of fixed points, when $T(\hat{h})>\max \left\{r_{\text {ver }}^{*}, r_{\text {noise }}^{*}\right\}$, we have

$$
\frac{c_{\kappa} \kappa^{2}}{4} T(\hat{h})>\frac{2}{\alpha\left(4\|\xi\|_{L_{2}} / \sqrt{c_{\kappa}}\right)} \varphi_{\text {noise }}\left(2 T(\hat{h}) ; \frac{\delta}{C_{r_{0}}}\right)
$$

and

$$
\frac{c_{\kappa} \kappa^{2}}{4} T(\hat{h})>4 \kappa \sqrt{2 T(\hat{h})} \varphi(2 T(\hat{h})) .
$$

Contrasting the above two inequalities with our previous result (A.4.22), on the event $\mathcal{A}$ we have

$$
T(\hat{h}) \leq \max \left\{r_{\mathrm{ver}}^{*}, r_{\xi}^{*}, r_{0}\right\} .
$$

We conclude that when $n>\max \left\{\bar{N}_{\delta, r_{0}}, \frac{72}{c_{\kappa}^{2}} \log \frac{C_{r_{0}}}{\delta}\right\}$, with probability at least $1-\delta$,

$$
\begin{equation*}
\left\|\hat{h}-h^{*}\right\|_{L_{2}}^{2} \leq \max \left\{r_{\text {noise }}^{*}, r_{\text {ver }}^{*}, r_{0}\right\} \tag{A.4.23}
\end{equation*}
$$

Finally, from the optimality condition on $h^{*}$ (Assumption 10), it is straightforward to prove that for all $h \in \mathcal{H}$,

$$
\mathcal{E}(h) \leq \frac{\beta_{\mathrm{SV}}}{2}\left\|h-h^{*}\right\|_{L_{2}}^{2} .
$$

Combining the above inequality with (A.4.23), we have

$$
\mathcal{E}(\hat{h}) \leq \frac{\beta_{\mathrm{sv}}}{2} \max \left\{r_{\text {noise }}^{*}, r_{\mathrm{ver}}^{*}, r_{0}\right\}
$$

This completes the proof.

Lemma 25 (lower bound of the residual of the Taylor expansion) Let $\ell_{\mathrm{sv}}$ be convex with respect to its first argument. Given $v>0$, for all $u_{1}, u_{2} \in \mathbb{R}$ and $y \in \mathcal{Y}$, we have

$$
\begin{equation*}
\ell_{\mathrm{sv}}\left(u_{1}, y\right)-\ell_{\mathrm{sv}}\left(u_{2}, y\right)-\partial_{1} \ell_{\mathrm{sv}}\left(u_{2}, y\right)\left(u_{1}-u_{2}\right) \geq \frac{\alpha(2 v)}{2} \min \left\{\left|u_{1}-u_{2}\right|^{2}, v^{2}\right\} \cdot \mathbb{1}\left\{\left|u_{2}-y\right| \leq v\right\} \tag{A.4.24}
\end{equation*}
$$

Proof of Lemma 25: we consider the following four cases: (1) $\left|u_{2}-y\right|>v$; (2) $\left|u_{2}-y\right| \leq v$ and $\left|u_{1}-u_{2}\right| \leq v$; (3) $\left|u_{2}-y\right| \leq v$ and $u_{1}-u_{2}>v$; and (4) $\left|u_{2}-y\right| \leq v$ and $u_{1}-u_{2}<-v$. It is straightforward to prove (A.4.24) in case (1) and case (2). In case (3), because
$\ell_{\mathrm{SV}}\left(u_{1}, y\right)-\ell_{\mathrm{SV}}\left(u_{2}, y\right)-\partial_{1} \ell_{\mathrm{SV}}\left(u_{2}, y\right)\left(u_{1}-u_{2}\right)=\int_{0}^{1}\left(\partial_{1} \ell_{\mathrm{SV}}\left(u_{2}+t\left(u_{1}-u_{2}\right)\right)-\partial_{1} \ell_{\mathrm{SV}}\left(u_{2}\right)\right)\left(u_{1}-u_{2}\right) d t$,
and $\left(\partial_{1} \ell_{\mathrm{SV}}\left(u_{2}+t\left(u_{1}-u_{2}\right)\right)-\partial_{1} \ell_{\mathrm{SV}}\left(u_{2}\right)\right)\left(u_{1}-u_{2}\right) \geq 0$ for all $t \in[0,1]$, we have

$$
\begin{aligned}
\ell_{\mathrm{sv}}\left(u_{1}, y\right)-\ell_{\mathrm{sv}}\left(u_{2}, y\right)-\partial_{1} \ell_{\mathrm{sv}}\left(u_{2}, y\right)\left(u_{1}-u_{2}\right) \geq & \int_{0}^{\frac{v}{u_{1}-u_{2}}}\left(\partial_{1} \ell_{\mathrm{sv}}\left(u_{2}+t\left(u_{1}-u_{2}\right)\right)-\partial_{1} \ell_{\mathrm{sv}}\left(u_{2}\right)\right)\left(u_{1}-u_{2}\right) d t \\
& =\ell_{\mathrm{sv}}\left(u_{2}+v, y\right)-\ell_{\mathrm{sv}}\left(u_{2}, y\right)-\partial_{1} \ell_{\mathrm{sv}}\left(u_{2}, y\right) v \geq \frac{\alpha(2 v)}{2} v^{2}
\end{aligned}
$$

Similarly, we can prove (A.4.24) in case (4). This completes the proof of Lemma 25.

## A.4.2 Proof of Corollary 21

The proof is nearly identical to the proof of Theorem 20, but with the following modifications. First, we only need to consider the hypothesis set $\mathcal{H}_{0}$. Second, based on the definition of $\varphi_{\text {noise }}$ in Corollary 21, we modify (A.4.4) to
$\sup _{h \in \mathcal{H}_{0}, \frac{r}{\lambda} \leq T(h) \leq r}\left(\mathbb{P}-\mathbb{P}_{n}\right)\left(f_{h}+g_{h}\right) \leq \sup _{\frac{r}{\lambda} \leq T(h) \leq r}\left(\mathbb{P}-\mathbb{P}_{n}\right) g_{h}+\frac{2}{\alpha\left(4\|\xi\|_{L_{2}} / \sqrt{c_{\kappa}}\right)}\left(\varphi_{\text {noise }}\left(r ; \frac{\delta}{2}\right)-\Phi\left(h^{*}\right)\right)$.

We also do similar modifications to (A.4.10) (A.4.11) (A.4.12) (A.4.14). Third, we modify (A.4.15) (note that this is the only place we use the property of empirical risk minimization) as follows:

$$
\begin{align*}
& \mathbb{P}_{n}\left(f_{\hat{h}}+g_{\hat{h}}\right) \leq \frac{2}{\alpha\left(4\|\xi\|_{L_{2}} / \sqrt{c_{\kappa}}\right)} \mathbb{P}_{n}\left[\ell_{\mathrm{SV}}(\hat{h}(x)-y)-\ell_{\mathrm{SV}}\left(h^{*}(x)-y\right)\right] \\
& \leq \frac{2}{\alpha\left(4\|\xi\|_{L_{2}} / \sqrt{c_{\kappa}}\right)} \Phi\left(h^{*}\right), \tag{A.4.26}
\end{align*}
$$

where the first inequality is due to (A.4.1) and the second inequality is due to the definition (A.1.10) of the estimator $\hat{h}$. After all these modifications, the inequality (A.4.17) still hold true, and the remaining proof is identical to that of Theorem 20.

## Appendix B: Appendix for Chapter 2

## B. 1 Proofs for the finite-action setting

## B.1.1 Proof of Theorem 8.

We prove the theorem on the clean event stated in Lemma 5, whose measure is at least $1-\delta / 2$. For all $t>K$,

$$
\begin{array}{r}
\mathbb{E}_{x}\left[f^{*}\left(x, \pi_{f^{*}}(x)\right] \leq \mathbb{E}_{x}\left[\widehat{f_{t}}\left(x, \pi_{f^{*}}(x)\right)\right]+\mathbb{E}_{x}\left[\frac{\beta_{t}}{\sum_{i=1}^{t} \mathbb{1}\left\{\pi_{f^{*}}(x)=\pi_{i}(x)\right\}}\right]+\frac{K \beta_{t}}{t}\right. \\
\leq \underset{\pi \in \Pi}{\arg \max }\left\{\mathbb{E}_{x}\left[\widehat{f_{t}}(x, \pi(x))\right]+\mathbb{E}_{x}\left[\frac{\beta_{t}}{\sum_{i=1}^{t} \mathbb{1}\left\{\pi(x)=\pi_{i}(x)\right\}}\right]\right\}+\frac{K \beta_{t}}{t} \\
\\
=\mathbb{E}_{x}\left[\widehat{f_{t}}\left(x, \pi_{t}(x)\right)\right]+\mathbb{E}_{x}\left[\frac{\beta_{t}}{\sum_{i=1}^{t} \mathbb{1}\left\{\pi_{t}(x)=\pi_{i}(x)\right\}}\right]+\frac{K \beta_{t}}{t}  \tag{B.1.1}\\
\leq \\
\leq \mathbb{E}_{x}\left[f^{*}\left(x, \pi_{t}(x)\right)\right]+\mathbb{E}_{x}\left[\frac{2 \beta_{t}}{\sum_{i=1}^{t} \mathbb{1}\left\{\pi_{t}(x)=\pi_{i}(x)\right\}}\right]+\frac{2 K \beta_{t}}{t},
\end{array}
$$

where the first and the last inequality are due to Lemma 5; the second inequality due to maximization over policies.

Therefore, we have the following:

$$
\begin{array}{r}
\sum_{t=1}^{T} \mathbb{E}\left[f^{*}\left(x_{t}, \pi_{f^{*}}\left(x_{t}\right)\right)-f^{*}\left(x_{t}, a_{t}\right) \mid H_{t-1}\right]=\sum_{t=1}^{T}\left(\mathbb{E}_{x}\left[f^{*}\left(x, \pi_{f^{*}}(x)\right)\right]-\mathbb{E}_{x}\left[f^{*}\left(x, \pi_{t}(x)\right)\right]\right) \\
\leq \sum_{t=K+1}^{T} \mathbb{E}_{x}\left[\frac{2 \beta_{t}}{\sum_{i=1}^{t} \mathbb{1}\left\{\pi_{t}(x)=\pi_{i}(x)\right\}}\right]+\sum_{t=K+1}^{T} \frac{2 K \beta_{t}}{t}+K \\
\leq 2 \beta_{T} \sum_{t=K+1}^{T} \mathbb{E}_{x}\left[\frac{1}{\sum_{i=1}^{t-1} \mathbb{1}\left\{\pi_{t}(x)=\pi_{i}(x)\right\}}\right]+2 \sqrt{17 K T} \log \left(|\mathcal{F}| T^{3} / \delta\right) \\
\leq  \tag{B.1.2}\\
\leq 2 \sqrt{17 K T \log \left(2|\mathcal{F}| T^{3} / \delta\right)}(\log (T / K)+1)+K
\end{array}
$$

where the first line uses the equivalence proved in Lemma 6; the second line is due to (B.1.1); the third line is due to $\beta_{t} \leq \beta_{T}$ and $\sum_{K+1}^{T} 1 / \sqrt{t} \leq \sqrt{T}$; and the last line is due to the contextual potential lemma (Lemma 4).

By Azuma's inequality, with probability at least $1-\delta / 2$, we can bound the regret by

$$
\begin{equation*}
\operatorname{Regret}(T, \operatorname{Algorithm} 1) \leq \sum_{t=1}^{T} \mathbb{E}\left[f^{*}\left(x_{t}, \pi_{f^{*}}\left(x_{t}\right)\right)-f^{*}\left(x_{t}, a_{t}\right) \mid H_{t-1}\right]+\sqrt{2 T \log (2 / \delta)} \tag{B.1.3}
\end{equation*}
$$

Therefore, by a union bound and inequalities (B.1.2) (B.1.3), with probability at least $1-\delta$, the regret of Algorithm 1 after $T$ rounds is upper bounded by
$\operatorname{Regret}(T, \operatorname{Algorithm} 1) \leq 2 \sqrt{17 K T \log \left(2|\mathcal{F}| T^{3} / \delta\right)}(\log (T / K)+1)+\sqrt{2 T \log (2 / \delta)}+K$.

## B.1.2 Analysis on the confidence

The main goal of this subsection is to prove Lemma 3. For a fixed $f$, we denote $Y_{f, i}=$ $\left(f\left(x_{i}, a_{i}\right)-r_{i}\left(x_{i}, a_{i}\right)\right)^{2}-\left(f^{*}\left(x_{i}, a_{i}\right)-r_{i}\left(x_{i}, a_{i}\right)\right)^{2}, i=1,2, \ldots$.

## Proof of Lemma 3.

For a fixed $f \in \mathcal{F}$, when conditioned on $\Upsilon_{i-1}$, we have

$$
\begin{array}{r}
\mathbb{E}_{x_{i}, a_{i}}\left[\left(f\left(x_{i}, a_{i}\right)-f^{*}\left(x_{i}, a_{i}\right)\right)^{2} \mid H_{i-1}\right]=\mathbb{E}_{x_{i}}\left[\left(f\left(x_{i}, \pi_{i}\left(x_{i}\right)\right)-f^{*}\left(x_{i}, \pi_{i}\left(x_{i}\right)\right)\right)^{2} \mid H_{i-1}\right] \\
=\mathbb{E}_{x}\left[\left(f\left(x, \pi_{i}(x)\right)-f^{*}\left(x, \pi_{i}(x)\right)\right)^{2} \mid H_{i-1}\right] \\
=\mathbb{E}_{x}\left[\left(f\left(x, \pi_{i}(x)\right)-f^{*}\left(x, \pi_{i}(x)\right)\right)^{2}\right],
\end{array}
$$

where the first equation is because $a_{i}=\pi_{i}\left(x_{i}\right)$ and the fact that $\pi_{i}$ is completely determined by $H_{t-1}$; the second equation is because the independence between $x_{i}$ and $H_{i-1}$; and the third inequality is because $\left(f\left(x_{i}, \pi_{i}(x)\right)-f^{*}\left(x_{i}, \pi_{i}\left(x_{i}\right)\right)\right)^{2}$ depends on $H_{i-1}$ only through $\pi_{i}$.

Therefore,

$$
\sum_{i=1}^{t-1} \mathbb{E}_{x_{i}, a_{i}}\left[\left(f\left(x_{i}, a_{i}\right)-f^{*}\left(x_{i}, a_{i}\right)\right)^{2} \mid H_{i-1}\right]=\sum_{i=1}^{t-1} \mathbb{E}_{x}\left[\left(f\left(x, \pi_{i}(x)\right)-f^{*}\left(x_{i}, \pi_{i}(x)\right)\right)^{2}\right]
$$

Applying Lemma 7 , we know that $\forall \delta \in(0,1)$, with probability at least $1-\delta / 2$,

$$
\begin{equation*}
\sum_{i=1}^{t-1} \mathbb{E}_{x}\left[\left(f_{t}\left(x, \pi_{i}(x)\right)-f^{*}\left(x_{i}, \pi_{i}(x)\right)\right)^{2}\right] \leq 68 \log \left(2|\mathcal{F}| t^{3} / \delta\right)+2 \sum_{i=1}^{t-1} Y_{f_{t}, i} \tag{B.1.4}
\end{equation*}
$$

uniformly over all $t \geq K$ and all fixed sequence $f_{K}, f_{K+1}, \cdots \in \mathcal{F}$.
Therefore, $\forall \pi \in \Pi$,

$$
\begin{array}{r}
\mathbb{E}_{x}\left[\sum_{i=1}^{t-1} \mathbb{1}\left\{\pi(x)=\pi_{i}(x)\right\}\left(f_{t}(x, \pi(x))-f^{*}(x, \pi(x))\right)^{2}\right] \\
=\mathbb{E}_{x}\left[\sum_{i=1}^{t-1} \mathbb{1}\left\{\pi(x)=\pi_{i}(x)\right\}\left(f_{t}\left(x, \pi_{i}(x)\right)-f^{*}\left(x, \pi_{i}(x)\right)\right)^{2}\right] \\
=\sum_{i=1}^{t-1} \mathbb{E}_{x}\left[\mathbb{1}\left\{\pi(x)=\pi_{i}(x)\right\}\left(f_{t}\left(x, \pi_{i}(x)\right)-f^{*}\left(x, \pi_{i}(x)\right)\right)^{2}\right] \\
\leq \sum_{i=1}^{t-1} \mathbb{E}_{x}\left[\left(f_{t}\left(x, \pi_{i}(x)\right)-f^{*}\left(x, \pi_{i}(x)\right)\right)^{2}\right] \\
\leq 68 \log \left(2|\mathcal{F}| t^{3} / \delta\right)+2 \sum_{i=1}^{t-1} Y_{f_{t}, i}, \tag{B.1.5}
\end{array}
$$

where the first inequalities are due to $\mathbb{1}\left\{\pi(x)=\pi_{i}(x)\right\} \leq 1$ and the second inequality is (B.1.4).
Since Algorithm 1 pick all actions exactly once during the first $K$ rounds, $t>K$ will ensure $\sum_{i=1}^{t-1} \mathbb{1}\{\pi(x)=\pi(x)\} \geq 1, \forall x \in \mathcal{X}$.

From Cauchy-Schwarz's inequality, $\forall t>K, \forall \pi \in \Pi$,

$$
\left|\mathbb{E}_{x}\left[f_{t}(x, \pi(x))-f^{*}(x, \pi(x))\right]\right|
$$

$$
\leq \sqrt{\mathbb{E}_{x}\left[\frac{1}{\sum_{i=1}^{t-1} \mathbb{1}\left\{\pi(x)=\pi_{i}(x)\right\}}\right]} \sqrt{\mathbb{E}_{x}\left[\sum_{i=1}^{t-1} \mathbb{1}\left\{\pi(x)=\pi_{i}(x)\right\}\left(f_{t}(x, \pi(x))-f^{*}\left(x_{i}, \pi(x)\right)\right)^{2}\right]} .
$$

Combine the above inequality with (B.1.5), we prove

$$
\leq \sqrt{\mathbb{E}_{x}\left[\frac{1}{\sum_{i=1}^{t-1} \mathbb{1}\left\{\pi(x)=\pi_{i}(x)\right\}}\right]} \sqrt{68 \log \left(2|\mathcal{F}| t^{3} / \delta\right)+2 \sum_{i=1}^{t-1} Y_{f_{t}, i}} .
$$

Taking $f_{t}=\widehat{f_{t}}$ in the above inequality, and use the fact $\sum_{i=1}^{t-1} Y_{\widehat{f_{t}}, i} \leq 0$ (as the least square solution $\widehat{f_{t}}$ minimizes $\left.\sum_{i=1}^{t-1}\left(f\left(x_{i}, a_{i}\right)-r_{i}\left(x_{i}, a_{i}\right)\right)^{2}\right)$, we obtain: with probability at least $1-\delta / 2, \forall t>K$, $\forall \pi \in \Pi$,

$$
\begin{array}{r}
\mid \mathbb{E}_{x}\left[\widehat{f f}_{t}(x, \pi(x))\right] \\
\leq \sqrt{\mathbb{E}_{x}\left[f^{*}(x, \pi(x))\right] \mid} \\
\left.\leq \sqrt{\mathbb{E}_{x}\left[\frac{1}{\sum_{i=1}^{t-1} \mathbb{1}\left\{\pi(x)=\pi_{i}(x)\right\}}\right.}\right] \sqrt{68 \log \left(2|\mathcal{F}| t^{3} / \delta\right)} \\
\mathbb{E}_{i=1}^{t-1} \mathbb{1}\left\{\pi(x)=\pi_{i}(x)\right\}
\end{array} \sqrt{68 \log \left(2|\mathcal{F}| t^{3} / \delta\right)}
$$

## Proof of Lemma 7

We now prove Lemma 7 and the supporting lemmas required to prove Lemma 7.

Proof of Lemma 7. Fix a $\delta \in(0,1)$. Take $\delta_{t}=\delta / 2 t^{3}$, and apply a union bound to Lemma 26 with all $t \geq 2$. From

$$
\sum_{t=1}^{\infty} \delta_{t} \log _{2}(t-1) \leq \sum_{t=2}^{\infty} \delta / 2 t^{2} \leq \delta / 2
$$

we know that with probability at least $1-\delta / 2$,

$$
\sum_{i=1}^{t-1} \mathbb{E}_{x_{i}, a_{i}}\left[\left(f_{t}\left(x_{i}, a_{i}\right)-f^{*}\left(x_{i}, a_{i}\right)\right)^{2} \mid H_{i-1}\right] \leq 68 \log \left(2|\mathcal{F}| t^{3} / \delta\right)+2 \sum_{i=1}^{t-1} Y_{f_{t}, i},
$$

uniformly over all $t \geq 2$ and all fixed sequence $f_{2}, f_{3}, \cdots \in \mathcal{F}$.

Lemma 26 (uniform convergence over $\mathcal{F}$ ) For a fixed $t \geq 2$ and a fixed $\delta_{t} \in\left(0,1 / e^{2}\right)$, with probability at least $1-\log _{2}(t-1) \delta_{t}$, we have

$$
\begin{equation*}
\sum_{i=1}^{t-1} \mathbb{E}_{x_{i}, a_{i}}\left[\left(f\left(x_{i}, a_{i}\right)-f^{*}\left(x_{i}, a_{i}\right)\right)^{2} \mid H_{i-1}\right] \leq 68 \log \left(|\mathcal{F}| / \delta_{t}\right)+2 \sum_{i=1}^{t-1} Y_{f, i} \tag{B.1.6}
\end{equation*}
$$

uniformly over all $f \in \mathcal{F}$.

Proof of Lemma 26. We have $\left|Y_{f, i}\right| \leq 1, \forall i$. From Lemma 27, for $\delta_{t} /|\mathcal{F}| \leq \delta_{t}<1 / e^{2}$, with probability at least $1-\log _{2}(t-1) \delta_{t} /|\mathcal{F}|$,

$$
\sum_{i=1}^{t-1} \mathbb{E}\left[Y_{f, i} \mid H_{i-1}\right]-\sum_{i=1}^{t-1} Y_{f, i} \leq 4 \sqrt{\sum_{i=1}^{t-1} \operatorname{Var}\left[Y_{f, i} \mid H_{i-1}\right] \log \left(|\mathcal{F}| / \delta_{t}\right)}+2 \log \left(|\mathcal{F}| / \delta_{t}\right)
$$

Applying union bound to all $f \in \mathcal{F}$, we obtain that with probability at least $1-\log _{2}(t-1) \delta_{t} \geq$ $1-\log _{2} t \delta_{t}$,

$$
\sum_{i=1}^{t-1} \mathbb{E}\left[Y_{f, i} \mid H_{i-1}\right]-\sum_{i=1}^{t-1} Y_{f, i} \leq 4 \sqrt{\sum_{i=1}^{t-1} \operatorname{Var}\left[Y_{f, i} \mid H_{i-1}\right] \log \left(|\mathcal{F}| / \delta_{t}\right)}+2 \log \left(|\mathcal{F}| / \delta_{t}\right), \quad \forall f \in \mathcal{F}
$$

From Lemma 28 we have $\operatorname{Var}\left[Y_{f, i} \mid H_{i}\right] \leq 4 \mathbb{E}\left[Y_{f, i} \mid H_{i}\right]$. Therefore

$$
\begin{aligned}
& \sum_{i=1}^{t-1} \mathbb{E}\left[Y_{f, i} \mid H_{i-1}\right] \leq 4 \sqrt{\sum_{i=1}^{t-1} \operatorname{Var}\left[Y_{f, i} \mid H_{i-1}\right]} \log \left(|\mathcal{F}| / \delta_{t}\right)
\end{aligned}+2 \log \left(|\mathcal{F}| / \delta_{t}\right)+\sum_{i=1}^{t-1} Y_{f, i} .
$$

This implies $\forall f \in \mathcal{F}$,

$$
\left(\sqrt{\sum_{i=1}^{t-1} \mathbb{E}\left[Y_{f, i} \mid H_{i-1}\right]}-4 \sqrt{\log \left(|\mathcal{F}| / \delta_{t}\right)}\right)^{2} \leq 18 \log \left(|\mathcal{F}| / \delta_{t}\right)+\sum_{i=1}^{t-1} Y_{f, i}
$$

which further implies $\forall f \in \mathcal{F}$,

$$
\sum_{i=1}^{t-1} \mathbb{E}\left[Y_{f, i} \mid H_{i-1}\right] \leq 68 \log \left(|\mathcal{F}| / \delta_{t}\right)+2 \sum_{i=1}^{t-1} Y_{f, i}
$$

From Lemma 28, we have

$$
\left.\sum_{i=1}^{t-1} \mathbb{E}_{x_{i}, a_{i}}\left[f\left(x_{i}, a_{i}\right)-f^{*}\left(x_{i}, a_{i}\right)\right)^{2} \mid H_{i-1}\right]=\sum_{i=1}^{t-1} \mathbb{E}\left[Y_{f, i} \mid H_{i-1}\right] \leq 68 \log \left(|\mathcal{F}| / \delta_{t}\right)+2 \sum_{i=1}^{t-1} Y_{f, i}
$$

This finish the proof to Lemma 26.

The following two lemmas are used in the proof of Lemma 26.

Lemma 27 (Freeman's inequality, [143]) Suppose $Z_{1}, Z_{2}, \ldots, Z_{t}$ is a martingale difference sequence with $\left|Z_{i}\right| \leq b$ for all $i=1, \ldots, t$. Then for any $\delta<1 / e^{2}$, with probability at least $1-\left(\log _{2} t\right) \delta$,

$$
\sum_{i=1}^{t} Z_{i} \leq 4 \sqrt{\sum_{i=1}^{t} \operatorname{Var}\left[Z_{i} \mid Z_{1}, \ldots, Z_{i-1}\right] \log (1 / \delta)}+2 b \log (1 / \delta)
$$

Lemma 28 (Lemma 4.2 in [76]) Fix a function $f \in \mathcal{F}$. Suppose we sample $x$ from the data distribution $\mathcal{D}_{\mathcal{X}}$, and $r(x, a)$ from $\mathcal{D}_{x, a}$. Define the random variable

$$
Y=(f(x, a)-r(x, a))^{2}-\left(f^{*}(x, a)-r(x, a)\right)^{2}
$$

Then we have

$$
\begin{gathered}
\mathbb{E}_{x, r, a}[Y]=\mathbb{E}_{x, a}\left[\left(f(x, a)-f^{*}(x, a)\right)^{2}\right], \\
\operatorname{Var}_{x, r, a}[Y] \leq 4 \mathbb{E}_{x, r, a}[Y] .
\end{gathered}
$$

## B.1.3 Proof of Lemma 4

Proof of Lemma 4. For any fixed $x \in \mathcal{X}$, we have

$$
\begin{aligned}
& \sum_{t=K+1}^{T} \frac{1}{\sum_{i=1}^{t-1} \mathbb{1}\left\{\pi_{t}(x)=\pi_{j}(x)\right\}} \leq \sum_{a \in \mathcal{A}} \sum_{i=1}^{\sum_{t=1}^{T} \mathbb{1}\left\{\pi_{t}(x)=a\right\}} \frac{1}{i} \\
\leq & \sum_{a \in \mathcal{A}}\left(1+\log \left(\sum_{t=1}^{T} \mathbb{1}\left\{\pi_{t}(x)=a\right\}\right)\right) \leq K+K \log (T / K),
\end{aligned}
$$

where the last inequality is due to Jensen's inequality. By taking expectation on both sides of the above inequality, we prove the lemma.

## B. 2 Proofs for the extensions to infinite function classes

## B.2.1 Proof of Corollary 9

From the well-known result on the covering of $d$-dimensional balls [144], the covering number of a $d$-dimensional ball with radius $\frac{\Delta}{2}$ and discretization error $\frac{1}{L t}$ is bounded by $(1+\Delta L t)^{d}$, so there exists a set $V_{t}$ of size no more than $(1+\Delta L t)^{d}+1 \leq(2+\Delta L t)^{d}$ that contains $\theta^{*}$ and satisfies

$$
\forall \theta \in \Theta \exists v \in V_{t} \text { s.t. }\|\theta-v\| \leq \frac{1}{L t}
$$

We see $\log \left|V_{t}\right| \leq d \log (2+\Delta L t) . \forall f_{\theta} \in \mathcal{F}, x \in \mathcal{X}, a \in \mathcal{A}$, take $v$ to be the closest point to $\theta$ in $V_{t}$, we have

$$
\begin{array}{r}
\left(f_{\theta}(x, a)-f^{*}(x, a)\right)^{2}=\left(f_{\theta}(x, a)-f_{v}(x, a)+f_{v}(x, a)-f_{\theta^{*}}(x, a)\right)^{2} \\
\leq 2\left(f_{\theta}(x, a)-f_{v}(x, a)\right)^{2}+2\left(f_{v}(x, a)-f_{\theta^{*}}(x, a)\right)^{2} \\
\leq 2 L^{2}\|\theta-v\|^{2}+2\left(f_{v}(x, a)-f_{\theta^{*}}(x, a)\right)^{2} \\
\leq \frac{2}{t^{2}}+2\left(f_{v}(x, a)-f_{\theta^{*}}(x, a)\right)^{2} .
\end{array}
$$

Sine $V_{t}$ is a finite function class, we can prove a slight modification of Lemma 26, with the result (B.1.6) becomes

$$
\sum_{i=1}^{t-1} \mathbb{E}_{x_{i}, a_{i}}\left[\left(f\left(x_{i}, a_{i}\right)-f^{*}\left(x_{i}, a_{i}\right)\right)^{2} \mid H_{i-1}\right] \leq 136 \log \left(\left|V_{t}\right| / \delta_{t}\right)+2+4 \sum_{i=1}^{t-1} Y_{f, i}
$$

Following the same path in the proof of Lemma 7, we can prove a slight modification of Lemma 7: with probability at least $1-\frac{\delta}{2}$,

$$
\begin{aligned}
\sum_{i=1}^{t-1} \mathbb{E}_{x_{i}, a_{i}}\left[\left(f_{t}\left(x_{i}, a_{i}\right)-f^{*}\left(x_{i}, a_{i}\right)\right)^{2} \mid H_{i-1}\right] & \leq 136\left(d \log (2+\Delta L t)+\log \frac{2 t^{3}}{\delta}\right)+2+4 \sum_{i=1}^{t-1} Y_{f_{t}, i} \\
& \leq 136\left(d \log (2+\Delta L t)+\log \frac{2 t^{3}}{\delta}+1\right)+4 \sum_{i=1}^{t-1} Y_{f_{t}, i}
\end{aligned}
$$

uniformly over all $t \geq 2$ and all fixed sequence $f_{2}, f_{3}, \cdots \in \mathcal{F}$.
By setting the parameter $\beta_{t}$ to be

$$
\beta_{t}=\sqrt{34 t / K} \sqrt{d \log (2+\Delta L t)+\log \left(2 t^{3} / \delta\right)+1}
$$

in Algorithm 1, we can prove a slight modification of Theorem 8, with the result being

$$
\operatorname{Regret}(T, \operatorname{Algorithm} 1) \leq 2 K \beta_{T}(\log (T / K)+1)+\sqrt{2 T \log (2 / \delta)}+K
$$

## B.2.2 Proof of Corollary 10

We introduce the following Lemma adopt from [73]:

Lemma 29 (a consequence of Lemma 4 in [73]) $\forall \delta \in(0,1)$, with probability at least $1-\delta$,

$$
\begin{array}{r}
\sum_{i=\tau_{1}}^{\tau_{2}} \mathbb{E}_{x_{i}, a_{i}}\left[\left(f\left(x_{i}, a_{i}\right)-f^{*}\left(x_{i}, a_{i}\right)\right)^{2} \mid H_{i-1}\right] \leq 2 \sum_{i=\tau_{1}}^{\tau_{2}} Y_{f, i}+ \\
K \cdot \inf _{\varepsilon>0}\left\{100 \varepsilon T+320 \log \left(\frac{4 K \mathbb{E}_{\left\{x_{i}\right\}_{i=1}^{T}} \mathcal{N}_{1}\left(\mathcal{G}, \varepsilon,\left\{x_{i}\right\}_{i=1}^{T}\right)}{\delta}\right)\right\} .
\end{array}
$$

for all $1 \leq \tau_{1} \leq \tau_{2} \leq T$ and $g \in \mathcal{G}$.

We then prove a slight modification of Lemma 7, with the result (2.2.7) becomes

$$
\begin{array}{r}
\sum_{i=1}^{t-1} \mathbb{E}_{x_{i}, a_{i}}\left[\left(f_{t}\left(x_{i}, a_{i}\right)-f^{*}\left(x_{i}, a_{i}\right)\right)^{2} \mid H_{i-1}\right] \leq 2 \sum_{i=1}^{t-1} Y_{f_{t}, i}+ \\
K \cdot \inf _{\varepsilon>0}\left\{100 \varepsilon T+320 \log \left(\frac{8 K T^{3} \mathbb{E}_{\left\{x_{i}\right\}_{i=1}^{T}} \mathcal{N}_{1}\left(\mathcal{G}, \varepsilon,\left\{x_{i}\right\}_{i=1}^{T}\right)}{\delta}\right)\right\},
\end{array}
$$

uniformly over all $t \geq 2$ and all fixed sequence $f_{2}, f_{3}, \cdots \in \mathcal{F}$.
By setting the parameter $\beta_{t}$ in Algorithm 1 to be the fixed value

$$
\beta=\sqrt{T K} \cdot \inf _{\varepsilon>0}\left\{25 \varepsilon T+80 \log \left(\frac{8 K T^{3} \mathbb{E}_{\left\{x_{i}\right\}_{i=1}^{T}} \mathcal{N}_{1}\left(\mathcal{G}, \varepsilon,\left\{x_{i}\right\}_{i=1}^{T}\right)}{\delta}\right)\right\},
$$

we can prove a slight modification of Theorem 8 , with the result being
$\operatorname{Regret}(T, \operatorname{Algorithm} 1) \leq 2 K \beta(\log (T / K)+1)+\sqrt{2 T \log (2 / \delta)}+K$.

## B. 3 Proofs for the infinite-action setting

Proof of Theorem 11. We prove the theorem on the clean event stated in Lemma 30, whose measure is at least $1-\delta / 2$. For all $t \geq 2$,

$$
\begin{array}{r}
\mathbb{E}_{x}\left[f^{*}\left(x, \pi_{f^{*}}(x)\right] \leq \mathbb{E}_{x}\left[1 \wedge \beta_{t} V_{x}\left(\pi_{f^{*}}(x) \|\left\{\pi_{i}(x)\right\}_{i=1}^{t-1}\right)\right]+\mathcal{E} \beta_{t} / t\right. \\
\leq \mathbb{E}_{x}\left[\widehat{f}_{t}\left(x, \pi_{t}(x)\right)\right]+\mathbb{E}_{x}\left[1 \wedge \beta_{t} V_{x}\left(\pi_{t}(x) \|\left\{\pi_{i}(x)\right\}_{i=1}^{t-1}\right)\right]+\mathcal{E} \beta_{t} / t \\
\leq \mathbb{E}_{x}\left[f^{*}\left(x, \pi_{t}(x)\right)\right]+2 \mathbb{E}_{x}\left[1 \wedge \beta_{t} V_{x}\left(\pi_{t}(x) \|\left\{\pi_{i}(x)\right\}_{i=1}^{t-1}\right)\right]+2 \mathcal{E} \beta_{t} / t
\end{array}
$$

where the first and the last inequalities are due to Lemma 30; the second inequality is due to the definition of $\pi_{t}$ in Lemma 31. The above argument implies that for all $t \geq 2$

$$
\begin{equation*}
\mathbb{E}_{x}\left[f ^ { * } \left(x, \pi_{f^{*}}(x)-f^{*}\left(x, \pi_{t}(x)\right] \leq 2 \mathbb{E}_{x}\left[1 \wedge \beta_{t} V_{x}\left(\pi_{t}(x) \|\left\{\pi_{i}(x)\right\}_{i=1}^{t-1}\right)\right]+2 \mathcal{E} \beta_{t} / t\right.\right. \tag{B.3.1}
\end{equation*}
$$

When $t=1$, inequality (B.3.1) trivially holds true, because $V_{x}\left(\pi_{1}(x) \| \emptyset\right)=\infty$ by definition. So inequality (B.3.1) holds true for all $t \geq 1$.

When $T \leq \mathcal{E}$, we can bound the regret by $\mathcal{E}$. We now give the regret bound for the case $T>\mathcal{E}$.

We have the following:

$$
\begin{array}{r}
\sum_{t=1}^{T} \mathbb{E}\left[f^{*}\left(x_{t}, \pi_{f^{*}}\left(x_{t}\right)\right)-f^{*}\left(x_{t}, a_{t}\right) \mid H_{t-1}\right]=\sum_{t=1}^{T}\left(\mathbb{E}_{x}\left[f^{*}\left(x, \pi_{f^{*}}(x)\right)\right]-\mathbb{E}_{x}\left[f^{*}\left(x, \pi_{t}(x)\right)\right]\right) \\
\leq \sum_{t=1}^{T} 2 \mathbb{E}_{x}\left[1 \wedge \beta_{t} V_{x}\left(\pi_{t}(x) \|\left\{\pi_{i}(x)\right\}_{i=1}^{t-1}\right)\right]+\sum_{t=1}^{T} 2 \mathcal{E} \beta_{t} / t \\
\leq 2 \sum_{t=1}^{T} \mathbb{E}_{x}\left[1 \wedge \beta_{t} V_{x}\left(\pi_{t}(x) \|\left\{\pi_{i}(x)\right\}_{i=1}^{t-1}\right)\right]+2 \sqrt{17 \mathcal{E} T \log \left(2|\mathcal{F}| T^{3} / \delta\right)} \\
\leq 2 \sum_{t=1}^{T} \mathbb{E}_{x}\left[\beta_{T} \wedge \beta_{T} V_{x}\left(\pi_{t}(x) \|\left\{\pi_{i}(x)\right\}_{i=1}^{t-1}\right)\right]+2 \sqrt{17 \mathcal{E} T \log \left(2|\mathcal{F}| T^{3} / \delta\right)} \\
=2 \beta_{T} \mathbb{E}_{x}\left[\sum_{t=1}^{T} 1 \wedge V_{x}\left(\pi_{t}(x) \|\left\{\pi_{i}(x)\right\}_{i=1}^{t-1}\right)\right]+2 \sqrt{17 \mathcal{E} T \log \left(2|\mathcal{F}| T^{3} / \delta\right)} \\
\leq 2 \beta_{T} \mathcal{E} \operatorname{poly}(\log T)+2 \sqrt{17 \mathcal{E} T \log \left(2|\mathcal{F}| T^{3} / \delta\right)} \\
=2 \sqrt{17 \mathcal{E} T \log \left(2|\mathcal{F}| T^{3} / \delta\right)}(\operatorname{poly}(\log T)+1), \tag{B.3.2}
\end{array}
$$

where the first line uses the equivalence proved in Lemma 31; the second line is due to (B.3.1); the third line is due to $\sum_{t=1}^{T} 1 / \sqrt{t} \leq \sqrt{T}$; the fourth line is due to $\beta_{T}>\beta_{t}$ and $\beta_{T}>1$ when $T>\mathcal{E}$; and the sixth line is due to the condition II in Assumption 8. By Azuma's inequality, with probability at least $1-\delta / 2$, we can bound the regret by

$$
\begin{equation*}
\operatorname{Regret}(T, \operatorname{Algorithm} 2) \leq \sum_{t=1}^{T} \mathbb{E}\left[f^{*}\left(x_{t}, \pi_{f^{*}}\left(x_{t}\right)\right)-f^{*}\left(x_{t}, a_{t}\right) \mid H_{t-1}\right]+\sqrt{2 T \log (2 / \delta)} \tag{B.3.3}
\end{equation*}
$$

Therefore, by a union bound and inequalities (B.3.2) (B.3.3), with probability at least $1-\delta$, the regret of Algorithm 1 after $T$ rounds is upper bounded by

$$
\operatorname{Regret}(T, \operatorname{Algorithm} 2) \leq 2 \sqrt{17 \mathcal{E} T \log \left(2|\mathcal{F}| T^{3} / \delta\right)}(\operatorname{poly}(\log T)+1)+\sqrt{2 T \log (2 / \delta)}
$$

Combine the case $T \leq \mathcal{E}$ and $T>\mathcal{E}$ we finish the proof.

Lemma 30 (counterfactual confidence bound) Consider a non-randomized contextual bandit algorithm that selects $\pi_{t}$ based on $H_{t-1}$ and chooses the action $a_{t}=\pi_{t}\left(x_{t}\right)$ at all rounds $t$. Then $\forall \delta \in(0,1)$, with probability at least $1-\delta / 2$, we have

$$
\mid \mathbb{E}_{x}\left[\widehat{f}_{t}(x, \pi(x)]-\mathbb{E}_{x}\left[f^{*}(x, \pi(x))\right] \mid \leq \mathbb{E}\left[1 \wedge \beta_{t} V_{x}\left(\pi(x) \|\left\{\pi_{i}(x)\right\}_{i=1}^{t-1}\right)\right]+\mathcal{E} \beta_{t} / t .\right.
$$

uniformly over all $\pi \in \Pi$ and all $t \geq 2$.

Proof of Lemma 30. For a fixed $f$, we denote $Y_{f, i}=\left(f\left(x_{i}, a_{i}\right)-r_{i}\left(x_{i}, a_{i}\right)\right)^{2}-\left(f^{*}\left(x_{i}, a_{i}\right)-\right.$ $\left.r_{i}\left(x_{i}, a_{i}\right)\right)^{2}, i=1,2, \ldots$ From Lemma $7, \forall \delta \in(0,1)$, with probability at least $1-\delta / 2$, we have

$$
\begin{equation*}
\sum_{i=1}^{t-1} \mathbb{E}_{x_{i}, a_{i}}\left[\left(f_{t}\left(x_{i}, a_{i}\right)-f^{*}\left(x_{i}, a_{i}\right)\right)^{2} \mid H_{i-1}\right] \leq 68 \log \left(2|\mathcal{F}| t^{3} / \delta\right)+\sum_{i=1}^{t-1} Y_{f_{t}, i} \tag{B.3.4}
\end{equation*}
$$

uniformly over all $t \geq 2$ and all fixed sequence $f_{2}, f_{3}, \ldots f_{t+1}, \cdots \in \mathcal{F}$.
Use the fact that $\pi_{i}$ is completely determined by $H_{i-1}$ and independent with $x_{i}$, we obtain:

$$
\begin{equation*}
\sum_{i=1}^{t-1} \mathbb{E}_{x}\left[\left(\widehat{f_{t}}\left(x, \pi_{i}(x)\right)-f^{*}\left(x, \pi_{i}(x)\right)\right)^{2}\right] \leq 68 \log \left(2|\mathcal{F}| t^{3} / \delta\right)+\sum_{i=1}^{t-1} Y_{f_{t}, i}=4 \mathcal{E} \beta_{t}^{2} / t+\sum_{i=1}^{t-1} Y_{f_{t}, i} \tag{B.3.5}
\end{equation*}
$$

uniformly over all $t \geq 2$.
From the definition of counterfactual action divergence, we know $\forall x \in \mathcal{X}$,

$$
\left|f_{t}(x, \pi(x))-f^{*}(x, \pi(x))\right| \leq \sqrt{V_{x}\left(\pi(x) \|\left\{\pi_{i}(x)\right\}_{i=1}^{t-1}\right)} \sqrt{\sum_{i=1}^{t-1}\left(f_{t}\left(x, \pi_{i}(x)\right)-f^{*}\left(x, \pi_{i}(x)\right)\right)^{2}} .
$$

Applying the AM-GM inequality to the above inequality, we obtain

$$
\left|f_{t}(x, \pi(x))-f^{*}(x, \pi(x))\right| \leq \beta_{t} V_{x}\left(\pi(x) \|\left\{\pi_{i}(x)\right\}_{i=1}^{t-1}\right)+\frac{1}{4 \beta_{t}} \sum_{i=1}^{t-1}\left(f_{t}\left(x, \pi_{i}(x)\right)-f^{*}\left(x, \pi_{i}(x)\right)\right)^{2}
$$

Since $\mathcal{F}$ is bounded by $[0,1]$, we further obtain

$$
\begin{align*}
&\left|f_{t}(x, \pi(x))-f^{*}(x, \pi(x))\right| \leq \max \left\{\beta_{t} V_{x}\left(\pi(x)| |\left\{\pi_{i}(x)\right\}_{i=1}^{t-1}\right)+\frac{1}{4 \beta_{t}} \sum_{i=1}^{t-1}\left(f_{t}\left(x, \pi_{i}(x)\right)-f^{*}\left(x, \pi_{i}(x)\right)\right)^{2}, 1\right\} \\
& \leq \max \left\{1, \beta_{t} V_{x}\left(\pi(x) \|\left\{\pi_{i}(x)\right\}_{i=1}^{t-1}\right)\right\}+\frac{1}{4 \beta_{t}} \sum_{i=1}^{t-1}\left(f_{t}\left(x, \pi_{i}(x)\right)-f^{*}\left(x, \pi_{i}(x)\right)\right)^{2} \\
&=1 \wedge \beta_{t} V_{x}\left(\pi(x) \|\left\{\pi_{i}(x)\right\}_{i=1}^{t-1}\right)+\frac{1}{4 \beta_{t}} \sum_{i=1}^{t-1}\left(\widehat{f_{t}}\left(x, \pi_{i}(x)\right)-f^{*}\left(x, \pi_{i}(x)\right)\right)^{2} \tag{B.3.6}
\end{align*}
$$

By taking expectation on both side of (B.3.6) and using (B.3.5), we obtain that with probability at least $1-\delta / 2$,

$$
\begin{array}{r}
\left|\mathbb{E}_{x}\left[f_{t}(x, \pi(x))-f^{*}(x, \pi(x))\right]\right| \leq \mathbb{E}_{x}\left|f_{t}(x, \pi(x))-f^{*}(x, \pi(x))\right| \\
\leq \mathbb{E}_{x}\left[1 \wedge \beta_{t} V_{x}\left(\pi(x)| |\left\{\pi_{i}(x)\right\}_{i=1}^{t-1}\right)\right]+\frac{1}{4 \beta_{t}} \mathbb{E}_{x}\left[\sum_{i=1}^{t-1}\left(f_{t}\left(x, \pi_{i}(x)\right)-f^{*}\left(x, \pi_{i}(x)\right)\right)^{2}\right] \\
=\mathbb{E}_{x}\left[1 \wedge \beta_{t} V_{x}\left(\pi(x)| |\left\{\pi_{i}(x)\right\}_{i=1}^{t-1}\right)\right]+\frac{1}{4 \beta_{t}} \sum_{i=1}^{t-1} \mathbb{E}\left[\left(f_{t}\left(x, \pi_{i}(x)\right)-f^{*}\left(x, \pi_{i}(x)\right)\right)^{2}\right] \\
\leq \mathbb{E}_{x}\left[1 \wedge \beta_{t} V_{x}\left(\pi(x)| |\left\{\pi_{i}(x)\right\}_{i=1}^{t-1}\right)\right]+\mathcal{E} \beta_{t} / t+\frac{1}{4 \beta_{t}} \sum_{i=1}^{t-1} Y_{f_{t}, i} .
\end{array}
$$

uniformly over all $\pi \in \Pi$, all $t \geq 2$ and all fixed sequence $f_{2}, f_{3}, \ldots, \in \mathcal{F}$. Here the first inequality is due to the triangle inequality; the second inequality is due to (B.3.6); the last inequality is due to (B.3.5).

By taking $f_{t}=\widehat{f}_{t}$ the the least square solution that minimizes $\sum_{i=1}^{t-1}\left(f\left(x_{i}, a_{i}\right)-r_{i}\left(x_{i}, a_{i}\right)\right)^{2}$, we have $Y_{f_{t}, i} \leq 0$ and finish the proof.

Lemma 31 Consider an algorithm that choose policy $\pi_{t}$ by

$$
\pi_{t}(x)= \begin{cases}A_{x, t} & \text { if } t \leq d_{x} \\ \underset{\mathcal{A}(x)}{\arg \max }\left\{\widehat{f}_{t}(x, \cdot)+\beta_{t} V_{x_{t}}\left(\cdot \|\{\pi(x)\}_{j=1}^{t-1}\right)\right\} & \text { if } t>d_{x}\end{cases}
$$

( $A_{x, t}$ is determined the initialization oracle and the input $\mathcal{A}(x)$; the "argmax" problem when $t>d_{x}$ is computed via the action maximization oracle.) Then this algorithm produces the same actions as those produced by Algorithm 2.

Proof. The proof to this lemma is straightforward.

## B. 4 Proofs for the "optimistic subroutine" in Section 2.5

In this subsection we prove Proposition 3. Our proof is motivated by Lemma 6 and Lemma 7 in [77].

We call $q$ an improper distribution if: 1) $\int_{\mathcal{A}} q(a) \mathrm{d} a$ is within $(0,1]$ but not necessarily equal to one; and 2) $\mathbb{E}_{a \sim q}\left[a a^{\top}\right]$ is invertible. We define the improper expectation $\mathbb{E}_{a \sim q}[W(a)]$ for any random variable $W: \mathcal{A} \rightarrow \mathbb{R}$ by the integral $\int_{a \in \mathcal{A}} W(a) q(a) \mathrm{d} a$.

We aim to minimize the potential function

$$
\begin{equation*}
\Phi(q):=-2 \log \left(\operatorname{det}\left(\mathbb{E}_{a \sim q}\left[b_{a} b_{a}^{\top}\right]\right)\right)+\mathbb{E}_{a \sim q}[2 d+(\widehat{h}(\widehat{a})-\widehat{h}(a)) / \beta], \tag{B.4.1}
\end{equation*}
$$

where $b_{a}$ is the coefficient vector of $a$ when the basis is the barycentric spanner $\left\{A_{i}\right\}_{i=1}^{d}$. We prove that after each iteration, either Algorithm 4 outputs a desired distribution that satisfies both (2.5.1) and (2.5.2), or

$$
\begin{equation*}
\Phi\left(q_{t}\right) \leq \Phi\left(q_{t-1}\right)-\frac{1}{4} \tag{B.4.2}
\end{equation*}
$$

Since $\Phi$ function is bounded the algorithm must halt within finite iterations. (B.4.2) is a consequence of the following two lemmas:

Lemma 32 When $\mathbb{E}_{a \sim q(a)}[2 d+(\widehat{h}(\widehat{a})-\widehat{h}(a)) / \beta] \geq 2 d$, the objective $\Phi(q)$ will not increase if we multiply q by $\frac{2 d}{2 d+\mathbb{E}_{a \sim q}[2 d+(\widehat{h}(\widehat{a})-\widehat{h}(a)) / \beta]}$. That is, after step $(2.5 .3)$ in Algorithm 4, we always have

$$
\Phi\left(q_{t-\frac{1}{2}}\right) \leq \Phi\left(q_{t-1}\right)
$$

Lemma 33 If Algorithm 4 does not halt at round $t$, then after the coordinate descent step (2.5.5) in Algorithm 4, we always have

$$
\Phi\left(q_{t}\right) \leq \Phi\left(q_{t-\frac{1}{2}}\right)-\frac{1}{4}
$$

Now we present the proof of Proposition 3, as well as proofs of Lemma 32 and Lemma 33.

Proof of Proposition 3. From Lemma 32 and Lemma 33 we know that if the algorithm does not halt at round $t$, then $\Phi\left(q_{t}\right) \leq \Phi\left(q_{t-\frac{1}{2}}\right)-\frac{1}{4}$. Assume Algorithm 4 does not halt after $t$ rounds. Then we have

$$
\begin{array}{r}
2 d \log d+2 d+\frac{1}{\beta} \geq 2 d \log d+\mathbb{E}_{a \sim q_{0}}[2 d+(\widehat{h}(\widehat{a})-\widehat{h}(a)) / \beta]=\Phi\left(q_{0}\right) \\
\geq \Phi\left(q_{t}\right)+\frac{t}{4}=-2 \log \left(\operatorname{det}\left(\mathbb{E}_{a \sim q_{t}}\left[b_{a} b_{a}^{\top}\right]\right)\right)+\mathbb{E}_{a \sim q_{t}}[2 d+(\widehat{h}(\widehat{a})-\widehat{h}(a)) / \beta]+\frac{t}{4} \\
\geq-2 \log \left(\operatorname{det}\left(\mathbb{E}_{a \sim q_{t}}\left[b_{a} b_{a}^{\top}\right]\right)\right)+\frac{t}{4} . \tag{B.4.3}
\end{array}
$$

where the first inequality is due to $\int_{A} q_{0}(a) \mathrm{d} a=1$; the first equation is due to $-2 \log \left(\operatorname{det}\left(\mathbb{E}_{a \sim q_{0}}\left[b_{a} b_{a}^{\top}\right]\right)\right)=$ $-2 \log \left(\operatorname{det}\left(\frac{1}{d} I\right)\right)=2 d \log d$; and the last inequality is due to $\mathbb{E}_{a \sim q_{t}}[2 d+(\widehat{h}(\widehat{a})-\widehat{h}(a)) / \beta] \geq 0$.

Since the initialization actions consist of a barycentric spanner of $\mathcal{A}$, all coordinates of $b_{a}$ is within $[-1,1], \forall a \in \mathcal{A}$. Clearly $\left\|b_{a}\right\| \leq \sqrt{d}$ for all $a \in \mathcal{A}$. We know that $q_{t}$ is a improper
distribution with at most $d+t$ non-zero supports, so we assume $q_{t}=\sum_{i=1}^{d+t} q_{t}\left(A_{i}\right) \mathbb{1}_{A_{i}}$.

$$
\begin{array}{r}
\operatorname{det}\left(\mathbb{E}_{a \sim q_{t}}\left[b_{a} b_{a}^{\top}\right]\right)=\operatorname{det}\left(\sum_{i=1}^{d+t} q_{t}\left(A_{i}\right) b_{A_{i}} b_{A_{i}}^{\top}\right) \\
\leq\left(\frac{\operatorname{tr}\left(\sum_{i=1}^{d+t} q_{t}\left(A_{i}\right) b_{A_{i}} b_{A_{i}}^{\top}\right)}{d}\right)^{d}=\left(\frac{\sum_{i=1}^{d+t} q_{t}\left(A_{i}\right) \operatorname{tr}\left(b_{A_{i}} b_{A_{i}}^{\top}\right)}{d}\right)^{d} \\
\leq\left(\frac{\sum_{i=1}^{d+t} q_{t}\left(A_{i}\right)\left\|b_{A_{i}}\right\|^{2}}{d}\right)^{d} \leq 1,
\end{array}
$$

where the first inequality is due to the AM-GM inequality; the last inequality is due to $\left\|b_{a}\right\| \leq \sqrt{d}$ for all $a \in \mathcal{A}$ and $\sum_{i=1}^{d+t} q_{t}\left(A_{i}\right)=\int_{\mathcal{A}} q_{t}(a) d a \leq 1$. As a result, we obtain $\log \left(\operatorname{det}\left(\mathbb{E}_{a \sim q_{t}}\left[b_{a} b_{a}^{\top}\right]\right)\right) \leq$ 0. Combine this result with (B.4.3), we obtain

$$
t \leq 8 d(\log d+1)+\frac{4}{\beta}
$$

So Algorithm 4 must halt within at most $\left\lceil\frac{4}{\beta_{m}}+8 d(\log d+1)\right\rceil$ iterations. When it halts, it is straightforward to verify that the output distribution is proper and satisfies both (2.5.1) and (2.5.2).

Proof of Lemma 32. Denote $w(a)=(\widehat{h}(\widehat{a})-\widehat{h}(a)) / \beta$. Given an arbitrary improper distribution $q$, we view $\Phi(c \cdot q)$ as a function on the scaling factor $c$. By the chain rule, we can compute the derivative of this function with respect to $c$,

$$
\begin{aligned}
& \partial_{c} \Phi(c \cdot q)=\int_{\mathcal{A}}\left[\partial_{c q(a)} \Phi(c \cdot q)\right]\left(\partial_{c} c q(a)\right) \mathrm{d} a \\
= & \int_{\mathcal{A}}\left[-2 a^{\top}\left(\mathbb { E } _ { \widetilde { a } \sim c q } \left[{\left.\left.\left.\widetilde{a} \widetilde{a}^{\top}\right]\right)^{-1} a+2 d+w(a)\right] q(a) \mathrm{d} a}_{=}^{c} \mathbb{E}_{a \sim q}\left[a^{\top}\left(\mathbb{E}_{a \sim q}\left[a a^{\top}\right]\right)^{-1} a\right]+2 d+\mathbb{E}_{a \sim q}[w(a)],\right.\right.\right.
\end{aligned}
$$

where the second equation use the fact that the partial gradient of $\log \left(\operatorname{det}\left(\mathbb{E}_{a \sim c q}\left[b_{a} b_{a}^{\top}\right]\right)\right)$ with respect to the coordinate $c q(a)$ is $b_{a}^{\top}\left(\mathbb{E}_{\widetilde{a} \sim c q}\left[b_{\widetilde{a}} b_{\widetilde{a}}^{\top}\right]\right)^{-1} b_{a}=a^{\top}\left(\mathbb{E}_{\widetilde{a} \sim c q}\left[\widetilde{a} \widetilde{a}^{\top}\right]\right)^{-1} a$.

By the "trace trick", we have

$$
\begin{aligned}
\mathbb{E}_{a \sim q}\left[a^{T}\left(\mathbb{E}_{a \sim q}\left[a a^{T}\right]\right)^{-1} a\right] & =\operatorname{tr}\left(\mathbb{E}_{a \sim q}\left[a^{T}\left(\mathbb{E}_{a \sim q}\left[a a^{T}\right]\right)^{-1} a\right]\right) \\
& =\mathbb{E}_{a \sim q}\left[\operatorname{tr}\left(a^{T}\left(\mathbb{E}_{a \sim q}\left[a a^{T}\right]\right)^{-1} a\right)\right] \\
& =\mathbb{E}_{a \sim q}\left[\operatorname{tr}\left(a a^{T}\left(\mathbb{E}_{a \sim q}\left[a a^{T}\right]\right)^{-1}\right)\right] \\
=\operatorname{tr} & \left.\left(\mathbb{E}_{a \sim q} a a^{T}\left(\mathbb{E}_{a \sim q}\left[a a^{T}\right]\right)^{-1}\right)\right)=d .
\end{aligned}
$$

So we have

$$
\partial_{c} \Phi(c \cdot q)=-\frac{2 d}{c}+2 d+\mathbb{E}_{a \sim q}[w(a)] .
$$

This means that for all $c \in\left[\frac{2 d}{2 d+\mathbb{E}_{a \sim q}[w(a)]}, 1\right], \partial_{c} \Phi(c \cdot q) \geq 0$.

Proof of Lemma 33. We define

$$
\Delta_{t}=\frac{-2 a_{t}^{\top}\left(\mathbb{E}_{a \sim q}\left[a a^{\top}\right]\right)^{-1} a_{t}+2 d+\left(\widehat{h}(\widehat{a})-\widehat{h}\left(a_{t}\right)\right) / \beta}{\left(a_{t}^{\top}\left(\mathbb{E}_{a \sim q}\left[a a^{\top}\right]\right)^{-1} a_{t}\right)^{2}}
$$

then the coordinate descent step (2.5.5) is $q_{t}=q_{t-\frac{1}{2}}+\Delta_{t} \mathbb{1}_{a_{t}}$.

$$
\begin{array}{r}
\Phi\left(q_{t-\frac{1}{2}}\right)-\Phi\left(q_{t}\right)=2 \log \left(\frac{\operatorname{det}\left(\mathbb{E}_{a \sim q^{+}}\left[b_{a} b_{a}^{\top}\right]\right)}{\operatorname{det}\left(\mathbb{E}_{a \sim q}\left[b_{a} b_{a}^{\top}\right]\right)}-\Delta_{t}\left(2 d+\left(\widehat{h}(\widehat{a})-\widehat{h}\left(a_{t}\right)\right) / \beta\right)\right. \\
=2 \log \left(1+\Delta_{t} a_{t}^{\top}\left(\mathbb{E}_{a \sim q}\left[a a^{\top}\right]\right)^{-1} a_{t}\right)-\Delta_{t}\left(2 d+\left(\widehat{h}(\widehat{a})-\widehat{h}\left(a_{t}\right)\right) / \beta\right) \\
\geq 2 a_{t}^{\top}\left(\mathbb{E}_{a \sim q}\left[a a^{\top}\right]\right)^{-1} a_{t} \Delta_{t}-\left(a_{t}^{\top}\left(\mathbb{E}_{a \sim q}\left[a a^{\top}\right]\right)^{-1} a_{t}\right)^{2} \Delta_{t}^{2}-\Delta_{t}\left(2 d+\left(\widehat{h}(\widehat{a})-\widehat{h}\left(a_{t}\right)\right) / \beta\right) \\
=  \tag{B.4.6}\\
\left(-2 a_{t}^{\top}\left(\mathbb{E}_{a \sim q}\left[a a^{\top}\right]\right)^{-1} a_{t}+2 d+\left(\widehat{h}(\widehat{a})-\widehat{h}\left(a_{t}\right)\right) / \beta\right)^{2} \\
4\left(a_{t}^{\top}\left(\mathbb{E}_{a \sim q}\left[a a^{\top}\right]\right)^{-1} a_{t}\right)^{2}
\end{array},
$$

where (B.4.4) is due to the matrix determinant lemma as well as $a_{t}^{\top}\left(\mathbb{E}_{a \sim q}\left[a a^{\top}\right]\right)^{-1} a_{t}=b_{a_{t}}^{\top}\left(\mathbb{E}_{a \sim q}\left[b_{a} b_{a}{ }^{\top}\right]\right)^{-1} b_{a_{t}}$; (B.4.5) is due to the inequality $\log (1+w) \geq w-\frac{w^{2}}{2}$ for all $w \geq 0$; (B.4.6) is due to the definition
of $\Delta_{t}$ which maximize the quadratic function in (B.4.5).
As the algorithm does not halt, we have $a_{t}^{\top}\left(\mathbb{E}_{a \sim q}\left[a a^{\top}\right]\right)^{-1} a_{t} \geq 2 d+\left(\widehat{h}(\widehat{a})-\widehat{h}\left(a_{t}\right)\right) / \beta$, so

$$
\begin{array}{r}
\left|-2 a_{t}^{\top}\left(\mathbb{E}_{a \sim q}\left[a a^{\top}\right]\right)^{-1} a_{t}+2 d+\left(\widehat{h}(\widehat{a})-\widehat{h}\left(a_{t}\right)\right) / \beta\right| \\
=2 a_{t}^{\top}\left(\mathbb{E}_{a \sim q}\left[a a^{\top}\right]\right)^{-1} a_{t}-2 d-\left(\widehat{h}(\widehat{a})-\widehat{h}\left(a_{t}\right)\right) / \beta \\
\geq a_{t}^{\top}\left(\mathbb{E}_{a \sim q}\left[a a^{\top}\right]\right)^{-1} a_{t}
\end{array}
$$

Combine this inequality with (B.4.6) we obtain

$$
\Phi\left(q_{t-\frac{1}{2}}\right)-\Phi\left(q_{t}\right) \geq \frac{1}{4}
$$

## Appendix C: Appendix for Chapter 3

## C. 1 Extensions and proofs for AIR

## C.1.1 Extensions of AIR

Extension to general Bregman divergence We can generalize AIR from using KL divergence to using general Bregman divergence. And all the results in Section 3.3 can be extended as well. This generalization is inspired by [107], which defines information ratio and studies algorithm design using general Bregman divergence.

Let $\Psi: \Delta(\Pi) \rightarrow \mathbb{R} \cup \infty$ be a convex, Legendre, and second-order differentiable function. Denote $D_{\Psi}$ to be the Bregman divergence of $\Psi$, and diam $(\Psi)$ to be the diameter of $\Psi$ (see Appendix C.4.3) for the background). Given a reference probability $q \in \operatorname{int}(\Delta(\Pi))$ in the interior of the simplex and learning rate $\eta>0$, we define the generalized Algorithmic Information Ratio with potential function $\Psi$ for decision $p$ and distribution $v$ by

$$
\begin{equation*}
\operatorname{AIR}_{q, \eta}^{\Psi}(p, v)=\mathbb{E}\left[f_{M}\left(\pi^{*}\right)-f_{M}(\pi)-\frac{1}{\eta} D \Psi\left(v_{\pi^{*} \mid \pi, o}, v_{\pi^{*}}\right)-\frac{1}{\eta} D_{\Psi}\left(v_{\pi^{*}}, q\right)\right] . \tag{C.1.1}
\end{equation*}
$$

We generalize Theorem 14 and Theorem 17 to second-order differentiable Bregman divergence for general convex Legendre function, with the $\frac{\log |\Pi|}{\eta}$ term in the regret bound be replaced by $\frac{\operatorname{diam}(\Psi)}{\eta}$. Using the extension, we can generalize Theorem 15 (regret of APS) and Theorem 16 (regret of AMS) to generalized Bregman divergence as well, where the definition of information ratio will also use the corresponding Bregman divergence as in [107]. We state the extension of Theorem 17 here.

Theorem 25 (Using general Bregman divergence) Assume $\Psi: \Delta(\Pi) \rightarrow \mathbb{R} \cup \infty$ is convex, Legendre, second-order differentiable, and has bounded diameter. Given a compact $\mathcal{M}$, an arbitrary
algorithm ALG that produces decision probability $p_{1}, \ldots, p_{T}$, and a sequence of beliefs $v_{1}, \ldots, v_{T}$ where $\left(v_{t}\right)_{\pi^{*} \mid \pi, o} \in \operatorname{int}(\Delta(\Pi))$ for all rounds, we have

$$
\begin{array}{r}
\boldsymbol{R}_{T} \leq \frac{\operatorname{diam}(\Psi)}{\eta}+\sum_{t=1}^{T}\left(\operatorname{AIR}_{q_{t}, \eta}^{\Psi}\left(p_{t}, v_{t}\right)\right. \\
\left.+\sup _{v^{*}}\left(\left.\frac{\partial \operatorname{AR}_{q_{t}, \eta}^{\Psi}\left(p_{t}, v\right)}{\partial v}\right|_{v=v_{t}}\right)^{\top}\left(v^{*}-v_{t}\right)\right) .
\end{array}
$$

Extension to high probability bound We conjecture that the results in Section 3.3 may be able to be extended to high probability bounds, with some modification in our algorithm and complexity measure. We refer to [108] for a possible approach to achieve this goal.

## C.1.2 Proof of Theorem 14

By the discussion in Section 3.5.1 and 3.5.2, we only need to prove Lemma 11 in order to prove Theorem 14.

Proof of Lemma 11: let $Q$ the space of all mappings from $\Pi \times O$ to $\Delta(\Pi)$. For a mapping $Q \in Q$, denote $Q\left[\pi_{t}, o_{t}\right] \in \Delta(\Pi)$ as the image of $(\pi, o)$. Define $B: \Delta(\mathcal{M} \times \Pi) \times Q \rightarrow \mathbb{R}$ by

$$
\begin{equation*}
B(v, Q)=\mathbb{E}\left[f_{M}\left(\pi^{*}\right)-f_{M}(\pi)-\frac{1}{\eta} \log \frac{Q[\pi, o]\left(\pi^{*}\right)}{q\left(\pi^{*}\right)}\right] . \tag{C.1.2}
\end{equation*}
$$

$B(v, Q)$ is linear with respect to $v$, convex with respect to $Q$. In order to apply minimax theorem to the concave-convex objective function $B$, we need to verify that the sets $Q$ and $\Delta(\mathcal{M} \times \Pi)$ are convex and compact sets, and $B$ is continuous with respect to both $Q \in Q$ and $v \in \Delta(\mathcal{M} \times \Pi)$. This verification step assumes a basic understanding of general topology, as it involves infinite sets (compactness and continuity for finite sets are trivial); and eager readers may choose to skip this step. For this reason we put the verification step to the end of the proof. By applying Sion's
minimax theorem (Lemma 37), we have

$$
\begin{equation*}
\sup _{v} \inf _{Q} B(v, Q)=\inf _{Q} \sup _{v} B(v, Q) . \tag{C.1.3}
\end{equation*}
$$

From the definition of AIR and first-order optimality condition, we have

$$
\begin{aligned}
& \operatorname{AIR}_{q, \eta}(p, v) \\
= & \mathbb{E}\left[f_{M}\left(\pi^{*}\right)-f_{M}(\pi)-\frac{1}{\eta} \operatorname{KL}\left(\mathbb{P}_{v}\left(\pi^{*} \mid \pi, o\right), q\right)\right] \\
= & \inf _{Q \in Q} B(v, Q),
\end{aligned}
$$

so $\bar{v}$ is the maximizer of $\sup _{v} \inf _{Q} B(v, Q)$. Define $Q_{\bar{v}}$ as the mapping that maps each $(\pi, o)$ to the conditional probability $\mathbb{P}_{\bar{v}}\left(\pi^{*} \mid \pi, o\right)$, then $Q_{\bar{v}}$ is the unique minimizer of $B(\bar{v}, Q)$. Denote $\bar{Q}$ to be the minimizer of $\inf _{Q} \sup _{v} B(v, Q)$. From the equality (C.1.3), $(\bar{v}, \bar{Q})$ must be a Nash equilibrium of $B$, i.e.

$$
\operatorname{AIR}_{q, \eta}(p, \bar{v})=\sup _{v} \inf _{Q} B(v, Q)=B(\bar{v}, \bar{Q})=\inf _{Q} \sup _{v} B(v, Q) .
$$

Then $\bar{Q}$ is a minimizer of $B(v, Q)$, which implies $Q_{\bar{v}}=\bar{Q}$ as the minimizer is unique. As a result, we have

$$
\begin{aligned}
& \operatorname{AIR}_{q, \eta}(p, \bar{v}) \\
= & \sup _{v} B\left(v, Q_{\bar{v}}\right) \\
= & \sup _{M, \pi^{*}} \mathbb{E}\left[f_{M}\left(\pi^{*}\right)-f_{M}(\pi)-\frac{1}{\eta} \log \frac{\mathbb{P}_{\bar{v}}\left(\pi^{*} \mid \pi, o\right)}{q\left(\pi^{*}\right)}\right] .
\end{aligned}
$$

Verification of the conditions of Sion's minimax theorem: It is straightforward to see convexity of the sets $Q$ and $\Delta(\mathcal{M} \times \Pi)$. As a collection of mappings, $Q$ is compact with respect to the product topology by Tychonoff's theorem, and $B$ is continuous with respect to $Q$ by the definition
of product topology. Because the probability measure on the compact set is compact with respect to the weak*-topology, $\Delta(\mathcal{M} \times \Pi)$ is a compact set. We refer to the book [145] for the basic background of general topology. Finally, $B$ is continuous with respect to $v$ because $B$ is linear in $v$.

## C.1.3 Proof of Theorem 16

Combining Theorem 14 and Lemma 8, we prove Theorem 16.

## C.1.4 Proof of Theorem 15

We prove the following lemma that upper bounds $\operatorname{AIR}_{q_{t}, \eta}\left(q_{t}, v_{t}\right)$ by DEC and information ratio for Thompson Sampling. Theorem 15 will be a straightforward consequence of the regret bound (3.3.1) in Theorem 14 and this lemma.

Lemma 34 (Bounding AIR by DEC and IR for TS) Assume that $f_{M}(\pi)$ is bounded in $[0,1]$ for all $M, \pi$. Then for $\eta \in(0,1 / 2]$ and all $q \in \operatorname{int}(\Delta(\Pi))$, we have

$$
\operatorname{AIR}_{q, \eta}(q, v) \leq \sup _{M \in \Delta(\mathcal{M})} \operatorname{DEC}_{2 \eta}(\Delta(\mathcal{M}), M)+2 \eta \leq \frac{\eta}{2} \cdot \mathrm{IR}^{\mathrm{TS}}+2 \eta
$$

Proof of Lemma 34: Given a probability measure $v$, Denote $v_{o \mid \pi^{*}, \pi}=\mathbb{E}_{M \sim v_{M \mid \pi^{*}}}[M(\pi)]$ to be the posterior belief of observation $o$ conditioned on $\pi^{*}$ and $\pi$, and $v_{o \mid \pi}=\mathbb{E}_{\left(M, \pi^{*}\right) \sim \nu}[M(\pi)]$ to be posterior belief of $o$ conditioned solely on $\pi$.

Denote the $|\Pi|$-dimensional vector $X, Y$ by

$$
\begin{aligned}
X(\pi) & =\mathbb{E}_{\left(M, \pi^{*}\right) \sim v}\left[f_{M}\left(\pi^{*}\right)-f_{M}(\pi)\right] \\
Y(\pi) & =\mathbb{E}_{\left(M, \pi^{*}\right) \sim v}\left[D_{\mathrm{H}}^{2}\left(v_{o \mid \pi^{*}, \pi}, v_{o \mid \pi}\right)\right]
\end{aligned}
$$

Note that the Algorithmic Information Ratio can always be written as

$$
\begin{align*}
& \operatorname{AIR}_{q, \eta}(p, v)=\mathbb{E}_{v, p}\left[f_{M}\left(\pi^{*}\right)-f_{M}(\pi)-\frac{1}{\eta} \operatorname{KL}\left(v_{\pi^{*} \mid \pi, o}, q\right)\right] \\
= & \mathbb{E}_{v, p}\left[f_{M}\left(\pi^{*}\right)-f_{M}(\pi)-\frac{1}{\eta} \operatorname{KL}\left(v_{\pi^{*} \mid \pi, o}, v_{\pi}^{*}\right)-\frac{1}{\eta} \operatorname{KL}\left(v_{\pi^{*}}, q\right)\right] \\
= & \mathbb{E}_{v, p}\left[f_{M}\left(\pi^{*}\right)-f_{M}(\pi)-\frac{1}{\eta} \operatorname{KL}\left(v_{o \mid \pi^{*}, \pi}, v_{o \mid \pi}\right)-\frac{1}{\eta} \operatorname{KL}\left(v_{\pi^{*}}, q\right)\right], \tag{C.1.4}
\end{align*}
$$

where the first equality is the definition of AIR; the second equality is because the expectation of posterior is equal to prior; and the third equality is due to the symmetry property of mutual information. By (C.1.4) we have that

$$
\begin{align*}
& \operatorname{AIR}_{q, \eta}(q, v) \\
& \leq \mathbb{E}_{v, q}\left[f_{M}\left(\pi_{M}\right)-f_{M}(\pi)-\frac{1}{\eta} D_{\mathrm{H}}^{2}\left(v_{o \mid \pi^{*}, \pi}, v_{o \mid \pi}\right)\right]-\frac{1}{\eta} \operatorname{KL}\left(v_{\pi}^{*}, \rho_{t}\right) \\
&=\langle q, X\rangle-\frac{1}{2 \eta} \mathrm{KL}\left(v_{\pi^{*}}, q\right)-\frac{1}{\eta}\langle q, Y\rangle-\frac{1}{2 \eta} \mathrm{KL}\left(v_{\pi^{*}}, q\right) \\
& \leq\left\langle v_{\pi^{*}}, X\right\rangle+2 \eta-\frac{1}{\eta}\langle q, Y\rangle-\frac{1}{2 \eta} \mathrm{KL}\left(v_{\pi^{*}}, q\right) \\
& \leq\left\langle v_{\pi^{*}}, X\right\rangle+2 \eta-\frac{1}{\eta}\langle q, Y\rangle-\frac{1}{2 \eta} D_{\mathrm{H}}^{2}\left(v_{\pi^{*}}, q\right) \\
& \leq\left\langle v_{\pi^{*}}, X\right\rangle-\frac{(1-\eta)}{(1+\eta) \eta}\left\langle v_{\pi^{*}}, Y\right\rangle+2 \eta \\
& \leq \mathbb{E}_{v, v_{\pi^{*}}}\left[f_{M}\left(\pi_{M}\right)-f_{M}(\pi)-\frac{1}{2 \eta} D_{\mathrm{H}}^{2}\left(v_{o \mid \pi^{*}, \pi}, v_{o \mid \pi}\right)\right]+2 \eta \tag{C.1.5}
\end{align*}
$$

where the first inequality is by Lemma 41 ; the second inequality is by Lemma 43 and the fact $f_{M}(\pi) \in[0,1]$ for all $M \in \mathcal{M}$ and $\pi \in \Pi$; the third inequality is by Lemma 41; the fourth inequality is a consequence of Lemma 42 and the AM-GM inequality; and the last inequality uses the condition $\eta \leq \frac{1}{2}$.

Finally, by combining (C.1.5) and Lemma 8, we have that

$$
\operatorname{AIR}_{q, \eta}(q, v) \leq \sup _{M \in \Delta(\mathcal{M})} \operatorname{DEC}_{2 \eta}(\Delta(\mathcal{M}), M)+2 \eta \leq \frac{\eta}{2} \cdot \mathrm{IR}+2 \eta
$$

## C.1.5 Proof of Theorem 17 and Theorem 25

In this section we prove Theorem 25, which is a more general extension to Theorem 17 (Theorem 25 applies to general Bregman divergence with second-order differentiable $\Psi$ while Theorem 17 is stated with the KL divergence). Theorem 17 and 25 are consequences of the following "envelop theorem," which shows that gradients of AIR with respect to $v$ is equal to the gradient of the adversary when one uses the posterior mapping as the decision rule.

Lemma 35 (Envelop theorem) Let $\mathcal{X}$ and $\mathcal{Y}$ be convex sets, and $\phi: \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ a function such that for all $y \in \mathcal{Y}, \phi(\cdot, y)$ is a Legendre function in $x$; and for all $x \in \mathcal{X},-\phi(x, \cdot)$ is a Legendre function in $y$. For each $y \in \mathcal{Y}$, let $x_{y}$ be the a minimizer of the convex optimization problem

$$
\min _{x \in \mathcal{X}} \phi(x, y)
$$

and assume that $x_{y}$ is differentiable with respect to $y$. Then for all $y \in \operatorname{int}(\operatorname{dom}(\psi))$, we have

$$
\frac{\partial \phi\left(x_{y}, y\right)}{\partial y}=\left.\frac{\partial \phi(x, y)}{\partial y}\right|_{x=x_{y}} .
$$

We consider generalized AIR, where $\Psi$ is a Legendre and second-order differentiable function. Recall in (C.1.1) we define

$$
\operatorname{AIR}_{q, \eta}^{\Psi}(p, v)=\mathbb{E}\left[f_{M}\left(\pi^{*}\right)-f_{M}(\pi)-\frac{1}{\eta} D \Psi\left(v_{\pi^{*}} \| \pi, o, v_{\pi^{*}}\right)-\frac{1}{\eta} D \Psi\left(v_{\pi^{*}}, q\right)\right] .
$$

And similar to (C.1.2), for all $v \in \Delta(\mathcal{M} \times \Pi)$ such that $\mathbb{P}_{v}\left(\pi^{*}\right) \in \operatorname{int}(\Delta(\Pi))$ and $R: \Pi \times O \rightarrow \mathbb{R}^{|\Pi|}$, we define

$$
\left.B(v, R)=\mathbb{E}\left[f_{M}\left(\pi^{*}\right)-f_{M}(\pi)+\frac{1}{\eta}\langle\nabla \Psi(q)-R[\pi, o]), \mathbb{1}\left(\pi^{*}\right)-q\right\rangle+\frac{1}{\eta} D_{\Psi^{*}}(R[\pi, o], \nabla \Psi(q))\right],
$$

where $\mathbb{1}\left(\pi^{*}\right)$ is the vector whose $\pi^{*}$-coordinate is 1 but all other coordinates are 0 . Note that $B(v, R)$ is a convex function with respect to $R$. By Lemma 39 and the property (b) in Lemma 38, we know that

$$
\operatorname{AIR}_{q, \eta}^{\Psi}(p, v)=B\left(v, \nabla \Psi\left(v_{\pi^{*} \mid, \cdot}\right)\right)=\inf _{R} B(v, R),
$$

where the last equality is by the first-order optimal condition. By Lemma 35, when $\Psi$ is secondorder differentiable, we further have

$$
\frac{\partial \mathrm{AIR}_{q, \eta}^{\Psi}(p, v)}{\partial v}=\left.\frac{\partial B(v, R)}{\partial v}\right|_{R=\nabla \Psi\left(v_{\pi^{*} \mid, .,}\right)}
$$

By the above identity and the linearity of $B(v, Q)$ with respect to $v$, we have

$$
B\left(v^{*}, v_{\pi^{*} \mid, \cdot}\right)=\operatorname{AIR}_{q, \eta}^{\Psi}(p, v)+\left\langle\frac{\partial \operatorname{AIR}_{q, \eta}^{\Psi}(p, v)}{\partial v}, v^{*}-v\right\rangle, \quad \forall v^{*} .
$$

Following the same steps in proving Theorem 14, we prove Theorem 25 (and consequently Theorem 17).

## C. 2 Details and proofs for bandit problems

## C.2.1 Concave parameterization with Bernoulli reward

We consider Bernoulli structured bandit with an action set $\Pi \subset \mathbb{R}^{d}$ and a mean reward function class $\mathcal{F} \subset(\Pi: \mapsto[0,1])$ that is convex. (As discussed in the beginning of Section 3.4, every bounded-reward bandit problem can equivalently be reduced to a Bernoulli bandit problem. For simplicity we make the standard assumption that $\Pi$ is finite, which can be removed using standard discretization and covering argument. The goal here is to make the computation complexity to be independent of the size of model class $\mathcal{M}$, but only depends on $|\Pi|$. The general principle to achieve this goal is as follows. For each possible value of $\pi^{*}$, we assign an "effective model" $M_{\pi^{*}}$ to $\pi^{*}$ so that the optimization problem (3.6.1) reduces to selecting those $|\Pi|$ "effective models," as
well as the probability distribution over them.
We introduce the following parametrization: $\forall a, a^{*} \in \Pi$ (we use notation $a^{*}, a$ as the index sometimes to avoid repetition of notation $\left.\pi^{*}, \pi\right)$,

$$
\begin{array}{r}
\theta_{a^{*}}(a)=\mathbb{E}\left[r(a) \mid \pi^{*}=a^{*}\right], \\
\alpha\left(a^{*}\right)=v_{a^{*} \mid \pi, o}\left(a^{*}\right), \\
\beta_{a^{*}}(a)=\alpha\left(a^{*}\right) \cdot \theta_{a^{*}}(a) .
\end{array}
$$

Then we have represent AIR by $\left(\alpha, \beta_{\pi^{*} \pi^{*} \in \Pi}\right)$ :

$$
\begin{array}{r}
\operatorname{AIR}_{q, \eta}(p, v)=\sum_{\pi^{*} \in \Pi} \beta_{\pi^{*}}\left(\pi^{*}\right)-\sum_{\pi^{*}, \pi \in[K]} p(\pi) \beta_{\pi^{*}}(\pi) \\
-\frac{1}{\eta} \sum_{\pi^{*}, \pi \in[K]} p(\pi) \alpha\left(\pi^{*}\right) \mathrm{kl}\left(\frac{\beta_{\pi^{*}}(\pi)}{\alpha(\pi)}, \sum_{\pi^{*} \in \Pi} \beta_{\pi^{*}}(\pi)\right)-\frac{1}{\eta} \mathrm{KL}(\alpha, q), \tag{C.2.1}
\end{array}
$$

and the constraint of $\left(\alpha, \beta_{\pi^{*} \pi^{*} \in \Pi}\right)$ is that the functions parameterized by $\theta_{a^{*}}$ belong to the mean reward function class $\mathcal{F}$. We know the constraint set of $\left(\alpha, \beta_{\pi^{*} \pi^{*} \in \Pi}\right)$ to be convex because the convexity of perspective function.

Now we want to prove that the AIR objective in the maximization problem (3.6.1) is concave. We have

$$
\begin{align*}
B(v, Q)=\mathbb{E}_{p, v}\left[f_{M}\left(\pi^{*}\right)-f_{M}(\pi)+\frac{1}{\eta} \int_{O} \mathbb{P}_{\nu}\left(\pi, o, \pi^{*}\right) \log \frac{q\left(\pi^{*}\right)}{Q[\pi, o]\left(\pi^{*}\right)} d o\right] \\
=\sum_{\pi^{*}} \beta_{\pi^{*}}\left(\pi^{*}\right)-\sum_{\pi, \pi^{*}} p(\pi) \beta_{\pi^{*}}(\pi)+\frac{1}{\eta} \sum_{\pi, \pi^{*}} p(\pi) \beta_{\pi^{*}}(\pi) \log \frac{q\left(\pi^{*}\right)}{Q[\pi, 1]\left(\pi^{*}\right)} \\
+\frac{1}{\eta} \sum_{\pi, \pi^{*}} p(\pi)\left(\alpha\left(\pi^{*}\right)-\beta_{\pi^{*}}\left(\pi^{*}\right)\right) \log \frac{q\left(\pi^{*}\right)}{Q[\pi, 0]\left(\pi^{*}\right)} \tag{C.2.2}
\end{align*}
$$

This means that after parameterizing $v$ with $\alpha$ and $\left\{\beta_{\pi^{*}}\right\}_{\pi^{*} \in \Pi}, B(v, Q)$ will be a linear function of
$\left(\alpha,\left\{\beta_{\pi^{*}}\right\}_{\pi^{*} \in \Pi}\right)$. As a result,

$$
\operatorname{AIR}_{q, \eta}(p, v)=\inf _{Q} B(v, Q)
$$

will be a concave function of $\left(\alpha,\left\{\beta_{\pi^{*}}\right\}_{\pi^{*} \in \Pi}\right)$. So the optimization problem to maximize AIR is a convex optimization problem, whose computational complexity will be poly-logarithmic to the cardinality of $\left(\alpha,\left\{\beta_{\pi^{*}}\right\}_{\pi^{*} \in \Pi}\right)$. As a result, the computational complexity to maximize AIR is polynomial in $|\Pi|$ and does not depends on cardinality of the model class. This discussion shows that we give finite-running-time algorithm with computational complexity poly $\left(e^{d}\right)$ even when the cardinality of model class is double-exponential. Still, the computation is only efficient for simple problems such as $K$-armed bandits, but we also give efficient algorithm for linear bandits in Appendix 3.6.2.

## C.2.2 Simplified APS for Bernoulli MAB

For Bernoulli $K$-armed bandits discussed in in Section 3.4.1, we give the details about how to use first-order optimality conditions to derive Algorithm 8.

We denote $v_{\pi^{*}}(i \mid j, 1)$ as the shorthand for $v_{\pi^{*}}(i \mid \pi=j, o=1)$, the conditional probability $\mathbb{P}\left(\pi^{*}=i \mid \pi=j, o=1\right)$ when the underlying probability measure is $v$.

By (C.2.2) and Lemma 35 (using the envelop theorem and the bivariate function (C.2.2) to calculate the derivatives is easier than directly calculating the derivatives of the AIR parameterization (C.2.1)), we have for each $i \in[K]$,

$$
\begin{equation*}
\frac{\partial \mathrm{AIR}_{q, \eta}(p, v)}{\partial \beta_{i}(i)}=(1-p(i))-\frac{1}{\eta} p(i)\left(\log v_{\pi^{*}}(i \mid i, 1)-\log v_{\pi^{*}}(i \mid i, 0)\right) . \tag{C.2.3}
\end{equation*}
$$

And for every $i \neq j \in[K]$,

$$
\begin{equation*}
\frac{\partial \mathrm{AIR}_{q, \eta}(p, v)}{\partial \beta_{i}(j)}=-p(j)-\frac{1}{\eta} p(j)\left(\log v_{\pi^{*}}(i \mid j, 1)-\log v_{\pi^{*}}(i \mid j, 0)\right) \tag{C.2.4}
\end{equation*}
$$

Lastly, for each $i \in[K]$,

$$
\begin{equation*}
\frac{\partial \operatorname{AIR}_{q, \eta}(p, v)}{\partial \alpha(i)}=\frac{1}{\eta} \sum_{j \in[K]} p(j)\left(\log q(i)-\log v_{\pi^{*}}(i \mid j, 0)\right) \tag{C.2.5}
\end{equation*}
$$

We let the derivatives in (C.2.3) and (C.2.4) be zero, which means that the derivatives with respect to all coordinates of $\beta$ are zero. We have for all $i \neq j \in[K]$

$$
\begin{aligned}
\log \frac{v_{\pi^{*}}(j \mid j, 1)}{v_{\pi^{*}}(j \mid j, 0)} & =\frac{\eta}{p(j)}-\eta, \\
\log \frac{1-v_{\pi^{*}}(i \mid j, 1)}{1-v_{\pi^{*}}(i \mid j, 0)} & =-\eta, \quad i \neq j
\end{aligned}
$$

Solving the above two equation we obtain

$$
\begin{align*}
& v_{\pi^{*}}(j \mid j, 1)=\frac{1-\exp (-\eta)}{1-\exp (-\eta / p(j))}, \quad \forall j \in[K], \\
& v_{\pi^{*}}(i \mid j, 1)=\frac{\exp (-\eta)-\exp (-\eta / p(j))}{1-\exp (-\eta / p(j))} \cdot \frac{\alpha(i)}{1-\alpha(j)}, \quad \forall i \neq j \in[K], \\
& v_{\pi^{*}}(j \mid j, 0)=\frac{\exp (\eta)-1}{\exp (\eta / p(j))-1}, \quad \forall j \in[K], \\
& v_{\pi^{*}}(i \mid j, 0)=\frac{\exp (\eta / p(j))-\exp (\eta)}{\exp (\eta / p(j))-1} \cdot \frac{\alpha(i)}{1-\alpha(j)}, \quad \forall i \neq j \in[K] . \tag{C.2.6}
\end{align*}
$$

Now we set $\alpha=p=q$ so that the posterior updates (C.2.6) all have closed forms. Now we want to prove that (C.2.5) (the derivatives with respect to coordinates of $\alpha$ ) are bounded by constants. As can be seen from (C.2.6), when the observed reward at the chosen action $j$ is $r_{t}=0$, the posterior $v_{\pi^{*}}(j \mid j, 0)$ for the chosen action will be smaller than its prior belief $q(j)$; and the posteriors $v_{\pi^{*}}(i \mid j, 0)$ will be larger than the prior beliefs $q(i)$ for all unchosen actions $i \neq j$. As a
result, we have for every $i \in[K]$,

$$
\begin{aligned}
\frac{\partial \operatorname{AR}_{q, \eta}(p, v)}{\partial \alpha(i)} & =\frac{1}{\eta} \sum_{j \in[K]} p(j) \log \frac{q(i)}{v_{\pi^{*}}(i \mid j, 0)} \\
& \leq \frac{1}{\eta} p(i) \log \frac{q(i)}{v_{\pi^{*}}(i \mid i, 0)} \\
& =\frac{1}{\eta} p(i) \log \frac{p(i)(\exp (\eta / p(i))-1)}{\exp (\eta)-1} \\
& \leq 1,
\end{aligned}
$$

where the first equality is by (C.2.5); the first inequality is because $q(i)<v_{\pi^{*}}(i \mid j, 0)$ for all $j \neq i$; the second equality because of (C.2.5) and $p=q$; and the last inequality is a consequence of the following application of Jensen's inequality:

$$
\frac{1}{1+p(i)} \exp (\eta)+\frac{p(i)}{1+p(i)} \exp \left(-\frac{\eta}{p(i)}\right) \geq 1
$$

Now we have shown that the derivatives of AIR with respect to all $\left\{\beta_{i}(j)\right\}_{i, j \in[K]}$ are zeros, and the derivatives of AIR with respect to all $\{\alpha(i)\}_{i \in[K]}$. We note that AIR is $\frac{1}{\eta}$-strongly convex with respect to $\alpha$ when the gradient with respect to $\left\{\beta_{i}(j)\right\}_{i, j \in[K]}$ are all zeros. Then by Theorem 17 and Theorem 15 we can prove that

Theorem 26 (Regret of Simplified APS for Bernoulli MAB) The regret of Algorithm 8 with $\eta=$ $\gamma=\sqrt{2 \log K /(K T+4 T)}$ is bounded as follows, for all $T \geq 2 K \log K+4$,

$$
\mathfrak{R}_{T} \leq \sqrt{(5 K+4) T \log K}
$$

## C.2.3 Surrogate concave objective with Gaussian reward

For structured bandit with Gaussian reward structure, we can formulate the optimization problem as a surrogate optimization problem, where all classical upper bounds about information ratio in practical applications apply (e.g., see the square-loss formulation of IR in [104, 117]). For
the simplicity of presentation, we restrict our attention to Gaussian reward with mean bounded in $[0,1]$ and variance $\sigma^{2} \leq 1$.

Denote $\alpha=\mathbb{P}_{v}\left(\pi^{*}\right)$ and $\beta_{\pi^{*}}=\mathbb{P}_{\nu}\left(\pi^{*}\right) \cdot \theta_{\pi^{*}}$. Define a variant of AIR as

$$
\begin{equation*}
\overline{\mathrm{AIR}}=\mathbb{E}\left[f_{M}\left(\pi^{*}\right)-f_{M}(\pi)-\frac{1}{\eta} \mathrm{KL}\left(N\left(\theta_{\pi^{*}}(\pi), 1\right), N\left(\theta_{\mathrm{avg}}(\pi), 1\right)\right)-\frac{1}{\eta} \mathrm{KL}(\alpha, q)\right], \tag{C.2.7}
\end{equation*}
$$

where we denote

$$
\theta_{\mathrm{avg}}=\sum_{\pi^{*} \in \Pi} \alpha\left(\pi^{*}\right) \theta_{\pi^{*}}
$$

And we define

$$
\begin{align*}
\bar{B}(v, \omega) & =\mathbb{E}_{\left(M, \pi^{*}\right) \sim v, \pi \sim p}\left[f_{M}\left(\pi^{*}\right)-f_{M}(\pi)-\frac{1}{2 \eta \sigma^{2}}\left(\left(\theta_{\pi^{*}}^{\omega}(\pi)-o\right)^{2}-\left(\theta_{\mathrm{avg}}^{\omega}(\pi)-o\right)^{2}\right)-\frac{1}{\eta} \log \frac{\alpha^{\omega}\left(\pi^{*}\right)}{q\left(\pi^{*}\right)}\right] \\
= & \mathbb{E}\left[f_{M}\left(\pi^{*}\right)-f_{M}(\pi)-\frac{1}{2 \eta \sigma^{2}}\left(\theta_{\pi^{*}}^{\omega}(\pi)^{2}-\theta_{\mathrm{avg}}^{\omega}(\pi)^{2}\right)-\frac{1}{\eta} \log \frac{\alpha^{\omega}\left(\pi^{*}\right)}{q\left(\pi^{*}\right)}\right]+\frac{1}{\eta \sigma^{2}} \mathbb{E}\left[o\left(\theta_{\pi^{*}}-\theta_{\mathrm{avg}}^{\omega}\right)\right] . \tag{C.2.8}
\end{align*}
$$

Then we have

$$
\overline{\operatorname{AIR}}_{q, \eta}(p, v)=\inf _{\omega} \bar{B}(v, \omega) .
$$

This means that $\overline{\operatorname{AIR}}_{q, \eta}(p, v)$ will be a concave function of $\left(\alpha,\left\{\beta_{\pi^{*}}\right\}_{\pi^{*} \in \Pi}\right)$. And we can develop a parallel theory for approximately optimizing $\overline{\operatorname{AIR}}$ as we have done for AIR in Section 3.3.4. In particular, we need to verify that $\bar{B}(v, \omega)$ is always a upper bound of $B\left(v, \mathbb{P}_{\omega}\left(\pi^{*} \mid \cdot, \cdot\right)\right)$ in order to derive regret bounds. This is true if we assume the variance $\sigma^{2} \leq 1$ as the normal probability density function is locally concave round its mean.

## C.2.4 Simplified APS for Gaussian linear bandits, relationship with IPW

In this subsection we derive Algorithm 9 for adversarial linear bandits with Gaussian reward. As the decision space is an $d$-dimensional action set $\Pi=\mathcal{A} \subseteq \mathbb{R}^{d}$, we will use the notations $\mathcal{A}$ (as action set), $a$ (as action) and $a^{*}$ (as optimal action) to follow the tradition of literature about linear bandits.

By (C.2.8) and Lemma 35, we have for each $a^{*} \in \mathcal{A}$,

$$
\begin{equation*}
\frac{\partial \overline{\operatorname{AIR}}_{q, \eta}(p, v)}{\partial \beta_{a^{*}}}=a^{*}-\mathbb{E}_{a \sim p}[a]-\frac{1}{\eta} \mathbb{E}_{a \sim p}\left[a a^{\top}\right]\left(\theta_{a^{*}}-\theta_{\text {avg }}\right) \tag{C.2.9}
\end{equation*}
$$

And for each $a^{*} \in \mathcal{A}$,

$$
\begin{equation*}
\frac{\partial \overline{\operatorname{AIR}}_{q, \eta}(p, v)}{\partial \alpha\left(a^{*}\right)}=-\frac{1}{2 \eta} \int_{\mathcal{A}} p(a)\left(\theta_{a^{*}}^{\top} a-\theta_{\mathrm{avg}}^{\top} a\right)^{2} d a-\frac{1}{\eta} \log \frac{\alpha\left(a^{*}\right)}{q\left(a^{*}\right)} . \tag{C.2.10}
\end{equation*}
$$

Let the derivatives in (C.2.9) be zero. If the matrix $\mathbb{E}_{a \sim p}\left[a a^{\top}\right]$ have full rank, then we have

$$
\begin{aligned}
\theta_{a^{*}}-\theta_{\text {avg }} & =\eta\left(\mathbb{E}_{a \sim p}\left[a a^{\top}\right]\right)^{-1}\left(a^{*}-\mathbb{E}_{a \sim p}[a]\right), \quad \forall a^{*} \in \mathcal{A}, \\
\alpha & =p .
\end{aligned}
$$

Taking the above relationship into (C.2.10), we have

$$
\frac{\partial \overline{\operatorname{AIR}}_{q, \eta}(p, v)}{\partial \alpha\left(a^{*}\right)}=-\frac{\eta}{2}\left(a^{*}-\mathbb{E}_{a \sim p}[a]\right)\left(\mathbb{E}_{a \sim p}\left[a a^{\top}\right]\right)^{-1}\left(a^{*}-\mathbb{E}_{a \sim p}[a]\right)-\frac{1}{\eta} \log \frac{\alpha\left(a^{*}\right)}{q\left(a^{*}\right)} .
$$

Assume the minimal eigenvalue of $\mathbb{E}_{a \sim p}\left[a a^{\top}\right]$ satisfies $\lambda_{\text {min }}\left(\mathbb{E}_{a \sim p}\left[a a^{\top}\right]\right) \geq \eta$, then one can verify that the following solution is approximately optimal to the problem (3.6.1) (with controllable precision):

$$
\begin{align*}
\alpha & =p, \\
\theta_{a^{*}} & =\eta\left(\mathbb{E}\left[a a^{\top}\right]\right)^{-1} a^{*}, \quad \forall a^{*} \in \mathcal{A} . \tag{C.2.11}
\end{align*}
$$

Note that this solution satisfies $\theta_{i} \in[0,1]^{K}$ for all $i \in[K]$.
By Bayes' rule and (C.2.11), the posterior update $\mathbb{P}_{\nu}\left(\pi^{*} \mid \pi, r(\pi)\right)$ can be expressed as follows. Given $a^{*} \in \mathcal{A}$, we have

$$
\begin{aligned}
\mathbb{P}_{v}\left(a^{*}=\bar{a} \mid a, r(a)\right) & =\frac{\alpha(\bar{a}) \exp \left(-\frac{1}{2}\left(r(a)-\theta_{\bar{a}}^{\top} a\right)^{2}\right)}{\int_{\mathcal{A}} p\left(a^{*}\right) \exp \left(-\frac{1}{2}\left(r(a)-\theta_{a^{*}}(a)\right)^{2}\right) d a^{*}} \\
= & \frac{\alpha(\bar{a}) \exp \left(r(a) \theta_{\bar{a}}^{\top} a-\frac{1}{2}\left(\theta_{\bar{a}}^{\top} a\right)^{2}\right)}{\int_{\mathcal{A}} \alpha\left(a^{*}\right) \exp \left(r(a) \theta_{a^{*}}^{\top} a-\frac{1}{2}\left(\theta_{a^{*}}^{\top} a\right)^{2}\right) d a^{*}} .
\end{aligned}
$$

The resulting algorithm is an exponential weight algorithm with a modified importance weight estimator

$$
\hat{r}_{t}(a)=a^{\top}\left(\mathbb{E}_{a \sim p}\left[a a^{\top}\right]\right)^{-1} a_{t} r_{t}\left(a_{t}\right)-\frac{\eta}{2}\left(a^{\top}\left(\mathbb{E}_{a \sim p}\left[a a^{\top}\right]\right)^{-1} a_{t}\right)^{2} .
$$

The forced exploration to ensure $\lambda_{\min }\left(\mathbb{E}_{a \sim p}\left[a a^{\top}\right]\right) \geq \eta$ can be done with the help of the volumetric spanners constructed in [118].

## C. 3 Proofs for MAIR

## C.3.1 Proof of Lemma 13

By Definition 10 we have

$$
\begin{aligned}
& \sup _{\rho \in \operatorname{int}(\Delta(\mathcal{M})))} \sup _{\mu \in \Delta(\mathcal{M})} \inf _{p \in \Delta(\Pi)} \operatorname{MAIR}_{\rho_{t}, \eta}(p, \mu) \\
= & \sup _{\rho \in \operatorname{int}(\Delta(\mathcal{M})))} \sup _{\mu \in \Delta(\mathcal{M})} \inf _{p \in \Delta(\Pi)} \mathbb{E}\left[f_{M}\left(\pi_{M}\right)-f_{M}(\pi)-\frac{1}{\eta} \operatorname{KL}\left(M(\pi), \mu_{o \mid \pi}\right)-\frac{1}{\eta} \operatorname{KL}(\mu, \rho)\right] \\
= & \sup _{\mu \in \Delta(\mathcal{M})} \inf _{p \in \Delta(\Pi)} \mathbb{E}\left[f_{M}\left(\pi_{M}\right)-f_{M}(\pi)-\frac{1}{\eta} \operatorname{KL}\left(M(\pi), \mu_{o \mid \pi}\right)\right] \\
\leq & \sup _{\bar{M} \in \operatorname{conv}(M)} \operatorname{DEC}_{\eta}^{\mathrm{KL}}(\mathcal{M}, \bar{M}),
\end{aligned}
$$

where the second equality is by (3.7.1) and the inequality is because Hellinger distance is bounded by KL divergence (Lemma 41).
C.3.2 Analysis of sequential estimation

Consider the optimistic Bayesian posterior update

$$
\begin{equation*}
\rho_{t+1}(M) \propto \exp \left(\sum_{s=1}^{t}\left(\log \left[M\left(\pi_{s}\right)\right]\left(o_{s}\right)+\eta W_{s}(M)\right)\right), \tag{C.3.1}
\end{equation*}
$$

where $\left\{W_{s}\right\}_{s=1}^{T}$ is a series of non-negative weights in $[0,1]^{M}$. When all $W_{s}(M)=0$ for all $M$ and $s$, the update reduces to the update of Bayesian posterior. We want to upper bound the cumulative estimation error of updating rule (C.3.1). We present the following theorem, whose proof is inspired by [119, 120].

Theorem 27 Applying the updating rule (C.3.1) with $W_{s}(M) \in[0,1]$ for all $M \in \mathcal{M}$ and $s=$ $1, \ldots, T$, we have

$$
\sum_{t=1}^{T} \mathbb{E}_{\mu_{t}, p_{t}}\left[D_{H}^{2}\left(M(\pi), M^{*}(\pi)\right)\right] \leq 2 \eta+1+2 \log (|\mathcal{M}| T)+4 \sum_{t=1}^{T} \mathbb{E}_{\mu_{t}, p_{t}}\left[W_{t}(M)\right]
$$

Proof of Theorem 27: denote $\mathbb{E}_{t}[\cdot]$ be the conditional expectation conditioned on the filtration from round 1 to round $t$. Denote

$$
Z_{t}(M)=\sum_{s=1}^{T}\left(\log \left[M\left(\pi_{s}\right)\right]\left(o_{s}\right)\right)+\eta W_{s}(M)
$$

We have

$$
\begin{aligned}
& \log \left(\sum_{M \in \mathcal{M}} \mathbb{E}_{t-1}\left[Z_{t}(M)\right]\right)-\log \left(\sum_{M \in \mathcal{M}} Z_{t-1}(M)\right) \\
= & \log \left(\sum_{M \in \mathcal{M}} \frac{Z_{t-1}(M)}{\sum_{M \in \mathcal{M}} Z_{t-1}(M)} \mathbb{E}_{t-1}\left[\exp \left(\log \left[M\left(\pi_{t}\right)\right]\left(o_{t}\right)+W_{t}(M)\right)\right]\right) \\
= & \log \left(\sum_{M \in \mathcal{M}} \mu_{t}(M) \cdot \mathbb{E}_{t-1}\left[\left[M\left(\pi_{t}\right)\right]\left(o_{t}\right)\right] \cdot \exp \left(W_{t}(M)\right)\right) .
\end{aligned}
$$

By the above equality, we have

$$
\begin{aligned}
& \log \left(\sum_{M \in \mathcal{M}} \mathbb{E}_{t-1}\left[Z_{t}(M)\right]\right)-\log \left(\sum_{M \in \mathcal{M}} Z_{t-1}(M)\right) \\
& \leq \sum_{M \in \mathcal{M}} \mu_{t}(M) \cdot \mathbb{E}_{t-1}\left[\left[M\left(\pi_{t}\right)\right]\left(o_{t}\right)\right] \cdot \exp \left(W_{t}(M)\right)-1 \\
&= \sum_{M \in \mathcal{M}} \mu_{t}(M) \cdot\left(\mathbb{E}_{t-1}\left[\left[M\left(\pi_{t}\right)\right]\left(o_{t}\right)\right]-1\right) \cdot \exp \left(W_{t}(M)\right) \\
&+\sum_{M \in \mathcal{M}} \mu_{t}(M) \cdot \exp \left(W_{t}(M)\right)-1 \\
& \leq \mathbb{E}_{\mu_{t}, p_{t}}\left[\int_{O}\left[M^{*}(\pi)\right](o)[M(\pi)](o) d o-1\right]+2 W_{t}(M) \\
& \leq \mathbb{E}_{\mu_{t}, p_{t}}\left[\int_{O} \sqrt{\left[M^{*}(\pi)\right](o)[M(\pi)](o)} d o-1\right]+2 W_{t}(M) \\
&=-\frac{1}{2} \mathbb{E}_{\mu_{t}, p_{t}}\left[D_{\mathrm{H}}^{2}\left(M(\pi), M^{*}(\pi)\right)\right]+2 W_{t}(M),
\end{aligned}
$$

where the first inequality is because $\log (1+z) \leq z$ for all $z \in \mathbb{R}$; the second inequality is because $e^{z} \leq 1+2 z$ for all $z \leq[0,1]$ and $W_{t}(M) \in[0,1]$; and the third inequality is because $\left[M^{*}(\pi)\right](o) \in$ [ 0,1 ] by the the fact $M^{*}(\pi)$ is a probability distribution over $O$.

Rearrange the above inequality, we conclude that

$$
\begin{align*}
& \mathbb{E}_{\mu_{t}, p_{t}}\left[D_{\mathrm{H}}^{2}\left(M(\pi), M^{*}(\pi)\right)\right] \\
\leq & 2 \log \left(\sum_{M \in \mathcal{M}} Z_{t-1}(M)\right)-2 \log \left(\sum_{M \in \mathcal{M}} \mathbb{E}_{t-1}\left[Z_{t}(M)\right]\right)+4 W_{t}(M) . \tag{C.3.2}
\end{align*}
$$

By lemma 40 , for any $\delta \in(0,1)$, with probability at least $1-\delta$,

$$
\begin{align*}
& \sum_{t=1}^{T}\left(\log \left(\sum_{M \in \mathcal{M}} Z_{t-1}(M)\right)-\log \left(\sum_{M \in \mathcal{M}} \mathbb{E}_{t-1}\left[Z_{t}(M)\right]\right)\right) \\
\leq & \sum_{t=1}^{T}\left(\log \left(\sum_{M \in \mathcal{M}} Z_{t-1}(M)\right)-\log \left(\sum_{M \in \mathcal{M}} Z_{t}(M)\right)\right)+\log \frac{1}{\delta} \\
= & \log \left(\sum_{M \in \mathcal{M}} Z_{0}(M)\right)-\log \left(\sum_{M \in \mathcal{M}} Z_{T}(M)\right)+\log \frac{1}{\delta} \\
\leq & \eta+\log |\mathcal{M}|+\log \frac{1}{\delta} . \tag{C.3.3}
\end{align*}
$$

Taking $\delta=1 / T$ in (C.3.3) and applying (C.3.2), we can show that

$$
\begin{aligned}
& \sum_{t=1}^{T} \mathbb{E}_{\mu_{t}, p_{t}}\left[D_{\mathrm{H}}^{2}\left(M(\pi), M^{*}(\pi)\right)\right] \\
\leq & \max \left\{1,2 \eta+2 \log (|\mathcal{M}| T)+4 \sum_{t=1}^{T} \mathbb{E}_{\mu_{t}, p_{t}}\left[W_{t}(M)\right]\right\} \\
\leq & 2 \eta+1+2 \log (|\mathcal{M}| T)+4 \sum_{t=1}^{T} \mathbb{E}_{\mu_{t}, p_{t}}\left[W_{t}(M)\right]
\end{aligned}
$$

## C.3.3 Proof for Theorem 18

Calculating the optimization error. Similar to Theorem 17, we can prove

$$
\begin{equation*}
\mathfrak{R}_{T} \leq \frac{\log |\mathcal{M}|}{\eta}+\sum_{t=1}^{T}\left(\operatorname{MAIR}_{\rho_{t}, \eta}\left(p_{t}, \mu_{t}\right)+\left\langle\left.\frac{\partial \operatorname{MAIR}_{\rho_{t}, \eta}\left(p_{t}, \mu\right)}{\partial \mu}\right|_{\mu=\mu_{t}}, \mathbb{1}\left(M^{*}\right)-\mu_{t}\right\rangle\right), \tag{C.3.4}
\end{equation*}
$$

where $\mathbb{1}\left(M^{*}\right)$ is the vector whose $M^{*}$ - coordinate is 1 but all other coordinates are 0 .

By Lemma 35 we have

$$
\begin{aligned}
& \frac{\partial \operatorname{MAIR}_{\rho, \eta}(p, \mu)}{\partial \mu(M)}=f_{M}\left(\pi_{M}\right)-\mathbb{E}_{\pi \sim p}\left[f_{M}(\pi)\right]+\frac{1}{\eta} \mathbb{E}_{\pi \sim p, o \sim M(\pi)}\left[\log \frac{\rho(M)}{\mu(M \mid \pi, o)}\right] \\
& =f_{M}\left(\pi_{M}\right)-\mathbb{E}_{\pi \sim p}\left[f_{M}(\pi)\right]-\frac{1}{\eta} \log \frac{\mu(M)}{\rho(M)}-\frac{1}{\eta} \mathbb{E}_{\pi \sim p, \bar{o} \sim M(\pi)}\left[\log \frac{[M(\pi)](\bar{o})}{\mu_{o}(\bar{o} \mid \pi)}\right] .
\end{aligned}
$$

By using the updating rule in (3.7.2), we have

$$
\frac{\partial \operatorname{MAIR}_{\rho, \eta}(p, \mu)}{\partial \mu(M)}=-\frac{1}{\eta} \mathbb{E}_{\pi \sim p, \bar{o} \sim M(\pi)}\left[\log \frac{[M(\pi)](o)}{\mu_{o}(\bar{o} \mid \pi)}\right]=-\frac{1}{\eta} \mathbb{E}_{p \sim \pi}\left[\operatorname{KL}\left(M(\pi), \mu_{o \mid \pi}\right)\right]
$$

which implies

$$
\begin{align*}
& \left.\left|\frac{\partial \operatorname{MAIR}_{\rho_{t}, \eta}\left(p_{t}, \mu\right)}{\partial \mu}\right|_{\mu=\mu_{t}}, \mathbb{1}\left(M^{*}\right)-\mu_{t}\right\rangle \\
= & \frac{1}{\eta} \mathbb{E}_{\mu_{t}, p_{t}}\left[\operatorname{KL}\left(M(\pi),\left(\mu_{t}\right)_{o \mid \pi}\right)\right]-\frac{1}{\eta} \mathbb{E}_{\pi \sim p}\left[\operatorname{KL}\left(M^{*}(\pi),\left(\mu_{t}\right)_{o \mid \pi}\right)\right] . \tag{C.3.5}
\end{align*}
$$

Taking (C.3.5) into (C.3.4), we have

$$
\begin{aligned}
& \mathfrak{R}_{T} \\
\leq & \frac{\log |\mathcal{M}|}{\eta}+\sum_{t=1}^{T} \mathbb{E}_{\mu_{t}, p_{t}}\left[f_{M}\left(\pi_{M}\right)-f_{M}(\pi)-\frac{1}{\eta} \operatorname{KL}\left(M^{*}(\pi),\left(\mu_{t}\right)_{o \mid \pi}\right)-\frac{1}{\eta} \operatorname{KL}\left(\mu_{t}, \rho_{t}\right)\right] .
\end{aligned}
$$

So we have

$$
\begin{equation*}
\mathfrak{R}_{T} \leq \frac{\log |\mathcal{M}|}{\eta}+\sum_{t=1}^{T} \mathbb{E}_{\mu_{t}, p_{t}}\left[f_{M}\left(\pi_{M}\right)-f_{M}(\pi)-\frac{1}{\eta} \mathrm{KL}\left(\mu_{t}, \rho_{t}\right)\right] . \tag{C.3.6}
\end{equation*}
$$

Refined analysis of Algorithm 10. At the same time, we have

$$
\begin{align*}
& \mathbb{E}_{\mu_{t}, p_{t}}\left[f_{M}\left(\pi_{M}\right)-f_{M}(\pi)-\frac{1}{\eta} \operatorname{KL}\left(\mu_{t}, \rho_{t}\right)\right] \\
= & \mathbb{E}_{\mu_{t}, p_{t}}\left[f_{M}\left(\pi_{M}\right)-f_{M}(\pi)-\frac{1}{\eta} D_{\mathrm{H}}^{2}\left(M^{*}(\pi), M(\pi)\right)\right] \\
& -\frac{1}{\eta} \operatorname{KL}\left(\mu_{t}, \rho_{t}\right)+\frac{1}{\eta} \mathbb{E}_{\pi \sim p_{t}, M \sim \mu_{t},}\left[D_{\mathrm{H}}^{2}\left(M(\pi), M^{*}(\pi)\right)\right] . \tag{C.3.7}
\end{align*}
$$

Applying Theorem 27 with

$$
W_{s}(M)=\eta\left(f_{M}\left(\pi_{M}\right)-\mathbb{E}_{p_{s}}\left[f_{M}(\pi)\right]\right),
$$

we have that

$$
\begin{equation*}
\frac{1}{\eta} \sum_{t=1}^{T} \mathbb{E}_{\mu_{t}, p_{t}}\left[D_{\mathrm{H}}^{2}\left(M(\pi), M^{*}(\pi)\right)\right] \leq 2+\frac{2 \log (|\mathcal{M}| T)+1}{\eta}+4 \sum_{t=1}^{T} \mathbb{E}_{\mu_{t}, p_{t}}\left[f_{M}\left(\pi_{M}\right)-f_{M}(\pi)\right] \tag{C.3.8}
\end{equation*}
$$

Combining (C.3.6), (C.3.7) and (C.3.8), we have

$$
\begin{aligned}
& \mathfrak{R}_{T} \\
\leq & \frac{\log |\mathcal{M}|}{\eta}+\sum_{t=1}^{T} \mathbb{E}_{\mu_{t}, p_{t}}\left[f_{M}\left(\pi_{M}\right)-f_{M}(\pi)-\frac{1}{\eta} \operatorname{KL}\left(\mu_{t}, \rho_{t}\right)\right] \\
\leq & \sum_{t=1}^{T} \mathbb{E}_{\mu_{t}, p_{t}}\left[5\left(f_{M}\left(\pi_{M}\right)-f_{M}(\pi)\right)-\frac{1}{\eta} D_{\mathrm{H}}^{2}\left(M(\pi), M^{*}(\pi)\right)-\frac{1}{\eta} \operatorname{KL}\left(\mu_{t}, \rho_{t}\right)\right] \\
& +2+\frac{2 \log (|\mathcal{M}| T)+1}{\eta}
\end{aligned}
$$

Therefore, we prove Theorem 18.

## C.3.4 Proof of Theorem 19

Consider the Bayesian posterior sampling strategy induced by $\mu \in \Delta(M)$, which samples $M \sim \mu$ and plays $\pi_{M}$. Denote the induced decision probability as

$$
\mu_{\pi_{M}}(\pi)=\sum_{M \in \mathcal{M}, \pi_{M}=\pi} \mu(M) .
$$

For arbitrary $\mu \in \Delta(M)$, denote the $|\Pi|$-dimensional vectors $X, Y$ by

$$
\begin{aligned}
X(\pi) & =\mathbb{E}_{\mu}\left[f_{M}\left(\pi_{M}\right)-f_{M}(\pi)\right], \\
Y(\pi) & =\mathbb{E}_{\mu}\left[D_{\mathrm{H}}^{2}\left(\mathbb{P}(o \mid M, \pi), \mathbb{P}_{\mu}(o \mid \pi)\right)\right] .
\end{aligned}
$$

Then

$$
\begin{aligned}
& \mathbb{E}_{\mu, \pi \sim \rho_{\pi_{M}}}\left[5\left(f_{M}\left(\pi_{M}\right)-f_{M}(\pi)\right)-\frac{1}{\eta} D_{\mathrm{H}}^{2}\left(M^{*}(\pi), M(\pi)\right)\right]-\frac{1}{\eta} \operatorname{KL}(\mu, \rho) \\
\leq & \left\langle\rho_{\pi_{M}}, 5 X\right\rangle-\frac{1}{\eta}\left\langle\rho_{\pi_{M}}, Y\right\rangle-\frac{1}{\eta} \operatorname{KL}\left(\mu_{\pi_{M}}, \rho_{\pi_{M}}\right) \\
\leq & \left\langle\mu_{\pi_{M}}, 5 X\right\rangle+10 \eta-\frac{1}{\eta}\left\langle\rho_{\pi_{M}}, Y\right\rangle-\frac{1}{2 \eta} \operatorname{KL}\left(\mu_{\pi_{M}}, \rho_{\pi_{M}}\right) \\
\leq & \left\langle\mu_{\pi_{M}}, 5 X\right\rangle+10 \eta-\frac{1}{\eta}\left\langle\rho_{\pi_{M}}, Y\right\rangle-\frac{1}{2 \eta} D_{\mathrm{H}}^{2}\left(\mu_{\pi_{M}}, \rho_{\pi_{M}}\right) \\
\leq & \left\langle\mu_{\pi_{M}}, 5 X\right\rangle-\frac{(1-\eta)}{(1+\eta) \eta}\left\langle\mu_{\pi_{M}}, Y\right\rangle+10 \eta \\
\leq & 5 \mathbb{E}_{\mu, \pi \sim \mu_{\pi_{M}}}\left[f_{M}\left(\pi_{M}\right)-f_{M}(\pi)-\frac{1}{2 \eta} D_{\mathrm{H}}^{2}\left(M^{*}(\pi), M(\pi)\right)\right]+2 \eta,
\end{aligned}
$$

where the first inequality is because KL divergence between induced decision distributions of two model distributions will be no larger than KL divergence between the two model distributions; the second inequality is by Lemma 43 and the fact $f_{M}(\pi) \in[0,1]$ for all $M$ and $\pi$; the third inequality is by Lemma 41; the fourth inequality is a consequence of Lemma 42 and the AM-GM inequality; and the last inequality uses the condition $\eta \leq \frac{1}{10}$.

Combining the above inequality with Theorem 18, we prove Theorem 19.

## C. 4 Technical backgrounds

## C.4.1 Conditional entropy

In the discussion after Definition 7, we utilize the following important result stating that conditional entropy is concave. The reference provides a succinct proof to this result.

Lemma 36 (Conditional entropy of a probability measure is concave, [146]) Let $\mathbb{P}$ be a probability measure on locally compact space $\mathcal{X}$, and let $\mathfrak{E}, \mathfrak{F}$ be countable partitions of the space. Define the entropy with respect to the partition $\mathfrak{E}$ as

$$
H(\mathbb{P}, \mathfrak{E})=-\sum_{E \in \mathfrak{E}} \mathbb{P}(E) \log \mathbb{P}(E)
$$

and the conditional entropy as

$$
H(\mathbb{P}, \mathfrak{E} \mid \mathfrak{F})=\sum_{F \in \mathfrak{F}} \mathbb{P}(F) H(\mathbb{P}(\cdot \mid F), \mathfrak{E})
$$

Then the conditional entropy $H(\mathbb{P}, \mathfrak{E} \mid \mathfrak{F})$ is a concave function with respect to $\mathbb{P}$.

## C.4.2 Minimax theorem

We introduce the classical minimax theorem for convex-concave game.

Lemma 37 (Sion's minimax theorem for values, [115]) Let $\mathcal{X}$ and $\mathcal{Y}$ be convex and compact sets, and $\psi: \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ a function which for all $y \in \mathcal{Y}$ is convex and continuous in $x$ and for all $x \in \mathcal{X}$ is concave and continuous in $y$. Then

$$
\min _{x \in \mathcal{X}} \max _{y \in \mathcal{Y}} \psi(x, y)=\max _{y \in \mathcal{Y}} \min _{x \in \mathcal{X}} \psi(x, y)
$$

## C.4.3 Convex analysis

Let $\mathcal{W}$ be a $d$-dimensional convex decision set. Let the potential function $\Psi: \mathbb{R}^{d} \rightarrow \mathbb{R} \cup \infty$ be a proper convex function that is Legendre (see §26 in [147] for more background). We assume $\mathcal{W} \subset \operatorname{dom}(\Psi):=\left\{u \in \mathbb{R}^{d}: \Psi(u)<\infty\right\}$ and bounded diameter of potential, e.g.,

$$
\operatorname{diam}(\mathcal{W}):=\sup _{u, v \in \mathcal{W}} \Psi(u)-\Psi(v)<\infty
$$

Define the Fenchel-Legendre dual of $\Psi$ as

$$
\Psi^{*}(a)=\sup _{u \in \mathbb{R}^{d}}\langle a, u\rangle-\Psi(u), \quad \forall a \in \mathbb{R}^{d}
$$

Define the Bregman divergences with respect to $\Psi$ and $\Psi^{*}$ as

$$
\begin{gathered}
D_{\Psi}(u, v)=\Psi(u)-\Psi(v)-\langle\nabla \Psi(v), u-v\rangle, \\
D_{\Psi *}(a, b)=\Psi^{*}(a)-\Psi^{*}(b)-\left\langle\nabla \Psi^{*}(b), a-b\right\rangle .
\end{gathered}
$$

Lemma 38 (Properties of Legendre function, [110]) If $\Psi$ is a Legendre function, then
(a) $\nabla \Psi$ is a bijection between $\operatorname{int}(\operatorname{dom}(\Psi))$ and $\operatorname{int}\left(\operatorname{dom}\left(\Psi^{*}\right)\right)$ with the inverse $(\nabla \Psi)^{-1}=\nabla \Psi^{*}$. That is, for $u \in \operatorname{int}(\operatorname{dom}(\Psi))$, if $a=\nabla \Psi(u)$, then $a \in \operatorname{int}\left(\operatorname{dom}\left(\Psi^{*}\right)\right)$ and $\nabla \Psi^{*}(a)=u$;
(b) $D_{\Psi}(u, v)=D_{\Psi_{*}}(\nabla \Psi(v), \nabla \Psi(u))$ for all $u, v \in \operatorname{int}(\operatorname{dom}(\Psi)) ;$ and
(c) the Fenchel conjugate $\Psi^{*}$ is Legendre.

Note that the property (a) in Lemma 38 is a foundational results in convex optimization-in order to optimize a convex function, one only needs to optimize its Fenchel dual function (in the sense of making gradient small). For example, mirror descent, dual averaging and follow the regularized leader are procedures based on this principle. This property is a special case of Lemma 35, the "envelop theorem."

We also introduce a property of Bregman divergence.

Lemma 39 (Generalized Pythagorean theorem) For a convex function $\Psi: \mathcal{W} \rightarrow \mathbb{R} \cup \infty$ and $u, v, w \in \mathcal{W}$, we have

$$
D_{\Psi}(u, v)-D_{\Psi}(v, w)-D_{\Psi}(w, u)=\langle u-w, \nabla \Psi(w)-\nabla \Psi(v)\rangle .
$$

## C.4.4 Concentration inequality

We introduce a one-sided martingale concentration inequality, Lemma A. 4 in [105], for a sequence of random variables.

Lemma 40 (Martingale concentration inequality) For any sequence of real-valued random variables $\left\{X_{t}\right\}_{t=1}^{T}$ adapted to a filtration $\left\{\mathfrak{F}_{t}\right\}_{t=1}^{T}$, it holds that for any $\delta \in(0,1)$, with probability at least $1-\delta$, for all $T^{\prime} \leq T$,

$$
\sum_{t=1}^{T^{\prime}} X_{t} \leq \sum_{t=1}^{T^{\prime}} \log \left(\mathbb{E}_{t-1}\left[\exp \left(X_{t}\right)\right]\right)+\log \frac{1}{\delta}
$$

## C.4.5 Information theory

We have the following result stating that the Hellinger distance between two probability measures are smaller than the KL divergence between those two probability measures.

Lemma 41 (Hellinger distance smaller than KL divergence) For probability measures $\mathbb{P}$ and $\mathbb{Q}$, the following inequalities hold:

$$
D_{\mathrm{H}}^{2}(\mathbb{P}, \mathbb{Q}) \leq \operatorname{KL}(\mathbb{P}, \mathbb{Q})
$$

We introduce a localized version of Pinsker-type inequality using Hellinger distance (which will be stronger than using the KL divergence).

Lemma 42 (Multiplicative Pinsker-type inequality for Hellinger distance, [105]) Let $\mathbb{P}$ and $\mathbb{Q}$ be probability measures on compact space $\mathcal{X}$. For all $h: \mathcal{X} \rightarrow \mathbb{R}$ with $0 \leq h(X) \leq R$ almost surely under $\mathbb{P}$ and $\mathbb{Q}$, we have

$$
\left|\mathbb{E}_{\mathbb{P}}[h(X)]-\mathbb{E}_{\mathbb{Q}}[h(X)]\right| \leq \sqrt{2 R\left(\mathbb{E}_{\mathbb{P}}[h(X)]+\mathbb{E}_{\mathbb{Q}}[h(X)]\right) \cdot D_{\mathrm{H}}^{2}(\mathbb{P}, \mathbb{Q})} .
$$

We introduce a standard one-sided bound using KL divergence. Compared with Lemma 42, the upper bound in Lemma 43 only depends on the probability measure $q$, while the bound is one-sided and it does not take the square-root from as in Lemma 42.

Lemma 43 (Drifted error bound using KL divergence) For any $p, q \in \Delta(\Pi), \eta>0$, and any vector $y \in \mathbb{R}^{\Pi}$ where $y(\pi) \leq 1 / \eta$ for all $\pi \in \Pi$, we have

$$
\langle y, p-q\rangle-\frac{1}{\eta} \operatorname{KL}(p, q) \leq \eta \sum_{\pi \in \Pi} q(\pi) y(\pi)^{2} .
$$

Proof of Lemma 43: consider the KL divergence $\psi_{q, \eta}(p)=\frac{1}{\eta} \mathrm{KL}(p \| q)$, it is known that the convex conjugate duality of $\psi_{q}$ is the log partition function

$$
\begin{align*}
\psi_{q, \eta}^{*}(y) & :=\sup _{p \in \Delta(\Pi)}\left\{\langle y, p\rangle-\frac{1}{\eta} \operatorname{KL}(p \| q)\right\} \\
& =\frac{1}{\eta} \log \left(\sum_{\pi \in \Pi} q(\pi) \exp (\eta y(\pi))\right) . \tag{C.4.1}
\end{align*}
$$

We have

$$
\begin{array}{r}
\langle y, p\rangle-\frac{1}{\eta} \mathrm{KL}(p \| q) \\
\leq \frac{1}{\eta} \log \left(\sum_{\pi \in \Pi} q \exp (\eta y(\pi))\right) \\
\leq \frac{1}{\eta} \log \left(\sum_{\pi} q(\pi)\left(1+\eta y(\pi)+\eta^{2} y(\pi)^{2}\right)\right) \\
=\frac{1}{\eta} \log \left(1+\eta\langle y, q\rangle+\eta^{2} \sum_{\pi \in \Pi} q(\pi) y(\pi)^{2}\right) \\
\leq\langle y, q\rangle+\eta \sum_{\pi \in \Pi} q(\pi) y(\pi)^{2}, \tag{C.4.2}
\end{array}
$$

where the first equation is because of (C.4.1); the second inequality is because $e^{z} \leq 1+z+z^{2}$ for all $z \leq 1$ and the last inequality is due to $\log (1+z) \leq z$ for all $z \in \mathbb{R}$. Therefore we have

$$
\langle y, p-q\rangle-\frac{1}{\eta} \operatorname{KL}(p \| q) \leq \eta \sum_{\pi \in \Pi} q(\pi) y(\pi)^{2}
$$

for all $y \in \mathbb{R}^{|\Pi|}$ where $y(\pi) \leq 1 / \eta$ for all $\pi$.


[^0]:    ${ }^{1}$ Here and in what follows we will assume that such suprema are measurable, namely, the required regularity conditions on the underlying function classes are met (see, e.g., Appendix C in [14], Section 1.7 in [15]).

[^1]:    ${ }^{1}$ we adopt non-asymptotic big-oh notation: for functions $h_{1}, h_{2}, h_{1}=O\left(h_{2}\right)$ if there exists constant $C>$ 0 such that $h_{1}$ is dominated by $C h_{2}$ with high probability (omitting $\log \frac{1}{\delta}$ factors); $h_{1}=\tilde{O}\left(h_{2}\right)$ if $h_{1}=$ $O\left(h_{2} \max \left\{1, \operatorname{polylog}\left(h_{2}\right)\right\}\right)$.

[^2]:    ${ }^{2}$ Lemma 9 in [89] holds for an arbitrary compact set $\mathcal{A} \subset \mathbb{R}^{d}$, as changing the coordinate system is without the loss of generality for this lemma.
    ${ }^{3}$ in formulation (2.4.2) we have asked $\varphi$ to preserve compactness with respect to $a$ (e.g. the continuous ones), so

[^3]:    such barycentric spanner must exists.

[^4]:    ${ }^{1}$ For technical reason we use the Hellinger distance to define $I R_{H}$ (instead of KL as in the definition (3.2.1) of $I R)$. There is no difference between the definitions of $I R$ and $I R_{H}$ in practical applications since all currently known bounds on the information ratio hold for the stronger definition $I R_{H}$.

