Contextual Decision-making under Parametric Uncertainty and Data-driven Optimistic Optimization

Junyu Cao and Rui Gao
Department of Information, Risk and Operations Management, University of Texas at Austin
junyu.cao, rui.gao@mccombs.utexas.edu

We consider decision-making problems with contextual information, in which the reward function involves uncertain parameters that can be predicted using covariates. To quantify the uncertainty of the reward, we propose a new parameter uncertainty set based on a supervised learning oracle. We show that the worst/best-case reward over the proposed parameter uncertainty set serves as a confidence bound on the reward by sizing the uncertainty set properly. Based on these results, we develop performance guarantees for robust contextual optimization in the offline setting, and propose data-driven optimistic optimization as a systematic tool for online contextual decision-making with provable performance guarantees.

Key words: contextual optimization; joint learning and optimization; data-driven optimization; robust optimization; optimism principle

1. Introduction

Decision-making models in prescriptive analytics often involve uncertain parameters that are not known exactly before the decisions are made. With the increasing availability of covariate data (a.k.a. contextual information), the uncertainty of these parameters can be reduced using predictive analytical tools and thereby enables better decision-making. For example, with the help of advanced machine learning models, the decision-maker for a service system can use customer profile data to predict the service time, so as to achieve better management of the system. There is a recent emerging interest in *contextual optimization*—that is, developing frameworks that jointly estimate the uncertain parameters and optimize the decisions—in both optimization community [15, 29] and machine learning community [57], as well as in applied domains such as finance [10], inventory management [12, 62], revenue management [26], transportation [61], etc.

In contextual optimization, it is crucial to understand how the uncertainty from the prediction model influences the downstream decision optimization problem. For the sake of presentation, we will consider reward maximization as the decision objective in the sequel. To quantify the uncertainty of the reward associated with a decision, one useful approach in the literature is to formulate an optimization problem that computes the worst/best-case reward when the uncertain parameters vary within some uncertainty set. For example, in robust optimization [14, 16], one finds the worst-case reward over uncertain parameters belonging in an ellipsoid that reflects the mean and covariance of the parameters; and in the optimism-in-the-face-of-uncertainty algorithm for linear bandits [1], one finds the best-case reward over an ellipsoidal confidence set. When the uncertainty set is chosen to contain the true parameter with high confidence, such an optimization-based framework is able to provide confidence bounds for any reward function so long as the optimization problem can be solved efficiently. In this case, a confidence interval of the reward can be constructed by solving the worst- and best-case problems with a properly chosen uncertainty set.

The crux of the above optimization-based uncertainty quantification framework is to design a parameter uncertainty set with nice computational tractability and provable statistical guarantees. In the offline setting where there is a given fixed dataset for prediction, a majority of works are developed in the robust optimization literature [90, 99, 46, 63, 84]. Since many of them have been focusing on the tractability of the problem, the statistical guarantees are either not provided or sub-optimal for parametric uncertainty. In the online setting where the design points are chosen sequentially and

adaptively to the historical observations, the parameter uncertainty sets in existing works have been mostly focusing on linear and generalized linear models [1, 8, 24, 80, 33, 58], and the response variable in the prediction model is assumed to be identical to the reward.

In this paper, we propose a new parameter uncertainty set for contextual optimization which is flexible to incorporate various prediction models, and derive its tractable approximation with provable performance guarantees in both offline and online settings. More specifically,

(I) We propose a novel supervised-learning-oracle-based parameter uncertainty set designed for contextual optimization (Section 2). It is centered at a nominal parameter value estimated by minimizing a statistical loss function dependent on decisions, observed contexts, and their associated responses; and it contains all parameters that satisfy some first-order optimality condition perturbed from the optimality condition for the nominal estimator. Our new uncertainty set is flexible to cover a large class of M-estimators in supervised learning. Particularly, for contextual linear optimization, when the parameters are estimated via least squares, our formulation recovers the classic ellipsoidal uncertainty set defined by the asymptotic covariance matrix of the least-square estimator.

(II) Using the new proposed uncertainty set, we derive a computationally efficient approximation to compute the worst/best-case reward based on first-order expansions (Section 2.4). Based on covering number and martingale arguments, for the fixed design and the adaptive design respectively, we show that one can size the uncertainty set properly so that it covers the true parameter value with high probability, thereby the worst/best-case reward provides a lower/upper confidence bound on the true reward (Section 3).

(III) Based on the results above, we develop a robust optimization framework for offline contextual decision-making with general prediction models and decision objectives with nearly optimal statistical guarantees. For the online contextual decision-making, we propose *data-driven optimistic optimization* (DOO) which computes the best-case reward at each round of the adaptive design (Section 4). Particularly, for the linear reward with the least-squares estimator, DOO is reduced to the classic optimism-in-the-face-of-uncertainty algorithm for linear bandits (e.g., Abbasi-Yadkori et al. [1]). We establish its performance guarantees by developing regret bounds for a wide class of contextual decision-making problems. As an illustration, we apply our results to two problems: (i) an optimal pricing problem for queuing service, and (ii) a rank-one matrix estimation problem.

Related Literature

Uncertainty Set for Robust Contextual Optimization. In the optimization community, there is a recent emerging interest in developing frameworks that integrate prediction and decision optimization [62, 64, 12, 15, 29, 28, 83, 31, 27, 38, 39, 45, 30], many of which builds upon ideas from robust optimization [90, 92, 99, 17, 88, 46, 63, 84], along with other principles.

Among these diverse uncertainty sets that have been proposed in the literature, the closest one to ours is the Estimate Uncertainty Set proposed in Zhu et al. [99, Definition 1], which contains parameters that yields a similar statistical loss—in other words, satisfies a similar zero-order optimality condition—as the nominal parameter value. From this aspect, both our work and theirs consider parameter uncertainty sets described in terms of the optimality condition of the prediction problem. In addition, as pointed out by Loke et al. [63], the joint estimation and robustness optimization model in Zhu et al. [99] is closely related to operational cost regularization proposed by Tulabandhula and Rudin [90], and can be cast as a special case of the decision-driven regularization proposed by Loke et al. [63]. Comparing to these works, we consider a somewhat different way to incorporate the residual uncertainty in the prediction model and thus derive different forms of uncertainty sets than theirs even in the simple case of least-square estimators. Moreover, we do not require the convexity of the statistical loss with respect to parameters without sacrificing the performance much.

Another related uncertainty set is the Estimator Uncertainty and Residual Ambiguity Set proposed in a recent work by Sim et al. [84]. Both their work and ours consider a joint uncertainty set of parameters and residuals, and in particular, their strict exogeneity condition of residuals shares a similar spirit as the first-order optimality condition in our formulation. However, their construction is limited to linear regression and, to our understanding, not straightforward to be extended to general prediction models. In addition, Kannan et al. [46] also considers uncertainty set induced from the residuals and the nominal estimator, but they consider a different class of decision objectives and the uncertainty sets are defined for responses rather than parameters in the prediction model. The statistical guarantee in this work is for non-parametric models and builds on the concentration inequality for empirical Wasserstein distance, thus suffers from the curse of dimensionality. Our result in Section 3.1 provides the first computationally tractable robust contextual decision-making framework for general reward functions and prediction models with optimal statistical guarantees.

Optimistic Optimization. In the literature on robust optimization, the optimistic counterpart was studied as the dual of robust optimization in robust linear optimization [13], robust vector optimization [19, 6] and distributionally robust optimization [97]. Norton et al. [72] studied optimistic robust optimization and made a connection with non-convex regularization in machine learning; and distributionally optimistic optimization has been studied recently in [69, 70, 71, 36, 42]. All these literature do not consider the parameter uncertainty involved in contextual optimization and focus on offline settings only.

In online settings, data stream arrives sequentially and the decision-maker can learn more about the uncertainty aided by new data. To balance between earning (reward in each round) and learning (information about the uncertainty), the principle of optimism in the face of uncertainty [67] is a well-known heuristic that has profound impact in online optimization, ranging from bandits learning [57] and reinforcement learning [68, 87] to Monte Carlo tree search [51] and experimental design [60]. Essentially, the optimism principle recommends a decision as if the unknown model parameters assume their best possible values in accordance with the historical data. Hierarchical optimistic optimization was systematically investigated in the monograph Munos et al. [67] that designed algorithms for function optimization in metric spaces, trees, graphs, etc. Our proposed data-driven optimistic optimization (DOO) follows the optimism principle, which will be reviewed in detail in the next paragraph.

Contextual-Bandits-Based Learning. Our formulation (DOO) is related to upper-confidence-bound (UCB) algorithms for contextual bandits when the response variable in the prediction model is the reward. A fair amount of works have been developed for linear bandits [9, 24, 80, 58, 23, 4, 5] and generalized linear models (GLM) [33, 59, 53]. When the reward function is linear, (DOO) is equivalent to the optimism-in-the-face-of-uncertainty algorithm for linear bandits (Abbasi-Yadkori et al. [1]; see also UCB for linear bandits [9, 24, 80]); and when the reward function is a generalized linear model, (DOO) is closely related to UCB for GLM [33, 59], for which we improve the regret bound by a constant related to the shape of the link function (Appendix A). Other parametric classes have also been recently studied, such as kernelized contextual bandits [91], low rank bandits [43, 66], and neural contextual bandits [95, 98]. While the regret analysis of these works relies essentially on some (generalized) linear structure in their considered problems, we do not have hidden linear structure in our model, and our results contribute to these literature by developing a new class of UCB algorithms for general parametric contextual bandits that achieves both computational efficiency and provable nearly-optimal regret in terms of the dependence on the time horizon.

There are also some recent works devoted in developing algorithms for general reward functions, based on UCB [81, 34, 94] and other algorithms [81, 35, 85]. Among them, Russo and Van Roy [81] derives a regret bound that scales linearly with the cardinality of contexts in general (see Appendix C.1 in their paper); Foster et al. [34] imposes strong assumptions on data distribution. The main difference between our result and the regret bounds developed in Foster and Rakhlin [35], Simchi-Levi and Xu [85], Xu and Zeevi [94] is that our bound does not depend on the number of decisions. Finally, there

are some works for non-parametric contextual bandits [78, 75, 86, 40] whose scaling of the regret bounds with respect to the time horizon depends on the covariate dimension, while in our parametric case the dependence is square-root and thus dimension-independent.

In addition to the major differences as above-mentioned, compared with contextual bandits, our formulation allows the observed response in the prediction model to be different than the reward. Such difference is also seen in assortment optimization [20, 79, 3, 41, 22, 73] and in dynamic pricing [18, 26, 77, 50, 11]. In these problems, the reward (such as revenue) is an explicit function of the response (such as demand), but our formulation allows the two functions to be related only implicitly by a common parameter; see Section 5.1 for an example.

The rest of the paper proceeds as follows. In Section 2, we introduce the setup for contextual decision-making model and propose a new parameter uncertainty set for contextual optimization based on a supervised-learning oracle, illustrated by the least-square estimate. In Section 3, we discuss how to choose the size of the uncertainty set so that it covers the true parameter with high probability, for offline and online settings, respectively. In Section 4, we propose the DOO formulation for sequential contextual decision-making and analyze its performance guarantees. In Section 5, we demonstrate our framework with two examples. Proofs and auxiliary results are deferred to the Appendix.

2. A New Uncertainty Set for Contextual Decision-making

In this section, we describe the decision-making problem with contextual information in Section 2.1, and propose a new uncertainty set for contextual decision-making based on a supervised-learning oracle in Section 2.2. We define our optimization-based uncertainty quantification model in Section 2.3, and develop a computationally-efficient approximation in Section 2.4.

2.1. Contextual Decision-making

Consider the following decision-making problem with contextual information. Prior to decision-making, the decision-maker observes a context $X = x$ from a set $\mathcal{X}$. Given this context, the conditional expected reward of a decision $A = a$ is parameterized as

$$\mathbb{E}[R|X = x, A = a] = r(x, a; \theta^*),$$

where $\theta^* \in \mathbb{R}^d$ is an unknown model parameter that can only be estimated from historical observations. We are interested in the following conditional reward maximization problem

$$\max_{a \in \mathcal{A}} \; r(x, a; \theta^*),$$

which finds the optimal decision from a decision set $\mathcal{A}$ to maximize the expected reward under a given context $X = x$.

Each historical observation is represented by a triple (Y, X, A), where $Y \in \mathcal{Y} \subset \mathbb{R}$ is the response variable. The statistical relationship between the response variable Y and the context-decision pair (X, A) is modeled as

$$Y = f_{\theta^*}(X, A) + \epsilon,$$

where ϵ represents the zero-mean noise. Note that the response Y and the reward R may be either identical, or related through θ^*. For example, in dynamic pricing, Y represents the product demand dependent on product features X and price level A in a parametric fashion with θ^*, while the objective is the total revenue R, which depends on both price and demand, thereby R is an (explicit) function of the demand Y that also involves the parameter θ^*. In Section 5.1, we provide an example where R and Y only relate to each other implicitly with a shared parameter θ^*. Given $n-1$ observations

$\{(Y_i, X_i, A_i)\}_{i=1}^{n-1}$, suppose the model parameter θ^* can be estimated by minimizing a statistical loss function $\ell : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$:

$$\hat{\theta}_n \in \arg\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n-1} \ell(Y_i, f_\theta(X_i, A_i)).$$

2.2. Supervised-Learning-Oracle-Based Parameter Uncertainty Set

In this subsection, we introduce a new parameter uncertainty set based on a supervised learning oracle, and define the associated optimistic optimization problem.

For simplicity, we denote

$$\psi(y, x, a, \varepsilon; \theta) := \nabla_\theta \ell(y, f_\theta(x, a) + \varepsilon) \in \mathbb{R}^d.$$

Assume ℓ is a differentiable function satisfying $\ell(y, y) = 0$ for all $y \in \mathcal{Y} \subset \mathbb{R}$. Let us consider a supervised learning oracle that outputs a root of the following equation (also known as M-estimators of ψ-type):

$$\sum_{i=1}^{n-1} \psi(Y_i, X_i, A_i, 0; \theta) = 0. \tag{Oracle}$$

By our assumption on ℓ we have $\ell(Y_i, f_{\theta^*}(X_i, A_i) + \epsilon_i) = 0$, hence θ^* satisfies

$$\sum_{i=1}^{n-1} \psi(Y_i, X_i, A_i, \epsilon_i; \theta^*) = 0.$$

Let $\hat{\theta}_n$ be a root to (Oracle), namely, $\hat{\theta}_n$ satisfies

$$\sum_{i=1}^{n-1} \psi(Y_i, X_i, A_i, 0; \hat{\theta}_n) = 0.$$

In view of the two equations above, we introduce the following uncertainty set involving the parameter θ and the noise ε:

$$\left\{ (\theta, \varepsilon) \in \mathbb{R}^d \times \mathbb{R}^{n-1} : \sum_{i=1}^{n-1} \psi(Y_i, X_i, A_i, \varepsilon_i; \theta) = 0, \ \sum_{i=1}^{n-1} \varepsilon_i^2 / \omega_{n,i}^2 \le \rho_n^2 \right\}.$$

Here, the projection on the ε-component $\{\varepsilon \in \mathbb{R}^{n-1} : \sum_{i=1}^{n-1} \varepsilon_i^2/\omega_{n,i}^2 \le \rho_n^2\}$ is an ellipsoidal uncertainty set for the noises $\{\varepsilon_i\}_{i=1}^{n-1}$, where $\omega_{n,i} > 0$ indicates the (estimated) standard deviation of ε_i. This captures the potential heterogeneity among the historical observations. We do not consider their covariance as the true noises $\{\epsilon_i\}_{i=1}^{n-1}$ are often independent of each other. Note that we use ϵ to denote the noise in the true model and use ε to denote the variable in the uncertainty set. When $\varepsilon = 0$, by definition $(\hat{\theta}_n, 0)$ is feasible, hence the projection of the uncertainty set on the θ-component contains $\hat{\theta}_n$. If ρ_n satisfies $\rho_n^2 \ge \sum_{i=1}^{n-1} \epsilon_i^2/\omega_{n,i}^2$, then (θ^*, ϵ) is feasible. Therefore, with proper selection of the radius ρ_n, the induced parameter uncertainty set of θ:

$$\left\{ \theta \in \mathbb{R}^d : \ \exists \varepsilon \in \mathbb{R}^{n-1} \ s.t. \ \sum_{i=1}^{n-1} \psi(Y_i, X_i, A_i, \varepsilon_i; \theta) = 0, \ \sum_{i=1}^{n-1} \varepsilon_i^2 / \omega_{n,i}^2 \le \rho_n^2 \right\} \tag{S_o}$$

can be viewed as a confidence region for θ^*. As such, the derivation above suggests a principled way to construct the confidence region of θ^* by virtue of an auxiliary variable ε representing the noise, which is novel in the literature to the best of our knowledge.

2.3. Worst/Best-case Reward

With the new uncertainty set (S_o), we consider the following data-driven robust/optimistic optimization problem

$$\begin{aligned} \min/\max_{\theta\in\mathbb{R}^d,\varepsilon\in\mathbb{R}^{n-1}} \quad & r(x,a;\theta) \\ s.t. \quad & \sum_{i=1}^{n-1}\psi(Y_i,X_i,A_i,\varepsilon;\theta)=0, \\ & \sum_{i=1}^{n-1}\varepsilon_i^2/\omega_{n,i}^2\le\rho_n^2, \end{aligned} \tag{UQ_o}$$

which provide a conservative/optimistic estimate for the reward associated with a given decision $a\in\mathcal{A}$.

Illustration for least-square estimates. As a simple illustration, consider a linear reward $r(x,a;\theta)=a^\top\theta$ and (Oracle) to be the least-square estimate for linear regression

$$\sum_{i=1}^{n-1}2(Y_i-A_i^\top\theta)A_i=0.$$

In this case, the maximization problem in (UQ_o) with $\omega_{n,i}=1$ becomes

$$\begin{aligned} &\max_{\theta\in\mathbb{R}^d,\varepsilon\in\mathbb{R}^{n-1}}\left\{a^\top\theta:\ \sum_{i=1}^{n-1}2(Y_i-\varepsilon_i-A_i^\top\theta)A_i=0,\ \sum_{i=1}^{n-1}\varepsilon_i^2\le\rho_n^2\right\} \\ &=\max_{\theta\in\mathbb{R}^d}\left\{a^\top\theta+\min_{\alpha\ge0,\beta\in\mathbb{R}^d}\max_{\varepsilon\in\mathbb{R}^{n-1}}\left\{\alpha\rho_n^2+\sum_{i=1}^{n-1}2(Y_i-A_i^\top\theta)\beta^\top A_i-\sum_{i=1}^{n-1}(2\beta^\top A_i\varepsilon_i+\alpha\varepsilon_i^2)\right\}\right\} \\ &=\max_{\theta\in\mathbb{R}^d}\left\{a^\top\theta+\min_{\alpha\ge0,\beta\in\mathbb{R}^d}\left\{\alpha\rho_n^2+\sum_{i=1}^{n-1}2(Y_i-A_i^\top\theta)\beta^\top A_i+\frac{1}{\alpha}\sum_{i=1}^{n-1}(\beta^\top A_i)^2\right\}\right\} \\ &=\max_{\theta\in\mathbb{R}^d}\left\{a^\top\theta+\min_{\alpha\ge0}\left\{\alpha\rho_n^2-\alpha\Big(\sum_{i=1}^{n-1}(Y_i-A_i^\top\theta)A_i\Big)^\top\Big(\sum_{i=1}^{n-1}A_iA_i^\top\Big)^{-1}\Big(\sum_{i=1}^{n-1}(Y_i-A_i^\top\theta)A_i\Big)\right\}\right\}. \end{aligned}$$

Observe that

$$\sum_{i=1}^{n-1}(Y_i-A_i^\top\theta)A_i=\sum_{i=1}^{n-1}(\hat\theta_n-\theta)^\top A_iA_i.$$

It follows that the problem above is equivalent to

$$\begin{aligned} &\max_{\theta\in\mathbb{R}^d}\left\{a^\top\theta+\min_{\alpha\ge0}\left\{\alpha\rho_n^2-\alpha(\hat\theta_n-\theta)^\top\Big(\sum_{i=1}^{n-1}A_iA_i^\top\Big)(\hat\theta_n-\theta)\right\}\right\} \\ &=\max_{\theta\in\mathbb{R}^d}\left\{a^\top\theta:\ \|\hat\theta_n-\theta\|_{V_n}\le\rho_n\right\},\qquad \text{where } V_n=\sum_{i=1}^{n-1}A_iA_i^\top, \end{aligned} \tag{1}$$

where $\|\vartheta\|_{V_n}=\sqrt{\vartheta^\top V_n\vartheta}$. Whenever V_n is invertible, the best-case reward equals

$$U:=a^\top\hat\theta_n+\rho_n\|a\|_{V_n^{-1}}. \tag{2}$$

Similarly, the worst-case reward would be equal to $a^\top\hat\theta_n-\rho_n\|a\|_{V_n^{-1}}$.

Let us compare the above result with the JERO framework proposed in Zhu et al. [99]. Set

$$\hat\ell:=\sum_{i=1}^{n-1}(Y_i-\hat\theta_n^\top A_i)^2,\quad A_{[n]}:=[A_1,\ldots,A_{n-1}].$$

Then the support function defined in their Propositions 3 essentially computes the best-case reward

$$\sup_{\theta\in\mathbb{R}^d}\left\{a^\top\theta:\ \sum_{i=1}^{n-1}(Y_i-\theta^\top A_i)^2\le\hat{\ell}+\rho_n\right\},$$

where ρ_n is interpreted as the size of the Estimate Uncertainty Set defined in their Definition 1. By their Proposition 4, this is equivalent to

$$U_{\mathsf{JERO}}:=\min_{w\in\mathbb{R}^{n-1}}\left\{w^\top\left(\sum_{i=1}^{n-1}Y_i\right)+\sqrt{\hat{\ell}+\rho_n}\cdot\|w\|_2:\ A_{[n]}^\top w=a\right\}.$$

Suppose $A_{[n]}$ has full column rank, which implies the invertibility of V_n. Then solving the linear system in the constraint above yields that

$$U_{\mathsf{JERO}}=\min_{w\in\mathbb{R}^{n-1}}\left\{w^\top\left(\sum_{i=1}^{n-1}Y_i\right)+\sqrt{\hat{\ell}+\rho_n}\cdot\|w\|_2:\ w=A^g a+(I-A^gA)w\right\},$$

where $A^g=(A_{[n]}A_{[n]}^\top)^{-1}A_{[n]}$. Observe that U_{JERO} is not equivalent to U defined in (2) in general. Indeed, suppose $\sum_{i=1}^{n-1}Y_i=0$, in which case U_{JERO} solves the minimum-norm least-square estimate:

$$U_{\mathsf{JERO}}=\sqrt{\hat{\ell}+\rho_n}\cdot\sqrt{a^\top A_{[n]}^\top(A_{[n]}A_{[n]}^\top)^{-2}A_{[n]}a}.$$

In comparison, in this case we have $\theta_n=0$, $V_n=A_{[n]}A_{[n]}^\top$, and U in (2) equals $\rho_n\sqrt{a^\top(A_{[n]}A_{[n]}^\top)^{-1}a}$. Hence U_{JERO} and U are not equal to each other even up to a constant scaling of the radius. Note that in our framework, for linear regression, the induced uncertainty set (1) has exactly the same form of confidence region under fixed and adaptive designs [55, 56]. In Section 4.1, we will exemplify that our proposed uncertainty set is related to the confidence region of θ^* for several problems.

2.4. Computationally Efficient Approximation

The problem ($\mathsf{UQ_o}$) may not have a tractable exact solution in general due to the potential non-convexity of the first-order optimality constraint. In this section, we derive a tractable reformulation through a sequence of first-order Taylor approximations.

To ease the notation, we denote by $\psi'(y,x,a,\varepsilon;\theta):=\frac{\partial\psi}{\partial\varepsilon}(y,x,a,\varepsilon;\theta)\in\mathbb{R}^d$ the derivative of the vector-valued function ψ with respect to ε. Replacing the first constraint in ($\mathsf{UQ_o}$) with its first-order approximation with respect to ε gives

$$\begin{aligned}\min/\max_{\theta\in\mathbb{R}^d,\varepsilon\in\mathbb{R}^{n-1}}\quad & r(x,a;\theta)\\ s.t.\quad & \sum_{i=1}^{n-1}\psi(Y_i,X_i,A_i,0;\theta)-\varepsilon_i\psi'(Y_i,X_i,A_i,0;\theta)=0,\\ & \sum_{i=1}^{n-1}\varepsilon_i^2/\omega_{n,i}^2\le\rho_n^2.\end{aligned}\tag{3}$$

For notational brevity, we define

$$\zeta_n(\theta):=\sum_{i=1}^{n-1}\psi(Y_i,X_i,A_i,0;\theta),\quad\text{and}\quad V_n(\theta):=\sum_{i=1}^{n-1}\omega_{n,i}^2\psi'(Y_i,X_i,A_i,0;\theta)\psi'(Y_i,X_i,A_i,0;\theta)^\top.\tag{4}$$

The following result provides an exact reformulation of (3), whose proof is given in Appendix EC.1.

LEMMA 1. *Assume $V_n(\theta)$ is invertible. Formulation (3) can be equivalently reformulated as*

$$\begin{aligned} \min/\max_{\theta\in\mathbb{R}^d} \quad & r(x,a;\theta) \\ s.t. \quad & \zeta_n(\theta)^\top V_n(\theta)^{-1}\zeta_n(\theta) \le \rho_n^2. \end{aligned} \tag{5}$$

Furthermore, the gap between the optimal values of ($\mathrm{UQ_o}$) *and* (5) *are bounded by $C\rho_n^2$ for some constant C that is independent of ρ_n.*

We remark that $\hat\theta_n$ is always a feasible solution to (5) since $\zeta_n(\hat\theta_n)=0$. In general, the formulation above may still be computationally challenging since it involves a possibly non-convex constraint. To obtain a formulation that is amenable to efficient computation, we propose a further approximation as follows. To ease the notation, we define

$$\hat V_n^\omega = \sum_{i=1}^{n-1} \omega_{n,i}^2 \psi'(Y_i,X_i,A_i,0;\hat\theta_n)\psi'(Y_i,X_i,A_i,0;\hat\theta_n)^\top. \tag{6}$$

By definition of $\zeta_n(\theta)$ and the constraint in (3), it holds that

$$\zeta_n(\theta) = \sum_{i=1}^{n-1}\epsilon_i\psi'(Y_i,X_i,A_i,0;\theta) = \sum_{i=1}^{n-1}\psi'(Y_i,X_i,A_i,0;\theta)(f_{\theta^*}(X_i,A_i) - f_{\hat\theta_n}(X_i,A_i)).$$

When $\hat\theta_n$ is close to θ^* and θ, we have the approximation

$$\begin{aligned} \zeta_n(\theta) &\simeq \sum_{i=1}^{n-1}\psi'(Y_i,X_i,A_i,0;\hat\theta_n)(f_\theta(X_i,A_i) - f_{\hat\theta_n}(X_i,A_i)) \\ &\simeq \sum_{i=1}^{n-1}\psi'(Y_i,X_i,A_i,0;\hat\theta_n)^\top(\theta-\hat\theta_n)\nabla f_{\hat\theta_n}(X_i,A_i) =: \sum_{i=1}^{n-1}\langle\psi'_{\hat\theta_n,i},\theta-\hat\theta_n\rangle\nabla f_{\hat\theta_n,i}. \end{aligned}$$

Under such approximation, it holds that

$$\begin{aligned} \zeta_n(\theta)^\top V_n(\theta)^{-1}\zeta_n(\theta) &\simeq (\theta-\hat\theta_n)^\top\Big(\sum_{i=1}^{n-1}\psi'_{\hat\theta_n,i}\nabla f_{\hat\theta_n,i}^\top\Big)\hat V_n^{\omega\,-1}\Big(\sum_{i=1}^{n-1}\psi'_{\hat\theta_n,i}\nabla f_{\hat\theta_n,i}^\top\Big)(\theta-\hat\theta_n) \\ &=: \|\theta-\hat\theta_n\|_{\bar V_n}^2, \end{aligned} \tag{7}$$

where

$$\bar V_n := \Big(\sum_{i=1}^{n-1}\psi'_{\hat\theta_n,i}\nabla f_{\hat\theta_n,i}^\top\Big)\hat V_n^{\omega\,-1}\Big(\sum_{i=1}^{n-1}\psi'_{\hat\theta_n,i}\nabla f_{\hat\theta_n,i}^\top\Big).$$

Define an uncertainty set

$$\Theta_n := \{\theta\in\mathbb{R}^d :\ \|\theta-\hat\theta_n\|_{\bar V_n}\le\rho_n\}. \tag{S}$$

The derivation above leads to the following approximation of ($\mathrm{UQ_o}$):

$$\min/\max_{\theta\in\Theta_n} r(x,a;\theta). \tag{UQ}$$

Moreover, whenever r is convex (resp. concave) in θ, then the min (resp. max) problem above can be evaluated efficiently using convex optimization. Nonetheless, when the reward r_n is a general function, the problem above can be intractable. Replacing the reward function by its first-order Taylor expansion and assuming $\bar V_n$ is invertible, we have

$$\min/\max_{\theta\in\Theta_n}\ \big\{r(x,a;\hat\theta_n) + (\theta-\hat\theta_n)^\top\nabla r(x,a;\hat\theta_n)\big\} \ =\ r(x,a;\hat\theta_n) \pm \rho_n\cdot\|\nabla r(x,a;\hat\theta_n)\|_{\bar V_n^{-1}}, \tag{$\widehat{\mathrm{UQ}}$}$$

where ∇ denotes the derivative with respect to θ, and the second equality holds by Lemma 2 below, which controls the gap between (UQ) and ($\widehat{\mathrm{UQ}}$). The proof is in Appendix EC.1.

LEMMA 2. *Assume $\bar{V}_n$ is invertible. Let $a \in \mathcal{A}$ and assume $\sup_{\theta\in\Theta} \|\nabla^2 r(x,a;\theta)\|_2 \le \hbar_r$. Then the optimal values of* (UQ) *and* ($\widehat{\text{UQ}}$) *differ by at most $\hbar_r \|\tilde{\theta}_n - \hat{\theta}_n\|_2^2$, where $\tilde{\theta}_n$ is an optimal solution to* (UQ).

The uncertainty set (S) falls into the category of ellipsoidal uncertainty sets in classic robust optimization, but here the ellipse defined by the matrix $\bar{V}_n$ is carefully chosen, tailored to the generic prediction model (Oracle). When the radius ρ_n is chosen properly, the optimal values of (UQ) and ($\widehat{\text{UQ}}$) can serve as an (approximately) confidence bound for general reward functions and prediction models. We will provide a statistical analysis in Section 3 that rigorously justifies (UQ) and ($\widehat{\text{UQ}}$) as approximations of (UQ_o).

3. Radius Selection and Uncertainty Quantification of the Reward

In this section, we derive our main result on how to choose the radius ρ_n of the uncertainty set Θ_n in (S) to ensure $\theta^* \in \Theta_n$ with high probability, thereby we can quantify the uncertainty of the reward by computing the worst/best-case reward using (UQ) or ($\widehat{\text{UQ}}$).

In the sequel, with slightly abuse of notations, we consider the prediction model of the form $Y = \phi \circ f_\theta(X, A) + \epsilon$, where $\phi : \mathbb{R} \to \mathbb{R}$ is a link function that accommodates prediction models like the generalized linear model. All results in Section 2 remain to hold by replacing f_θ with $\phi \circ f_\theta$. We will focus on (Oracle) of the form

$$\sum_{i=1}^{n-1} \big(Y_i - \phi \circ f_\theta(X_i, A_i)\big) \nabla f_\theta(X_i, A_i) = 0, \tag{8}$$

and instances of this oracle will be provided in Section 4.1, including the least-squares for linear regression and the quasi-likelihood estimation for generalized linear models and beyond. We assume that there exists a compact set Θ such that it contains the solution $\hat{\theta}_n$ to (8) for all n. To ease the exposition, we set $\omega_{n,i} = 1$ in (UQ_o) (and thereby (6) with $\omega_{n,i} = 1$), under which it holds that $\psi(y, x, a, \varepsilon; \theta) = (\phi \circ f_\theta(x, a) + \varepsilon - y)\nabla f_\theta(x, a)$ and $\psi'(Y_i, X_i, A_i, 0; \hat{\theta}_n) = \nabla f_{\hat{\theta}_n}(X_i, A_i)$, thereby in ($\widehat{\text{UQ}}$) we have

$$\bar{V}_n = \hat{V}_n := \sum_{i=1}^{n-1} \nabla f_{\hat{\theta}_n}(X_i, A_i) \nabla f_{\hat{\theta}_n}(X_i, A_i)^\top. \tag{9}$$

The results still hold similarly if $\{\omega_{n,i}\}_{n,i}$ has a uniform strictly positive lower and upper bounds; see Remark 1 and Appendix A. Under such setting, to ensure $\theta^* \in \Theta_n$ with high probability, it amounts to providing a high-probability bound on $\|\theta^* - \hat{\theta}_n\|_{\hat{V}_n}$.

In what follows, we consider two settings: offline setting where the uncertain parameters are estimated based on a fixed dataset, and online setting where the uncertain parameters are estimated based on adaptively chosen decisions.

3.1. Fixed Design with an Offline Dataset

In the offline setting, the decision-maker is given a fixed set of observations $\{(Y_i, A_i, X_i)\}_{i=1}^{n-1}$, and the goal is to seek a decision to maximize the reward given a new context x:

$$\max_{a\in\mathcal{A}} r(x, a; \theta^*).$$

To bound this probability, we make the following assumptions. The first is a standard sub-Gaussian assumption.

ASSUMPTION 1. *$\{\epsilon_i\}_{i=1}^n$ are independent and there exists $\sigma > 0$ such that for every $i = 1, \ldots, n$ and every $u \in \mathbb{R}$, $\mathbb{E}[\exp(u\epsilon_i)] \le \exp(u^2\sigma^2/2)$.*

To ease the notation, we define $g_\theta : \Theta \times \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$ as

$$g_\theta(x,a) := \nabla_\theta(\phi \circ f_\theta(x,a)). \tag{10}$$

In what follows, we use a shorthand notation ϕ' to denote the derivative of ϕ. The second assumption is on the boundedness.

Assumption 2. *The following holds:*

(i) $\kappa_f := \sup_{1\le i\le n-1,\theta\in\Theta} \|\nabla f_\theta(X_i,A_i)\|_2 < \infty$.
(ii) $\kappa_g := \sup_{1\le i\le n-1,\theta\in\Theta} \|\nabla g_\theta(X_i,A_i)\|_2 < \infty$.
(iii) $\underline{\kappa}_\phi := \inf_{1\le i\le n-1,\theta\in\Theta} \phi'(f_\theta(X_i,A_i)) > 0$.
(iv) $\beta_\Theta := \sup_{\theta\in\Theta} \|\theta\|_2 < \infty$.
(v) $\hbar_r := \sup_{1\le i\le n-1,\theta\in\Theta} \|\nabla^2 r(X_i,A_i;\theta)\|_2 < \infty$.

Let us define

$$\xi_n(\theta) := \sum_{i=1}^{n-1} \epsilon_i \nabla f_\theta(X_i,A_i), \quad \hat{\xi}_n := \xi_n(\hat{\theta}_n).$$

In the remainder of this subsection, we assume $\hat{V}_n$ defined in (9) is invertible. It follows that

$$\begin{aligned}
\|\hat{\xi}_n\|_{\hat{V}_n^{-1}} &= \|\sum_{i=1}^{n-1}(Y_i - \phi\circ f_{\theta^*}(X_i,A_i))\nabla f_{\hat{\theta}_n}(X_i,A_i)\|_{\hat{V}_n^{-1}} \\
&\overset{(8)}{=} \|\sum_{i=1}^{n-1}(\phi\circ f_{\hat{\theta}_n}(X_i,A_i) - \phi\circ f_{\theta^*}(X_i,A_i))\nabla f_{\hat{\theta}_n}(X_i,A_i)\|_{\hat{V}_n^{-1}} \\
&\overset{(10)}{\simeq} \|(\hat{\theta}_n - \theta^*)^\top \sum_{i=1}^{n-1} g_{\hat{\theta}_n}(X_i,A_i)\nabla f_{\hat{\theta}_n}(X_i,A_i)^\top\|_{\hat{V}_n} \\
&\ge \underline{\kappa}_\phi \|\theta^* - \hat{\theta}_n\|_{\hat{V}_n}.
\end{aligned}$$

In light of this, we first bound $\|\hat{\xi}_n\|_{\hat{V}_n^{-1}}$ in the following key lemma.

Lemma 3. *Assume Assumptions 1 and 2(i) are in force. Let $\delta \in (0,1)$. Then with probability at least $1-\delta$, it holds that*

$$\|\hat{\xi}_n\|^2_{\hat{V}_n^{-1}} \le 16\kappa_f^2\sigma^2 \log(5\kappa_f)(d + \log(1/\delta)).$$

We sketch the proof here, and a full proof can be found in Appendix EC.2.1. The sub-Gaussian Assumption 1 implies that for every $x \in \mathbb{R}^d$,

$$\mathbb{P}\left\{\left\langle x, \hat{V}_n^{-1/2}\hat{\xi}_n\right\rangle > c\right\} \le \exp\left(-\frac{c^2}{8\sigma^2\|x\|_2^2}\right).$$

Let $\mathcal{C}_{1/2}$ be a 1/2-covering set of the ball with radius κ_f. One can show that

$$\mathbb{P}\{\|\hat{\xi}_n\|_{\hat{V}_n^{-1}} > c\} \le \sum_{x\in\mathcal{C}_{1/2}} \mathbb{P}\left\{\left\langle x, \hat{V}_n^{-1/2}\hat{\xi}_n\right\rangle > c/2\right\}.$$

Then the result is obtained by taking the union bound.

To bound $\|\theta^* - \hat{\theta}_n\|_{\hat{V}_n}$ using Lemma 3, we impose the following assumption, which is closely related to the notion of one-point convexity in non-convex optimization [47]. Based on Assumption 3, proximity in function values implies proximity in parameters.

Assumption 3. *There exists $\alpha > 0$ such that for every $x \in \mathcal{X}$, $a \in \mathcal{A}$ and $\theta \in \Theta$,*

$$\big(\phi\circ f_\theta(x,a) - \phi\circ f_{\theta^*}(x,a)\big)\nabla f_\theta(x,a)^\top(\theta-\theta^*) \ge \alpha\big(\nabla f_\theta(x,a)^\top(\theta-\theta^*)\big)^2.$$

This assumption generalizes strong convexity to non-convex functions. For example, consider the statistical loss function ℓ to be the square loss $\ell(y, f_\theta(x,a)) = (y - f_\theta(x,a))^2$ or the negative log-likelihood function $\ell(y, f_\theta(x,a)) = -y\log(\phi \circ f_\theta(x,a)) - (1-y)\log(1-\phi \circ f_\theta(x,a))$ where $y \in \{0,1\}$. If for every $x \in \mathcal{X}$, $a \in \mathcal{A}$ and $y \in \mathcal{Y}$, $\ell(y, f_\theta(x,a))$ satisfies the *restricted secant inequality* at θ^*[1] [96], then Assumption 3 holds (see Lemma EC.1 in Appendix EC.2.2). Note that this can occur even when $\ell(y, f_\theta(x,a))$ is nonconvex in θ. To ensure the global optimality of gradient-based algorithms for non-convex optimization, restricted secant inequality is a weak condition, and thus in this respect, Assumption 3 is mild.

With the three assumptions above, we can bound $\|\theta^* - \hat{\theta}_n\|_{\hat{V}_n}$ in the following result, whose proof is given in Appendix EC.2.3.

THEOREM 1 (**RADIUS SELECTION FOR FIXED DESIGN**). *Assume Assumptions 1, 2, 3 are in force. Let $\delta \in (0,1)$. Then with probability at least $1-\delta$,*

$$\|\theta^* - \hat{\theta}_n\|_{\hat{V}_n}^2 \le \frac{16\kappa_f^2\sigma^2\log(5\kappa_f)(d+\log(1/\delta))}{\min(\underline{\kappa}_\phi^2, \alpha^2)}.$$

When the offline dataset is collected from i.i.d. samples, the largest eigenvalue of $\hat{V}_n$ grows at a linear rate in n, thereby Theorem 1 implies $\|\hat{\theta}_n - \theta^*\|_2 = O(\sqrt{d/n})$. Such rate is optimal in terms of both the dimension d and the sample size n, as can be seen from the case of linear regression (see Appendix EC.2.4). An immediate corollary of Theorem 1 and Lemma 2 is that the worst/best-case reward in (UQ) and ($\widehat{\text{UQ}}$) provides simultaneous confidence intervals of the true reward for all decisions.

COROLLARY 1 (**UNCERTAINTY QUANTIFICATION FOR OFFLINE CONTEXTUAL OPTIMIZATION**). *Assume Assumptions 1, 2, 3 are in force. Let $L_n(a)$ (resp. $U_n(a)$) be the optimal value of the minimization (resp. maximization) problem in (UQ), and let $\widehat{L}_n(a)$ (resp. $\widehat{U}_n(a)$) be the corresponding optimal value. Setting*

$$\rho_n = \frac{4\sigma\kappa_f\sqrt{\log(5\kappa_f)(d+\log(1/\delta))}}{\min(\underline{\kappa}_\phi, \alpha)}, \quad \varrho_n = \rho_n/\sqrt{\lambda_{\min}(\hat{V}_n)}. \tag{11}$$

Let $\delta \in (0,1)$. Then with probability at least $1-\delta$, for every $a \in \mathcal{A}$, it holds that

$$\begin{aligned} &r(x,a;\theta_*) \in [L_n(a), U_n(a)], \\ &r(x,a;\theta_*) \in \left[\widehat{L}_n(a) - \hbar_r\varrho_n^2,\ \widehat{U}_n(a) + \hbar_r\varrho_n^2\right]. \end{aligned} \tag{12}$$

The difference $\hbar_r\varrho_n^2$ in the two confidence intervals in (12) holds because $|L_n(a) - \widehat{L}_t(a)|$ and $|U_n(a) - \widehat{U}_t(a)|$ are bounded by $\hbar_r\varrho_n^2$ due to Lemma 2. Note that when the offline dataset is collected from i.i.d. samples, the length of the interval $[\widehat{L}_n(a), \widehat{U}_n(a)]$ is $O(1/\sqrt{n})$, while $\varrho_n^2 = O(1/n)$ since $\lambda_{\min}(\hat{V}_n) = O(n)$. Hence, the dominant term in (12) is $[\widehat{L}_n(a), \widehat{U}_n(a)]$.

REMARK 1. If the weights $\{\omega_{n,i}\}_{n,i}$ are not equal to 1 but have uniformly bounded: $\underline{\omega} \le \omega_{n,i} \le \overline{\omega}$ for all n and i, then $\underline{\omega}^2\hat{V}_n \preceq \hat{V}_n^\omega \preceq \overline{\omega}^2\hat{V}_n$, where $\hat{V}_n^\omega$ is defined in (9). This implies that

$$\|\theta^* - \hat{\theta}_n\|_{\hat{V}_n^\omega}^2 \le \overline{\omega}^2\|\theta^* - \hat{\theta}_n\|_{\hat{V}_n}^2, \quad \text{and} \quad \lambda_{\min}(\hat{V}_n^\omega) \ge \underline{\omega}\lambda_{\min}(\hat{V}_n).$$

Thereby Theorem 1 and Corollary 1 hold similarly.

[1] A function $h : \Theta \to \mathbb{R}$ is said to satisfy restricted secant inequality at θ^* if $\nabla h(\theta)^\top(\theta - \theta^*) \ge \alpha\|\theta - \theta^*\|$ for all $\theta \in \theta$.

3.1.1. Robust Contextual Optimization Define $R^* = \max_{a\in\mathcal{A}} r(x,a;\theta^*)$. Consider the following max-min robust contextual optimization formulation

$$R_{\text{rob}} = \max_{a\in\mathcal{A}} \min_{\theta\in\Theta_n} r(x,a;\theta), \quad \text{where } \Theta_n \text{ is defined in (S).} \tag{RCO}$$

If we are able to derive an upper bound on the probability that $\theta^* \in \Theta_n$, then R_{rob} provides a lower bound on R^*. An immediate consequence of Corollary 1 is the following result.

Corollary 2 **(Performance guarantees for robust contextual optimization).** *Let $\delta \in (0,1)$. Under the setting in Corollary 1, the robust optimal value R_{rob} of (RCO) satisfies $R^* \geq R_{\text{rob}}$ with probability at least $1-\delta$. Moreover, if there exists $\kappa_r > 0$ such that $|r(x,a;\theta) - r(x,a;\theta^*)| \leq \kappa_r \|\theta - \theta^*\|_2$ for all $a \in \mathcal{A}$, then with probability at least $1-\delta$, $R^* \leq R_{\text{rob}} + 2\kappa_r \varrho_n$.*

Define

$$\widehat{R}_{\text{rob}} = \max_{a\in\mathcal{A}} \widehat{L}_n(a).$$

Then by Corollaries 1 and 2, when r is Lipschitz in θ, it holds with high probability that

$$R^* \in [R_{\text{rob}},\, R_{\text{rob}} + 2\kappa_r\varrho_n], \quad R^* \in [\widehat{R}_{\text{rob}} - \hbar_r\varrho_n^2,\, \widehat{R}_{\text{rob}} + 2\kappa_r\varrho_n + \hbar_r\varrho_n^2]. \tag{13}$$

When the offline dataset is collected from i.i.d. samples, the length of this interval is $O(1/\sqrt{n})$. As far as we know, the formulation (RCO) is the first robust contextual optimization framework for general reward functions and prediction models with provable guarantees. Note that the confidence interval $R^* \in [R_{\text{rob}}, R_{\text{rob}} + 2\kappa_r\varrho_n]$ has an essential difference compared to the classical confidence interval based on sample average approximation (SAA). Indeed, let

$$\widehat{R}_n = \max_{a\in\mathcal{A}} r(x,a;\hat{\theta}_n).$$

Then a confidence interval based on SAA typically has a two-sided form $[\widehat{R}_n - \widehat{\rho}_n, \widehat{R}_n + \widehat{\rho}_n]$ with some half-length $\widehat{\rho}_n = O(1/\sqrt{n})$ for i.i.d. samples. In contrast, (13) is one-sided and the robust optimal value R_{rob} by itself serves as a lower bound on the true optimal reward R^*.

3.2. Adaptive Design in an Online Setting

In the online setting, the decision is chosen sequentially and adaptively. Suppose there are N rounds. At the beginning of round $n = 1,\ldots,N$, the decision-maker observes a context X_n from a set $\mathcal{X}_n$, then chooses a decision A_n from a set $\mathcal{A}_n$, observes a response Y_n and receives a reward R_n. Note that unlike the offline setting, here we allow $\mathcal{X}_n$, $\mathcal{A}_n$ to be random and r_n to be time-dependent. Suppose the conditional expected reward is parameterized as

$$\mathbb{E}[R_n | X_n = x, A_n = a] = r_n(x,a;\theta^*),$$

where $\theta^* \in \mathbb{R}^d$ is an unknown parameter that can only be estimated from historical observations. Let

$$\mathcal{H}_n := \sigma(X_1, \mathcal{A}_1, A_1, Y_1, \cdots, X_{n-1}, \mathcal{A}_{n-1}, A_{n-1}, Y_{n-1}, \mathcal{A}_n, X_n, A_n)$$

be the σ-algebra summarising the information available just before Y_n is observed. The following assumptions are similar to those in the fixed design setting.

Assumption 1'. *There exists $\sigma > 0$ such that for every $n = 1,\ldots,N$ and every $u \in \mathbb{R}$, $\mathbb{E}[\exp(u\epsilon_n) \mid \mathcal{H}_n] \leq \exp(u^2\sigma^2/2)$.*

Assumption 2'. *The following holds:*

(i) $\kappa_f := \sup_{x\in\mathcal{X}, a\in\mathcal{A}, \theta\in\Theta} \|\nabla f_\theta(x,a)\|_2 < \infty$.

(ii) $\hbar_f := \sup_{x\in\mathcal{X},a\in\mathcal{A},\theta\in\Theta} \|\nabla^2 f_\theta(x,a)\|_2 < \infty$.
(iii) $\kappa_g := \sup_{x\in\mathcal{X},a\in\mathcal{A},\theta\in\Theta} \|\nabla g_\theta(x,a)\|_2 < \infty$.
(iv) $\underline{\kappa}_\phi := \inf_{x\in\mathcal{X},a\in\mathcal{A},\theta\in\Theta} \phi'(f_\theta(x,a)) > 0$.
(v) $\beta_\Theta := \sup_{\theta\in\Theta} \|\theta\|_2 < \infty$.
(vi) $0 \le r_n(x,a;\theta) \le \beta_r < \infty$, $\forall x \in \mathcal{X}, a \in \mathcal{A}, \theta \in \Theta, 1 \le n \le N$.
(vii) $\hbar_r := \sup_{x\in\mathcal{X},a\in\mathcal{A},\theta\in\Theta} \|\nabla^2 r(x,a;\theta)\|_2 < \infty$.

Assumption 1' is a standard sub-Gaussian assumption in the bandit literature. Note that we allow the noise term in the prediction model to be dependent on the historical decisions and contexts. Assumption 2' extends the boundedness assumptions of model parameters from the linear and generalized bandits to our general parametric setting. Similar to the fixed design, we next analyze $\|\xi_n(\theta)\|^2_{V_n(\theta)^{-1}}$ and $\|\hat{\xi}_n\|^2_{\hat{V}_n^{-1}}$, through which we derive the bound for $\|\hat{\theta}_n - \theta^*\|^2_{\hat{V}_n}$.

In the remainder of this subsection, we assume that there exists $1 \le \tau_1 \le N$ such that

$$\lambda_0 := \inf_{\theta\in\Theta} \lambda_{\min}(V_{\tau_1}(\theta)) > 0$$

holds with high probability. This can be assured by, for example, initial random exploration, as indicated in the following corollary.

Corollary 3. *Suppose $\mathcal{A}_n = \mathcal{A}$ for all n, the contexts $\{X_n\}_{n=1}^N$ are independent and identically distributed from a distribution $\mathcal{D}_\mathcal{X}$, and there exist strictly positive constants $\underline{\lambda} < \bar{\lambda}$ and a distribution $\mathcal{D}_\mathcal{A}$ on $\mathcal{A}$ such that*

$$\lambda_{\min}(\mathbb{E}_{X\sim\mathcal{D}_\mathcal{X}}[XX^\top]),\ \lambda_{\min}(\mathbb{E}_{X\sim\mathcal{D}_\mathcal{X},A\sim\mathcal{D}_\mathcal{A}}[\nabla f_\theta(X,A)\nabla f_\theta(X,A)^\top]) \in [\underline{\lambda},\bar{\lambda}],\ \forall\theta\in\Theta.$$

Let $\delta \in (0,1)$. Set $\tau_1 = \max(\lceil \frac{\lambda_0}{\underline{\lambda}} \rceil, \frac{2\bar{\lambda}\log(d/\delta)}{\underline{\lambda}\log(e/2)})$. Then with probability at least $1-\delta$, the event $\inf_{\theta\in\Theta}\lambda_{\min}(V_{\tau_1}(\theta)) \ge \lambda_0$ holds.

When generating random contexts and actions is not a feasible option, an alternative is to consider the ridge regularization in the prediction. For brevity, we do not consider such a choice in this work.

Lemma 4. *Assume Assumptions 1' and 2'(i) are in force. Let $\delta \in (0,1)$ and $\tau_1 \le n \le N$.*

(I) *For every $\theta \in \Theta$, with probability at least $1-\delta$, it holds that*

$$\|\xi_n(\theta)\|^2_{V_n(\theta)^{-1}} \le 16 d\eta^2\sigma^2\log(n)\log(d/\delta),$$

where $\eta = \sqrt{3 + 2\log(1 + 2\kappa_f^2/\lambda_0)}$.

(II) *Assume additionally that Assumptions 2'(ii)-(vi) are in force. Then with probability at least $1-3\delta$, it holds that*

$$\|\hat{\xi}_n\|^2_{\hat{V}_n^{-1}} \le (35\eta)^2\sigma^2 d^2\log(n)\log(nd\varsigma\log(2/\delta)/\delta),$$

where $\varsigma = \frac{8\beta_\Theta\sigma\kappa_f\hbar_f}{\lambda_0}\cdot(1+\frac{\hbar_f}{\lambda_0})$.

Lemma 4 shows that $\|\xi_n(\theta)\|_{V_n(\theta)^{-1}}$ is $\tilde{O}(\log(n))$. It is parallel to the results for linear regression and generalized linear regression under adaptive designs, which are built on self-normalized processes [25, 74, 2]. But unlike these results, to accommodate for general parametric prediction modes, we consider a new exponential super-martingale based on the gradient $\nabla f_\theta(X_i, A_i)$ to prove (I):

$$\prod_{i=1}^n \exp\left(u\epsilon_i \nabla f_\theta(X_i,A_i)^\top x - u^2\sigma^2(\nabla f_\theta(X_i,A_i)^\top x)^2\right),$$

where $x \in \mathbb{R}^d$ and $u \in \mathbb{R}$ are fixed. The concentration of this super-martingale renders a high-probability bound for $x^\top \xi_n(\theta)$. Taking $x = V_n(\theta)^{-1/2} e_j$, $j = 1, \ldots, d$ and using union bound yields a high-confidence bound for $\|\xi_n(\theta)\|^2_{V_n(\theta)^{-1}}$. For (II), to take into account the randomness of $\hat{\theta}_n$, we adopt a covering number argument, using the fact that $\|\xi_n(\theta)\|^2_{V_n(\theta)^{-1}}$ is a Lipschitz function of θ.

Compared with Lemma 3, the bound in Lemma 4(II) has an additional factor d. The main reason is that $\hat{V}_n$ is deterministic in the fixed design but is random in the adaptive design. As such, we have to use a more involved super-martingale argument combined with a covering number argument as described above. Note that this argument is not needed in the special cases of linear and generalized linear models, because the gradients $\nabla f_\theta(X_i, A_i)$ and $\nabla f_{\hat{\theta}_n}(X_i, A_i)$ are equal or parallel to each other under proper conditions, and thereby $\|\hat{\xi}_n\|_{\hat{V}_n^{-1}}$ can be controlled directly by $\|\xi_n(\theta)\|_{V_n(\theta)^{-1}}$ in Lemma 4(I) up to a constant factor. In the general parametric case, however, we no longer have such a nice property. A complete proof of Lemma 4 is given in Appendix EC.2.5.2.

To derive a bound on the radius ρ_n, in the following, we bound $\|\theta^* - \hat{\theta}_n\|_{\hat{V}_n}$. Similar to the the fixed design, we also make Assumption 3, which appears to be new in the literature on parametric contextual learning literature, as most existing parametric bandit problems have a convex statistical loss. In Section 5.2, we will provide an illustrative example on low-rank matrix estimation whose statistical loss function is non-convex in θ while still satisfying Assumption 3.

THEOREM 2 **(RADIUS SELECTION FOR ADAPTIVE DESIGN).** *Let $\delta \in (0, 1)$. Assume Assumptions 1', 2', 3 are in force. Let $\tau_1 < n \le N$. Then with probability at least $1 - 3\delta$,*

$$\|\theta^* - \hat{\theta}_n\|^2_{\hat{V}_n} \le \frac{(35\eta)^2 \sigma^2 d^2 \log(n) \log(\varsigma n d \log(2/\delta)/\delta)}{\min(\underline{\kappa}_\phi^2, \alpha^2)},$$

where η and ς are defined in Lemma 4.

The proof is similar to the fixed design setting and is given in Appendix EC.2.5.3. According to Theorem 2, replacing δ with δ/n and choosing the radius

$$\rho_n = \frac{35\sigma\eta d\sqrt{2\log(n)\log(\varsigma n d \log(2n/\delta)/\delta)}}{\min(\underline{\kappa}_\phi, \alpha)}, \quad \forall \tau_1 \le n \le N, \tag{14}$$

we can ensure that with probability at least $1 - \delta$, $\theta^* \in \Theta_n$ simultaneously for all $\tau_1 \le n \le N$. The additional factor $\sqrt{d}$ in Theorem 2 compared to Theorem 1 results from the extra factor d in Lemma 4(II) compared with Lemma 3, as we discussed above. Based on this result, in the next section we will develop a new algorithm for online contextual optimization with provable performance guarantees.

4. Data-driven Optimistic Optimization (DOO)

In this section, we apply our results in Section 3.2 to sequential contextual decision-making problems, in which the goal of the decision-maker is to choose a sequence of decisions $\{A_n\}_{n=1}^N$ to maximize the cumulative expected reward

$$\sum_{n=1}^N r_n(X_n, A_n; \theta^*).$$

A general principle of solving such a problem is the optimism in the face of uncertainty [67], which balances the exploration-exploitation trade-off by acting as if the environment is as nice as plausibly possible. Based on this principle, in this section, we propose a new framework, termed as *data-driven optimistic optimization* (DOO), to solve sequential contextual decision-making problems. At each round, based on the historical estimate using (Oracle), it solves the following problem

$$\max_{a \in \mathcal{A}} U_n(a) := \max_{\theta \in \Theta_n} r_n(X_n, a; \theta), \tag{DOO}$$

or its approximation

$$\max_{a\in\mathcal{A}} r_n(X_n, a; \hat{\theta}_n) + \rho_n \cdot \|\nabla r_n(x, a; \hat{\theta}_n)\|_{\hat{V}_n^{-1}}, \tag{$\widehat{\text{DOO}}$}$$

where $\{\rho_n\}_{n=1}^N$ is chosen according to (14). Namely, in round n, for each decision $a \in \mathcal{A}_n$, we compute its upper confidence bound on the true reward and then choose the decision with the largest upper confidence bound. This leads to the following online algorithm. The number of initial exploration periods τ and the uncertainty set radii $\{\rho_n\}_{n=\tau}^N$ will be specified in Section 4.2.

Algorithm 1

Input: $\tau > 0$, $\{A_n\}_{n\in[\tau]}$
1: **for** $n = 1, \cdots, N$ **do**
2: If $n \le \tau$, play a decision A_n for exploration;
3: Compute $\hat{\theta}_n$ using (Oracle) and update $\hat{V}_n$ using (6);
4: Play a decision A_n by solving ($\widehat{\text{DOO}}$), observe a response Y_n and receive reward R_n;

Below, we relate this algorithm to existing bandit algorithms in Section 4.1, and the performance guarantees are established in Section 4.2.

4.1. Connection with UCB Algorithms

In this subsection, we give three examples showing that our proposed Algorithm 1 is reduced to classical bandit algorithms when the reward R coincides with the response Y.

Example 1 (Linear Model). Suppose a linear model between the reward and the decision

$$R_i = Y_i = A_i^\top \theta^* + \epsilon_i.$$

Consider the square loss $\frac{1}{2}(y - a^\top\theta)^2$. The least-square estimate $\hat{\theta}_n$ satisfies the first-order optimality condition $\sum_{i=1}^{n-1}(Y_i - A_i^\top\hat{\theta}_n)A_i = 0$. It follows that $\bar{V}_n = \hat{V}_n = \sum_{i=1}^{n-1} A_i A_i^\top$. Formulations ($\widehat{\text{DOO}}$) reads

$$U_n(a) = \max_{\theta\in\mathbb{R}^d}\left\{a^\top\theta :\ \|\theta - \hat{\theta}_n\|_{\hat{V}_n} \le \rho_n\right\} = a^\top\hat{\theta}_n + \rho_n \cdot \|a\|_{\hat{V}_n^{-1}},\ \text{where } \hat{V}_n = \sum_{i=1}^{n-1} A_i A_i^\top.$$

Thus, Algorithm 1 recovers the Optimism-in-Face-of-Uncertainty for Linear bandits (OFUL) algorithm [1]. ♣

Example 2 (Stochastic i.i.d. bandits). The classical d-armed bandit problem can be stated in the following form. Let $\mathcal{A}_n = \{1, \cdots, d\}$ and $X_n = \{e_a\}_{a\in[d]}$ be the standard basis in $\mathbb{R}^d$. Suppose

$$R_i = Y_i = e_{A_i}^\top\theta^* + \epsilon_i = \theta^*_{A_i} + \epsilon_i.$$

Let $N_a(n) = \sum_{i=1}^{n-1}\mathbf{1}(A_i = a)$ be the empirical frequency of arm a, and let $\hat{\theta}_n = (\hat{\theta}_{n,a})_{a\in[d]}$ be the empirical mean vector, where $\hat{\theta}_{n,a} = N_a(n)^{-1}\sum_{i=1}^{n-1} Y_i\mathbf{1}(A_i = a)$. Note that $\hat{\theta}_n$ solves the least square problem $\min_\theta \sum_{i=1}^{n-1}(Y_i - e_{A_i}^\top\theta)^2$. Thus $\bar{V}_n = \hat{V}_n = \sum_{i=1}^{n-1} e_{A_i}e_{A_i}^\top = \sum_{a=1}^d N_a(n)e_a e_a^\top$, whence ($\widehat{\text{DOO}}$) become

$$U_n(a) = \max_{\theta\in\mathbb{R}^d}\left\{e_a^\top\theta : \sum_{a=1}^d N_a(n)(\theta_a - \hat{\theta}_{n,a})^2 \le \rho_n^2\right\} = \hat{\theta}_{n,a} + \sqrt{\frac{\rho_n^2}{N_a(n)}}.$$

This recovers the classical UCB algorithm [54] for multi-armed bandits. ♣

Example 3 (Generalized linear model). Suppose the reward and the decision satisfy a generalized linear model (GLM)

$$R_i = Y_i = \phi(A_i^\top \theta^*) + \epsilon_i,$$

where ϕ is the inverse link function in the generalized linear model. Suppose the (Oracle) is given by the maximum quasi-likelihood estimation with $\psi(y, a, \varepsilon; \theta) = (y - \phi(\theta^\top a) - \varepsilon)a$:

$$\sum_{i=1}^{n-1} \big(Y_i - \phi(A_i^\top \theta)\big) A_i = 0.$$

For the canonical exponential family, the density of Y given A equals $\exp(yA^\top\theta^* - b(A^\top\theta^*) + c(y))$, where b is a twice continuously differentiable function satisfying $b' = \phi$ and c is a real function. This indicates that the conditional variance of ϵ given A equals $\phi'(A^\top\theta^*)$. Thereby we set $\omega_{n,i}^2 = \phi'(A_i^\top \hat{\theta}_n)$ in the formulation ($\mathsf{UQ_o}$). Then $\bar{V}_n = \hat{V}_n = \sum_{i=1}^{n-1} \phi'(A_i^\top \hat{\theta}_n) A_i A_i^\top$, and (DOO) is equivalent to

$$U_n(a) = \max_{\theta \in \mathbb{R}^d} \Big\{ \phi(a^\top \theta) : \ \|\theta - \hat{\theta}_n\|_{\hat{V}_n} \le \rho_n \Big\},$$

and ($\widehat{\text{DOO}}$) becomes

$$\phi(a^\top \hat{\theta}_n) + \rho_n \cdot |\phi'(a^\top \hat{\theta}_n)| \|a\|_{\hat{V}_n^{-1}}.$$

This form is closely related to but slightly different from existing algorithms for GLM bandits [33, 53, 59]. They did not consider constructing Θ_n and solving the joint optimization over decisions and parameters directly, but instead building the UCB using a proxy $\rho_n \cdot \|a\|_{\hat{V}_n^{-1}}$ originated from linear bandits. In Appendix A, we will show that our algorithm improves their regret bounds by a constant term that was discussed in Section 4.2 in [32]. ♣

4.2. Performance Guarantees for Online Contextual Optimization

In this section, we analyze the performance of Algorithm 1. Consistent with the literature on online optimization, we consider the cumulative (pseudo-)regret as a performance measure

$$\mathsf{Regret}_N = \mathbb{E}\left[\sum_{n=1}^{N} \left(\sup_{a \in \mathcal{A}_n} r_n(X_n, a; \theta^*) - R_n \right) \right],$$

where the expectation is taken with respect to all history $\mathcal{H}_N$. We aim to provide a high-probability regret bound for Algorithm 1.

Let $A_n^* \in \arg\max_{a \in \mathcal{A}_n} r_n(X_n, a; \theta^*)$ be a true optimal decision, which is assumed to exist, since we can argue by approximation otherwise. Observe that the one-step regret, denoted by reg_n, can be upper bounded in the following way. According to the radius selection rule proposed in Section 3.2 and our formulation (DOO), it holds that with probability at least $1 - 3\delta$,

$$U_n(a) \ge r_n(X_n, a; \theta^*), \ \forall a \in \mathcal{A}_n, \ n = 1, \ldots, N.$$

Whenever this holds, we have

$$\mathsf{reg}_n = r_n(X_n, A_n^*; \theta^*) - r_n(X_n, A_n; \theta^*) \le U_n(A_n^*) - r_n(X_n, A_n; \theta^*).$$

Since $A_n \in \arg\max_{a \in \mathcal{A}_n} U_n(a)$, we have $U_n(A_n^*) \le U_n(A_n)$. Let $\tilde{\theta}_n \in \arg\max_{\theta \in \Theta_n} r_n(X_n, A_n; \theta)$. It follows that

$$\mathsf{reg}_n \le U_n(A_n) - r_n(X_n, A_n; \theta^*) = r_n(X_n, A_n; \tilde{\theta}_n) - r_n(X_n, A_n; \theta^*).$$

If we choose decision A_n based on ($\widehat{\text{DOO}}$) instead of (DOO), then it follows from Lemma 2 that

$$\mathsf{reg}_n \le U_n(A_n) + \hbar_r \|\tilde{\theta}_n - \hat{\theta}_n\|_2^2 - r_n(X_n, A_n; \theta^*) = r_n(X_n, A_n; \tilde{\theta}_n) - r_n(X_n, A_n; \theta^*) + \hbar_r \|\tilde{\theta}_n - \hat{\theta}_n\|_2^2.$$

To bound the right-hand side of the inequality above, we impose the following assumption.

Assumption 4. *There exist constants $\mu, \hbar, \gamma_r > 0$ such that for every θ satisfying $\|\theta - \theta^*\|_2 \le \gamma_r$, and for every $x \in \mathcal{X}$, $a \in \mathcal{A}$ and $1 \le n \le N$,*

$$|r_n(x, a; \theta) - r_n(x, a; \theta^*)| \le \mu |f_\theta(x, a) - f_{\theta^*}(x, a)| + \hbar \|\theta - \theta^*\|_2^2.$$

With this assumption, the one-step regret reg_n of (DOO) can be upper bounded as

$$\mathsf{reg}_n \le \mu |f_{\tilde{\theta}_n}(X_n, A_n) - f_{\theta^*}(X_n, A_n)| + \hbar \|\tilde{\theta}_n - \theta^*\|_2^2 + \hbar_r \|\tilde{\theta}_n - \hat{\theta}_n\|_2^2. \tag{15}$$

Note that Assumption 4 is a local assumption that is imposed only on points that are within γ_r-neighborhood of θ^*. The following example shows that if Assumption 4 fails to hold, one may not achieve a sub-linear regret even in linear bandits with two decisions.

Example 4. Let $\mathcal{A} = [0, 1] \times [0, 1]$, $a = (a_1, a_2)$ and $\theta^* = (\theta_1^*, \theta_2^*) > 0$. Suppose $r_n(a; \theta^*) = \theta_1^* a_1$ and $f_{\theta^*}(a) = \theta_2^* a_2$. Then we have $|r_n(a; \theta) - r_n(a; \theta^*)| = |\theta_1 - \theta_1^*| a_1$, whereas $|f_\theta(a) - f_{\theta^*}(a)| = |\theta_2 - \theta_2^*| a_2$, which apparently violates Assumption 4. Observe that the response does not provide any information on θ_1^*. Therefore, no algorithm can achieve a sublinear regret even when θ_2^* is known exactly. ♣

Using Cauchy-Schwarz inequality, we further bound the one-step regret in (15) as

$$\mathsf{reg}_n \le \mu \|\nabla f_{\tilde{\theta}_n}(X_n, A_n)\|_{V_{n,*}^{-1}} \|\tilde{\theta}_n - \theta^*\|_{V_{n,*}} + \hbar \|\tilde{\theta}_n - \theta^*\|_2^2 + \hbar_r \|\tilde{\theta}_n - \hat{\theta}_n\|_2^2, \tag{16}$$

where $\bar{\theta}_n$ is determined by the mean value theorem. In light of this, the rest of this subsection is mainly devoted to providing bounds on $\|\theta^* - \hat{\theta}_n\|_{V_{n,*}}$ and $\|\nabla f_\theta(X_n, A_n)\|_{V_{n,*}^{-1}}$. Recall that $V_{n,*} = \sum_{i=1}^{n-1} \nabla f_{\theta^*}(X_i, A_i) \nabla f_{\theta^*}(X_i, A_i)^\top$, and $\hat{V}_n = \sum_{i=1}^{n-1} \nabla f_{\hat{\theta}_n}(X_i, A_i) \nabla f_{\hat{\theta}_n}(X_i, A_i)^\top$. In Theorem 2, we have already bounded $\|\hat{\theta}_n - \theta^*\|_{\hat{V}_n}$, so it remains to relate $\|\cdot\|_{V_{n,*}}$ with $\|\cdot\|_{\hat{V}_n}$. We impose the following assumption.

Assumption 5. *There exist constants $\alpha_f \in [0, \frac{1}{8})$, $\tau_2 \in \mathbb{N}_{\ge 1}$ and a sequence $\{\gamma_n\}_n$ satisfying $\gamma_n \ge \frac{\gamma_0}{n^{1/4}}$ for some $\gamma_0 > 0$, such that for all $n \ge \tau_2$, $x \in \mathcal{X}$, $a \in \mathcal{A}$, $\vartheta \in \mathbb{R}^d$, and $\theta \in \Theta_n$ with $\|\theta - \theta^*\|_2 \le \gamma_n$,*

$$\sum_{i=1}^{n-1} (\vartheta^\top (\nabla f_\theta(X_i, A_i) - \nabla f_{\theta^*}(X_i, A_i)))^2 \le \alpha_f^2 \sum_{i=1}^{n-1} (\vartheta^\top \nabla f_{\theta^*}(X_i, A_i))^2. \tag{17}$$

For linear models, Assumption 5 is satisfied trivially with $\alpha_f = 0$, $\tau_2 = 1$ and $\gamma_n = \infty$. For general prediction models, under the assumption in Corollary 3, Assumption 5 can be satisfied by exploring $\tau_2 = \max((16\hbar_f)^2, 1) \max(\rho_N^2, 1) d\sqrt{N}$ times during initialization. Indeed, this guarantees that with high probability $\lambda_{\min}(V_{n,*}) \ge \max((16\hbar_f)^2, 1) \max(\rho_N^2, 1) \underline{\lambda} n^{1/2}/2$ and thereby $\|\hat{\theta}_n - \theta^*\|_2 \le \rho_n / (16\hbar_f \sqrt{\rho_N^2 \underline{\lambda} n^{1/2}/2}) \le 1/(16\hbar_f \sqrt{\underline{\lambda} n^{1/2}/2})$, for all $\tau_2 \le n \le N$. Set $\gamma_n = \frac{\sqrt{2}}{16\hbar_f \sqrt{\underline{\lambda}} n^{1/4}}$, $\gamma_0 = \sqrt{\frac{2}{\underline{\lambda}}}$, and $\alpha_f = \frac{1}{16}$. It follows that $\|\hat{\theta}_n - \theta^*\|_2 \le \gamma_n$. By the mean value theorem, there exists $\bar{\theta}$ such that for every $\vartheta \in \mathbb{R}^d$,

$$(\vartheta^\top (\nabla f_\theta(x, a) - \nabla f_{\theta^*}(x, a)))^2 = (\vartheta^\top \nabla^2 f_{\bar{\theta}}(x, a)(\theta - \theta^*))^2 \le \|\vartheta\|_2^2 \hbar_f^2 \gamma_n^2 = \frac{\gamma_0^2 \alpha_f^2}{\sqrt{n}} \|\vartheta\|_2^2.$$

Meanwhile, we have

$$\frac{1}{n} \sum_{i=1}^{n} (\vartheta^\top \nabla f_{\theta^*}(X_i, A_i))^2 \ge \frac{1}{n} \lambda_{\min}(V_{n,*}) \|\vartheta\|_2^2 \ge \frac{\gamma_0^2 n^{1/2}}{n} \|\vartheta\|_2^2 = \frac{\gamma_0^2}{\sqrt{n}} \|\vartheta\|_2^2.$$

Combining the two inequalities above yields that for all $x \in \mathcal{X}$ and $a \in \mathcal{A}$,

$$\frac{(\vartheta^\top (\nabla f_\theta(x, a) - \nabla f_{\theta^*}(x, a)))^2}{\frac{1}{n} \sum_{i=1}^{n-1} (\vartheta^\top \nabla f_{\theta^*}(X_i, A_i))^2} \le \alpha_f^2,$$

which implies (17). We remark that the left-hand side of the inequality above can be viewed as a first-order counterpart of the counterfactual action divergence introduced by Xu and Zeevi [94]:

$$\sup_{\theta\in\Theta} \frac{|f_\theta(x,a)-f_{\theta^*}(x,a)|^2}{\sum_{i=1}^n (f_\theta(X_i,A_i)-f_{\theta^*}(X_i,A_i))^2}.$$

The discussion above provides an implementable way of ensuring Assumption 2 in their paper for general parametric prediction models. In this setting, our Assumption 5 is weaker and only requires a local condition around θ^*.

Observe that when $\theta=\hat{\theta}_n$, condition (17) becomes

$$\|\vartheta\|_{\hat{V}_n}^2 + \|\vartheta\|_{V_{n,*}}^2 - 2\sum_{i=1}^{n-1} \vartheta^\top \nabla f_{\hat{\theta}_n}(X_i,A_i)\nabla f_{\theta^*}(X_i,A_i)^\top \vartheta \le \alpha_f^2 \|\vartheta\|_{V_{n,*}}^2.$$

Then by Cauchy-Schwarz inequality, we have $(\|\vartheta\|_{\hat{V}_n} - \|\vartheta\|_{V_{n,*}})^2 \le \alpha_f^2\|\vartheta\|_{V_{n,*}}^2$, which implies the sandwich inequality $(1-\alpha_f)\|\vartheta\|_{V_{n,*}} \le \|\vartheta\|_{\hat{V}_n} \le (1+\alpha_f)\|\vartheta\|_{V_{n,*}}$. Thereby a consequence of Assumption 5 is the following result, whose proof can be found in Appendix EC.3.1.

Lemma 5. *Assume Assumption 5 is in force. For all $n \ge \tau_2$, it holds that for every $\vartheta\in\mathbb{R}^d$,*

$$\|\vartheta\|_{V_{n,*}}^2 \mathbf{1}\{\|\hat{\theta}_n-\theta^*\|_2 \le \gamma_n\} \le \frac{1}{1-2\alpha_f}\|\vartheta\|_{\hat{V}_n}^2.$$

Lemma 4 and Lemma 5 imply that under the radius selection rule (14), the term $\|\tilde{\theta}_n-\theta^*\|_{V_{n,*}}$ in (16) can be bounded via

$$\|\tilde{\theta}_n-\theta^*\|_{V_{n,*}}\mathbf{1}\{\|\hat{\theta}_n-\theta^*\|_2\le\gamma_n\} \le \frac{1}{1-2\alpha_f}\|\tilde{\theta}_n-\theta^*\|_{\hat{V}_n} \le \frac{2}{1-2\alpha_f}\rho_n. \tag{18}$$

It remains to bound the other term $\|\nabla f_{\theta^*}(X_n,A_n)\|_{V_{n,*}^{-1}}$ in (16). We invoke the following lemma that generalizes the Elliptical Potential Lemma [7, 1] in linear bandits, whose proof is given in Appendix EC.3.2.

Lemma 6. *Assume Assumption 2' (i) is in force. For any $\tau_1 \le \tau \le N$, it holds that*

$$\sum_{n=\tau}^N \left(1 \wedge \|\nabla f_{\theta^*}(X_n,A_n)\|_{V_{n,*}^{-1}}^2\right) \le 2d\log\left(\frac{N\kappa_f^2}{d\det(V_{\tau,*})^{1/d}}\right).$$

Combining the results above, we obtain the following result, whose proof can be found in Appendix EC.3.3.

Theorem 3. *Assume Assumptions 1', 2', 3, 4, 5 are in force. Let $\delta\in(0,1)$. Define an event $\mathcal{E}_n = \{\lambda_{\min}(\hat{V}_n) \ge \rho_n^2/\gamma_n^2\}$. Then with probability at least $1-3\delta$, the regret of Algorithm 1 is upper bounded by*

$$\mathsf{Regret}_N \mathbf{1}\{\cap_{n=\tau}^N \mathcal{E}_n\} \;\le\; \beta_r\tau + \frac{3}{1-2\alpha_f}\rho_N(\beta_r+\mu+\hbar\gamma_0+\hbar_r\gamma_0)\sqrt{Nd\log\left(\frac{N\kappa_f^2}{d\lambda_0}\right)},$$

where $\rho_N = \frac{35\sigma\eta d\sqrt{2\log(N)\log(\varsigma Nd\log(2N/\delta)/\delta)}}{\min(\underline{\kappa}_\phi,\alpha)}$.

Theorem 3 shows that with high probability, when the event $\cap_{n=\tau}^{N}\mathcal{E}_n$ holds, Algorithm 1 has an $\tilde{O}(\sqrt{N})$-regret bound. To assure this event, under the assumption in Corollary 3, we can explore $\tau_3 = \lceil \max(\frac{2\rho_N^2}{\gamma_N^2 \underline{\lambda}^2}, \frac{2\bar{\lambda}\log(dN/\delta)}{\underline{\lambda}\log(e/2)})\rceil$ times during initiation; see Appendix EC.3.4 for a proof.

To summarize, in Algorithm 1, we set $\tau = \max(\tau_1, \tau_2, \tau_3)$ and choose ρ_n according to (14). Then by Theorem 3, the regret is $\tilde{O}(d^{3/2}\sqrt{N})$. This bound matches the best bound in terms of the order of N but does not match the best order of d for some special cases. For instance, the optimal regret for linear bandits and generalized linear bandits is $O(d\sqrt{N})$ (e.g., Abbasi-Yadkori et al. [1], Filippi et al. [33]). This additional $\sqrt{d}$ factor results from the bound in Lemma 4(II) that can be improved for linear bandits and generalized linear bandits, as mentioned after Lemma 4. In Appendix A, we derive a regret bound for generalized linear models that has an optimal dependence on d. We remark that the optimality as discussed above concerns with regret bounds that do not depend on the number of decisions, which are useful in the setting when the number of decisions is large. This is different from the optimal regret bounds developed in recent works for contextual bandits with general reward functions [35, 85, 94], that typically has a linear or super-linear dependence on the number of decisions. Moreover, in these works the reward and the response in the prediction model are identical.

5. Applications

In this section, we illustrate our results with two applications. The first one is an optimal dynamic pricing problem for queuing service, in which the response variable in the prediction model for estimating the unknown parameter is different than the reward. The second one is a rank-one matrix estimation problem, which involves non-convex statistical loss functions. We will demonstrate the sizing of the uncertainty set in both offline and online settings, as well as our proposed robust and optimistic frameworks with performance guarantees. We remark that, in the online setting, both of them cannot be handled directly by existing UCB algorithms for contextual bandits with efficient implementation.

5.1. Pricing for Queueing Service

In this subsection, we consider an optimal pricing problem for an $M/G/1$ queue. Consider a service system in which the demand can be controlled by adjusting the price. Let $\mathcal{A} \subset [\underline{p}, \overline{p}]$ be the set of admissible prices with strictly positive lower and upper bounds. The decision-maker sets a price $p_n \in \mathcal{A}$ at the beginning of the time period n (such as one day). Within the time period n, the customers arrive according to a Poisson process with a known rate function $\Lambda(\cdot)$; let M_n be the number of customers arrived during the time period n. The decision-maker receives reward p_n per serving one customer and the goal is to maximize the reward while reducing the average waiting time.

For many practical problems, the service time is heterogeneous among customers. Consider the following service model involving two potential service tasks. Let $X_{nj} \subset \mathcal{X}$ be the feature vector of the j-th customer arrived during the time period n, where X_{nj}, $j = 1, \ldots, M_n$, are independently and identically distributed as X and has a positive definite covariance matrix $\lambda_{\min}(\mathbb{E}[XX^\top])$; while with probability $\phi(\theta^{*\top} X_{nj})$, where $\theta^* \in \mathbb{R}^d$ and $\phi(\cdot)$ is the logistic function $1/(1+\exp(-\cdot))$, the customer requests only one service, with an exponential service time with rate u_1. With probability $1 - \phi(\theta^{*\top} X_{nj})$, the customer requests an additional service task with an exponential service time with rate u_2, independent from the service time of the first task. Suppose M_n is reasonably large so that it makes sense to consider the steady state limit. In this case, the mean and variance of the service time can be estimated by (see Appendix EC.4.1)

$$m(\theta^*) = \frac{1}{u_1} + \frac{1}{u_2}\frac{1}{M_n}\sum_{j=1}^{M_n}\phi(\theta^{*\top} X_{nj}), \quad v(\theta^*) = \frac{1}{u_1^2} + \frac{1}{u_2^2}\frac{1}{M_n}\sum_{j=1}^{M_n}\phi(\theta^{*\top} X_{nj})(2 - \phi(\theta^{*\top} X_{nj})).$$

The expected waiting time of a customer equals [37, p.222]

$$W(p_n;\theta^*) := \frac{\Lambda(p_n)v(\theta^*)}{2(1-\Lambda(p_n)m(\theta^*))}.$$

Consider the reward as a weighted combination of the expected revenue and the expected waiting time $M_n(p_n - cW(p_n;\theta^*))$, where $c>0$ is a given weight parameter, thus the expected reward equals

$$r_n(p_n;\theta^*) = \Lambda(p_n)(p_n - cW(p_n;\theta^*)).$$

To estimate the unknown parameter θ^*, let Y_{nj} be a binary response variable that indicates whether the j-th customer needs an additional service task and consider a regression model $Y_{nj} = \phi({\theta^*}^\top X_{nj}) + \epsilon_{nj}$, where $\epsilon_{nj} = 1-\phi({\theta^*}^\top X_{nj})$ with probability $\phi({\theta^*}^\top X_{nj})$ and $\epsilon_{nj} = -\phi({\theta^*}^\top X_{nj})$ with probability $1-\phi({\theta^*}^\top X_{nj})$. Then the unknown parameter θ^* can be estimated by solving the quasi-likelihood maximization

$$\sum_{i=1}^{n-1}\sum_{j=1}^{M_i}(Y_{ij} - \phi(X_{ij}^\top\hat{\theta}_n))X_{ij} = 0. \tag{19}$$

Note that in this problem, the response $Y_{n,j}$ and the expected reward r_n are different and are linked with each other through a shared parameter θ^*. As such, this problem is not fitted into the classical contextual bandit problems. We also note that this model generalizes the model introduced in Section 2.1 in the sense that there are multiple observations during one time period, although Theorem 1 remains to hold and Theorem 2 holds similarly. Assume $\mathcal{X}$ and Θ are compact, and for every admissible price $p\in\mathcal{A}$, the system is stable, i.e., $\min_{\theta\in\Theta,p\in\mathcal{A}} 1-\Lambda(p)m(\theta) > 0$.

Similar to Example 3, let $f_\theta(x,a) = \theta^\top x$, and set

$$\hat{V}_n = \sum_{i=1}^{n-1}\sum_{j=1}^{M_i}\phi'(\hat{\theta}_n^\top X_{ij})X_{ij}X_{ij}^\top,$$

where corresponding weights $\omega_{n,ij} = \phi'(\hat{\theta}_n^\top X_{ij})$. We arrive at the following form of the uncertainty set $\Theta_n = \{\theta\in\Theta:\ \|\theta-\hat{\theta}_n\|_{\hat{V}_n}\le\rho_n\}$.

We first discuss the radius selection and the performance guarantees in an offline setting. To this end, let us verify the assumptions required by Theorem 1. Recall ϕ is the logistic function. Assumption 1 holds since ϵ_{nj}'s are bounded and thus sub-Gaussian. In Assumption 2, $\kappa_f = \sup_{x\in\mathcal{X}}\|x\|_2$, $\kappa_g = \sup_{x\in\mathcal{X},\theta\in\Theta}\frac{\exp(\theta^\top x)}{(1+\exp(\theta^\top x))^2}\|x\|_2$, $\underline{\kappa}_\phi = \inf_{x\in\mathcal{X},\theta\in\Theta}\frac{\exp(\theta^\top x)}{(1+\exp(\theta^\top x))^2}$, $\bar{\kappa}_\phi = \sup_{x\in\mathcal{X},\theta\in\Theta}\frac{\exp(\theta^\top x)}{(1+\exp(\theta^\top x))^2}$. These parameters are all finite due to the compactness of $\mathcal{X}$ and Θ. In Appendix EC.4.2, we verify Assumption 3 that for all $\theta\in\Theta$ and $x\in\mathcal{X}$,

$$\big(\phi(x^\top\theta)-\phi(x^\top\theta^*)\big)x^\top(\theta-\theta^*) \ge\ \phi'(x^\top\theta)\cdot\big(x^\top(\theta-\theta^*)\big)^2 \ge \underline{\kappa}_\phi\big(x^\top(\theta-\theta^*)\big)^2.$$

Given a fixed set of observations $\{(p_i, X_{i\cdot}, Y_{i\cdot})\}_{i=1}^{n-1}$, set ρ_n as

$$\rho_n = \frac{4\sigma\bar{\kappa}_\phi\kappa_f\sqrt{\log(5\kappa_f)(d+\log(1/\delta))}}{\underline{\kappa}_\phi}.$$

According to Corollary 1 and Remark 1, it holds with probability at least $1-\delta$ that for every $p\in[\underline{p},\overline{p}]$, $r(p;\theta^*)\in[\widehat{L}_n(p)-\hbar_r\varrho_n^2,\ \widehat{U}_n(p)+\hbar_r\varrho_n^2]$, where

$$\widehat{L}_n(p) = r(p;\hat{\theta}_n) - \rho_n\cdot\|\nabla r(p;\hat{\theta}_n)\|_{\hat{V}_n^{-1}},$$
$$\widehat{U}_n(p) = r(p;\hat{\theta}_n) + \rho_n\cdot\|\nabla r(p;\hat{\theta}_n)\|_{\hat{V}_n^{-1}}.$$

Consider the robust problem

$$\widehat{R}_{\text{rob}} = \max_{p\in[\underline{p},\overline{p}]} \widehat{L}_n(p).$$

Then by Corollary 2, when the samples are i.i.d., the true optimal reward R^* satisfies with probability $1-\delta$,

$$R^* \in [\widehat{R}_{\text{rob}} - O(1/n),\ \widehat{R}_{\text{rob}} + O(1/\sqrt{n})].$$

Next, we discuss the radius selection and the performance guarantees in an online setting with N rounds. In Appendix EC.4.2, we verify Assumption 4 for multiple observations using the fact that

$$|W(p;\theta) - W(p;\theta^*)| \le \mu |\phi(X_{nj}^\top \theta) - \phi(X_{nj}^\top \theta^*)|,$$

where $\mu = \max_{p\in[\underline{p},\overline{p}]} \frac{2c\Lambda(p)}{u_2^2 \min_{\theta\in\Theta} 1-\Lambda(p)m(\theta)} + \max_{\theta\in\Theta, p\in[\underline{p},\overline{p}]} \frac{c\Lambda(p)^2}{u_2^2} \frac{v(\theta)}{(1-\Lambda(p)m(\theta))(1-\Lambda(p)m(\theta^*))}$. Set

$$\rho_n = \sqrt{\frac{320}{7}\frac{\bar{\kappa}_\phi}{\underline{\kappa}_\phi} d\eta\sigma \log\Big(\sum_{i=1}^{n-1} M_i\Big)\log(dn/\delta)}.$$

Applying results in generalized lienar models (Appendix A), by Proposition 1, its regret is $\tilde{O}(d\sqrt{N})$; a detailed analysis is provided in in Appendix EC.4.3.

5.2. Rank-One Matrix Estimation

In this subsection, we consider a rank-one matrix estimation problem. Let $\mathcal{X}, \mathcal{A} \subset \mathbb{R}^d$. Suppose there exists $\Gamma_* \in \mathbb{R}^{d\times d}$ such that the reward R_i satisfies

$$R_i = Y_i = A_i^\top \Gamma_* X_i + \epsilon_i,$$

where Γ_* is an (unknown) underlying rank-one matrix; and ϵ_i is the noise that satisfies Assumption 1 or 1'. This problem finds numerous applications in e-commerce, recommendation system, choice modeling, etc. [82, 65, 44]. Consider the following square loss minimization

$$\min_{\substack{\Gamma\in\mathbb{R}^{d\times d}\\ \text{rank}(\Gamma)\le 1}} \sum_{i=1}^{n-1} (Y_i - A_i^\top \Gamma X_i)^2.$$

One classical way to deal with the rank-one constraint $\text{rank}(\Gamma)\le 1$ is to re-parameterize the matrix Γ as $\Gamma = \theta\theta^\top$, where $\theta\in\mathbb{R}^d$, and the loss function above becomes

$$\sum_{i=1}^{n-1} (Y_i - A_i^\top \theta\theta^\top X_i)^2.$$

Suppose there exists $\theta^* \in \Theta = [\underline{\theta}, \overline{\theta}]^d$ such that Γ_* admits a non-negative matrix factorization $= \theta^*\theta^{*\top}$. We estimate $\hat{\theta}_n$ using the following rank-one matrix estimation

$$\hat{\theta}_n \in \arg\min_{\theta\in\mathbb{R}^d} \sum_{i=1}^{n-1} (Y_i - A_i^\top \theta\theta^\top X_i)^2. \tag{20}$$

Although being non-convex (as shown in Appendix EC.5), the global minimizer $\hat{\theta}_n$ can be computed efficiently using stochastic gradient descent (e.g., [47]). Assume the decision set $\mathcal{A}_n$ and the set of contexts $\mathcal{X}_n$ are bounded by 1 in Euclidean norm. Then Assumption 2 or 2' holds.

In the offline setting, we are given a fixed set of observations $\{(Y_i, X_i, A_i)\}_{i=1}^{n-1}$. Set $f_\theta(x,a) = x^\top \theta \theta^\top a$. To verify Assumption 3, recall that we have assumed $\hat{\theta}_n \in \Theta$ for all n. Then from the mean value theorem, we have that

$$\begin{aligned}\big(f_{\hat{\theta}_n}(x,a) - f_{\theta^*}(x,a)\big)\nabla f_{\hat{\theta}_n}(x,a)^\top(\hat{\theta}_n - \theta^*) &= (\hat{\theta}_n - \theta^*)^\top \nabla f_{\tilde{\theta}_n}(x,a)\nabla f_{\hat{\theta}_n}(x,a)(\hat{\theta}_n - \theta^*) \\ &= (\hat{\theta}_n - \theta^*)^\top (ax^\top + xa^\top)\tilde{\theta}_n\hat{\theta}_n^\top(ax^\top + xa^\top)(\hat{\theta}_n - \theta^*),\end{aligned}$$

where $\tilde{\theta}_n$ is a convex combination of $\hat{\theta}_n$ and θ^*. Observe from the boundedness of $\hat{\theta}_n$ that $\theta^*\hat{\theta}_n^\top \geq \underline{\theta}/\overline{\theta} \cdot \hat{\theta}_n\hat{\theta}_n^\top$. Hence $\tilde{\theta}_n\hat{\theta}_n^\top \geq \underline{\theta}/\overline{\theta} \cdot \hat{\theta}_n\hat{\theta}_n^\top$, and it follows that

$$\begin{aligned}\big(f_{\hat{\theta}_n}(x,a) - f_{\theta^*}(x,a)\big)\nabla f_{\hat{\theta}_n}(x,a)^\top(\hat{\theta}_n - \theta^*) &\geq \underline{\theta}/\overline{\theta}(\hat{\theta}_n - \theta^*)^\top (ax^\top + xa^\top)\hat{\theta}_n\hat{\theta}_n^\top(ax^\top + xa^\top)(\hat{\theta}_n - \theta^*) \\ &= \underline{\theta}/\overline{\theta} \cdot (\nabla f_\theta(x,a)^\top(\hat{\theta}_n - \theta^*))^2,\end{aligned}$$

which verifies Assumption 3. Define

$$\hat{V}_n = 2\sum_{i=1}^{n-1}(Y_i - A_i^\top\hat{\theta}_n\hat{\theta}_n^\top X_i)^2(\hat{\theta}_n^\top X_i A_i + A_i^\top\hat{\theta}_n X_i)(\hat{\theta}_n^\top X_i A_i + A_i^\top\hat{\theta}_n X_i)^\top,$$

and consider the uncertainty set $\|\theta - \hat{\theta}_n\|_{\hat{V}_n} \leq \rho_n$ with

$$\rho_n = \frac{4\sigma\kappa_f\sqrt{\log(5\kappa_f)(d + \log(1/\delta))}}{\underline{\theta}/\overline{\theta}},$$

where $\kappa_f = \sup_{x\in\mathcal{X}, a\in\mathcal{A}, \theta\in\Theta}\|\theta^\top ax + x^\top\theta a\|_2$. Let $\tilde{a}$ be the optimal solution to the robust problem

$$\widehat{R}_{\mathsf{rob}} = \max_{a\in\mathcal{A}} r(x,a;\hat{\theta}_n) - \rho_n \cdot \|\nabla r(x,a;\hat{\theta}_n)\|_{\hat{V}_n^{-1}}.$$

Then by Corollary 2, when the samples are i.i.d., the true optimal reward R^* satisfies that with probability $1-\delta$,

$$R^* \in [\widehat{R}_{\mathsf{rob}} - O(1/n),\ \widehat{R}_{\mathsf{rob}} + O(1/\sqrt{n})].$$

In the online setting, let us verify the assumptions required by Theorem 3. Assumptions 1', 2', 3 hold in a similar fashion as in the offline setting. Assumption 4 holds trivially with $\mu = 1$ and $\hbar_r = 0$. Assumption 5 can be ensured by the initial exploration as discussed after Theorem 3. Therefore, by Theorem 3, setting

$$\rho_n = \frac{35\sigma\eta d\sqrt{2\log(n)\log(\varsigma nd\log(2n/\delta)/\delta)}}{\underline{\theta}/\overline{\theta}},$$

and applying Algorithm 1 with would yield $\tilde{O}(d^{3/2}\sqrt{N})$ regret. Rank-one and low-rank bandits have been studied recently [48, 49, 52, 43, 66]. Among them, Katariya et al. [48, 49], Kveton et al. [52] did not consider features and restrict the decision space so that only certain entries in the matrix Γ^* are selected in each round. Our result is mostly related to [43, 66] and yield the same regret bound. The LowLOC and LowGLOC algorithms proposed in [66] are UCB type algorithms, but they are not efficiently implementable because of the exponentially weighted average forecaster for solving the online low-rank linear prediction problem. The ESTR algorithm in [43] and LowESTR algorithm in [66] are "explore-then-commit" type algorithms that use a novel way of exploiting the subspace so as to reduce the problem to linear bandits after exploration. LowESTR improves on ESTR by solving an NP-hard sub-problem in the exploration stage, but relies on an additional assumption on the decision set. Comparing to these works, our algorithm is conceptually easier, efficient to implement, and does not restrict the decision space.

6. Concluding Remarks

In this paper, we study contextual optimization with parametric uncertainty. We propose a new parameter uncertainty set based on a generic supervised-learning oracle. We investigate its computational tractability and statistical properties in offline and online settings. Our analysis inspires a new data-driven optimistic optimization framework for online contextual decision-making. For future work, it would be intriguing to see how our design principle can be applied to other areas such as offline and online reinforcement learning.

Appendix A: Results of DOO for Generalized Linear Models

In this appendix, we derive the regret bound for the generalized linear model considered in Example 3. The main purpose of this section is to demonstrate the treatment for (DOO) in which (6) has coefficients $\omega_{n,i}$ not equal to 1, since in the main body we mainly worked with ($\widehat{\text{DOO}}$) with $\omega_{n,i} = 1$. Proofs for this subsection are given Appendix EC.6.

The following assumption specializes Assumption 2' for generalized linear bandits and are standard in the literature [33, 59].

Assumption 2". *The following holds:*

(i) $\beta_{\mathcal{A}} := \sup_{a\in\mathcal{A}} \|a\|_2 < \infty$.

(ii) $\hbar_\phi := \sup_{a\in\mathcal{A},\theta\in\Theta} \|\phi''(\theta^\top a)\|_2 < \infty$.

(iii) $\underline{\kappa}_\phi := \inf_{a\in\mathcal{A},\theta\in\Theta} \phi'(\theta^\top a) > 0$, *and* $\bar{\kappa}_\phi := \sup_{a\in\mathcal{A},\theta\in\Theta} \phi'(\theta^\top a) < \infty$.

(iv) $\beta_\Theta := \sup_{\theta\in\Theta} \|\theta\|_2 < \infty$.

(v) $0 \le \phi(\theta^\top a) \le \beta_\phi < \infty$, $\forall a \in \mathcal{A}, \theta \in \Theta$.

Assumption 2"(iii) implies the monotonicity of ϕ. Hence, (DOO) is tractable by simply dropping ϕ in the objective. In what follows, we simply assume Θ is compact without loss of generality, since the estimator converges to the neighborhood of the true parameter (Lemma 4 in [53]).

Consistent with the notations in Example 3, we set $f_\theta(a) = \theta^\top a$, $\omega_{n,i} = \sqrt{\phi'(A_i^\top \hat{\theta}_n)}$, $\xi_n = \sum_{i=1}^{n-1} \epsilon_i A_i$, $V_{n,*} = \sum_{i=1}^{n-1} \phi'(A_i^\top \theta^*) A_i A_i^\top$, $\hat{V}_n = \sum_{i=1}^{n-1} \omega_{n,i}^2 A_i A_i^\top$. Similar to (16), we can bound one-step regret by

$$\mathsf{reg}_n \le \phi(A_n^\top \tilde{\theta}_n) - \phi(A_n^\top \theta^*) = \phi'(A_n^\top \bar{\theta}_n)^\top A_n (\tilde{\theta}_n - \theta^*) \le \sqrt{\bar{\kappa}_\phi} \cdot \|\sqrt{\phi'(A_n^\top \bar{\theta}_n)} A_n\|_{\hat{V}_n^{-1}} \|\tilde{\theta}_n - \theta^*\|_{\hat{V}_n},$$

where $\bar{\theta}_n$ is an intermediate value on the line segment connecting $\tilde{\theta}_n$ and θ^*. Note that here we consider $\|\cdot\|_{\hat{V}_n}$ instead of $\|\cdot\|_{V_{n,*}}$ in order to achieve a regret with better constants. Also note that $\omega_{n,i} \le \bar{\kappa}_\phi$ for all n and i. We apply Algorithm 1 by setting the radius

$$\rho_n = \sqrt{\frac{320}{7} \frac{\bar{\kappa}_\phi}{\underline{\kappa}_\phi} d\eta\sigma \log(n) \log(dN/\delta)}.$$

Compared with (14), the radius ρ_n has an additional constant $\bar{\kappa}_\phi$, resulting from the upper bound on $\omega_{n,i}$; meanwhile, ρ_n shrinks from $O(d)$ to $O(\sqrt{d})$, since we avoid the covering number argument in Lemma 4(II) by exploiting the fact that $\|\cdot\|_{V_n(\theta)}$ can be bounded by $\|\cdot\|_{V_n(\theta')}$ for any $\theta, \theta' \in \Theta$; see Appendix EC.6.1. Define

$$\bar{\lambda}_n = \max\left(\left(16\eta^2\sigma^2 \bar{\kappa}_\phi \hbar \beta_{\mathcal{A}} / \underline{\kappa}_\phi^3\right)^2 d\log(N)\log(dN/\delta),\ \left(4\beta_{\mathcal{A}}\hbar/\underline{\kappa}_\phi^2\right)^2 \frac{320}{7} \frac{\bar{\kappa}_\phi}{\underline{\kappa}_\phi^3} d\eta^2\sigma^2 \log(n)\log(dN/\delta) \right),$$

$$\mathcal{E}_n = \left\{ \lambda_{\min}\Big(\sum_{i=1}^{n-1} A_i A_i^\top\Big) \ge \bar{\lambda}_n \right\},\ \text{and}\ \mathcal{E}_n^V = \{\|\hat{\theta}_n - \theta^*\|_{\hat{V}_n} \le \rho_n\}.$$

The counterpart results of Theorem 2 and Lemma 6 are given as follows.

Lemma 7. *Assume Assumptions 1' and 2" are in force. Then with probability at least* $1-\delta$, *for every* $\tau_1 \le n \le N$, *it holds that*

$$\|\hat{\theta}_n - \theta^*\|_{\hat{V}_n}^2 \mathbf{1}(\mathcal{E}_n) \le \rho_n^2.$$

LEMMA 8. *Assume Assumption 2" is in force. Then for any $\tau_1 \le \tau \le N$ it holds that*

$$\sum_{n=\tau}^{N} \left(1 \wedge \|\sqrt{\phi'(A_n^\top \hat{\theta}_n)} A_n\|^2_{\tilde{V}_n^{-1}}\right) \mathbf{1}(\mathcal{E}_n^V \cap \mathcal{E}_n) \le \frac{9}{2} d\bar{\kappa}_\phi \log\left(\frac{N\bar{\kappa}_\phi^2 \beta_{\mathcal{A}}^2}{d \det(V_{\tau,*})^{1/d}}\right).$$

Lemma 7 bounds the radius from above whenever the event $\mathcal{E}_n$ holds, which can be ensured using the exploration methods described in Corollary 3 with $\tau_1 = O(d \log^2 N)$. Lemma 8 is a variant of the elliptical potential lemma for the case that $\omega_{n,i}$ not equal to 1.

PROPOSITION 1. *Assume Assumption 1' and 2" is in force. Then with probability at least $1 - 2\delta$, we have*

$$\mathrm{Regret}_N \mathbf{1}(\cap_{n=\tau}^N \mathcal{E}_n) \le \beta_r \tau + \frac{15}{2\sqrt{2}} \rho_N \beta_r \sqrt{N d \log\left(\frac{N\bar{\kappa}_\phi^2 \beta_{\mathcal{A}}^2}{d \underline{\kappa}_\phi^2 \lambda_{\min}(V_{\tau,*})}\right)}.$$

The regret bound is $\tilde{O}(d\sqrt{N})$, which matches the optimal regret for GLM [33, 59]. As pointed out by Lattimore and Szepesvári [57, Section 19.4.7], the bound developed in [33] depends on the problem parameters $\bar{\kappa}_\phi$ and $\frac{\underline{\kappa}_\phi}{\bar{\kappa}_\phi}$ in an unpleasant manner, here our regret bound improves the dependence on the function parameters from $\frac{\bar{\kappa}_\phi}{\underline{\kappa}_\phi}$ to $\sqrt{\frac{\bar{\kappa}_\phi}{\underline{\kappa}_\phi}}$.

Finally, we remark that GLM models satisfy Assumptions 3-5. Though we did not explicitly state them in the results above, their proofs implicitly exploits these properties. To verify Assumption 3, observe that

$$\big(\phi(a^\top\theta) - \phi(a^\top\theta^*)\big) a^\top(\theta - \theta^*) = \phi'(a^\top\bar{\theta})(\theta - \theta^*) a a^\top (\theta - \theta^*) \ge \underline{\kappa}_\phi \big(a^\top(\theta - \theta^*)\big)^2.$$

Assumption 4 holds trivially because the reward function and the response function coincides. To verify Assumption 5, recall that $f_\theta = \theta^\top a$, thereby

$$(\vartheta^\top(\nabla f_\theta(x,a) - \nabla f_{\theta^*}(x,a)))^2 = 0,$$

hence Assumption 5 holds trivially with $\alpha_f = 0$, $\tau_2 = 1$ and $\gamma_n = \infty$.

References

[1] Abbasi-Yadkori Y, Pál D, Szepesvári C (2011) Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems*, 2312–2320.

[2] Abbasi-Yadkori Y, Pál D, Szepesvári C (2011) Online least squares estimation with self-normalized processes: An application to bandit problems. *arXiv preprint arXiv:1102.2670* .

[3] Agrawal S, Avadhanula V, Goyal V, Zeevi A (2019) Mnl-bandit: A dynamic learning approach to assortment selection. *Operations Research* 67(5):1453–1485.

[4] Agrawal S, Devanur NR (2014) Bandits with concave rewards and convex knapsacks. *Proceedings of the fifteenth ACM conference on Economics and computation*, 989–1006.

[5] Agrawal S, Devanur NR (2019) Bandits with global convex constraints and objective. *Operations Research* 67(5):1486–1502.

[6] Anh LQ, Duy TQ, Hien DV (2020) Well-posedness for the optimistic counterpart of uncertain vector optimization problems. *Annals of Operations Research* 295(2):517–533.

[7] Auer P (2002) Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research* 3(Nov):397–422.

[8] Auer P, Cesa-Bianchi N, Fischer P (2002) Finite-time analysis of the multiarmed bandit problem. *Machine learning* 47(2-3):235–256.

[9] Auer P, Cesa-Bianchi N, Freund Y, Schapire RE (2002) The nonstochastic multiarmed bandit problem. *SIAM journal on computing* 32(1):48–77.

[10] Ban GY, El Karoui N, Lim AE (2018) Machine learning and portfolio optimization. *Management Science* 64(3):1136–1154.

[11] Ban GY, Keskin NB (2020) Personalized dynamic pricing with machine learning: High dimensional features and heterogeneous elasticity. *Forthcoming, Management Science* .

[12] Ban GY, Rudin C (2019) The big data newsvendor: Practical insights from machine learning. *Operations Research* 67(1):90–108.

[13] Ben-Tal A, El Ghaoui L, Nemirovski A (2009) *Robust optimization* (Princeton university press).

[14] Ben-Tal A, Nemirovski A (1999) Robust solutions of uncertain linear programs. *Operations research letters* 25(1):1–13.

[15] Bertsimas D, Kallus N (2020) From predictive to prescriptive analytics. *Management Science* 66(3):1025–1044.

[16] Bertsimas D, Sim M (2004) The price of robustness. *Operations research* 52(1):35–53.

[17] Bertsimas D, Van Parys B (2021) Bootstrap robust prescriptive analytics. *Mathematical Programming* 1–40.

[18] Besbes O, Zeevi A (2009) Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Operations Research* 57(6):1407–1420.

[19] Caprari E, Baiardi LC, Molho E (2019) Primal worst and dual best in robust vector optimization. *European Journal of Operational Research* 275(3):830–838.

[20] Caro F, Gallien J (2007) Dynamic assortment with demand learning for seasonal consumer goods. *Management Science* 53(2):276–292.

[21] Chan J, Pacchiano A, Tripuraneni N, Song YS, Bartlett P, Jordan MI (2021) Parallelizing contextual linear bandits. *arXiv preprint arXiv:2105.10590* .

[22] Chen X, Wang Y, Zhou Y (2020) Dynamic assortment optimization with changing contextual information. *J. Mach. Learn. Res.* 21:216–1.

[23] Chu W, Li L, Reyzin L, Schapire R (2011) Contextual bandits with linear payoff functions. *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 208–214.

[24] Dani V, Hayes TP, Kakade SM (2008) Stochastic linear optimization under bandit feedback. *Conference on Learning Theory* .

[25] de la Pena VH, Klass MJ, Lai TL (2004) Self-normalized processes: exponential inequalities, moment bounds and iterated logarithm laws. *Annals of probability* 1902–1933.

[26] den Boer AV, Zwart B (2015) Dynamic pricing and learning with finite inventories. *Operations research* 63(4):965–978.

[27] Diao S, Sen S (2020) Distribution-free algorithms for learning enabled optimization with non-parametric estimation. *Management Science* 66(3):1025–1044.

[28] Donti PL, Amos B, Kolter JZ (2017) Task-based end-to-end model learning in stochastic optimization. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 5490–5500.

[29] Elmachtoub AN, Grigas P (2021) Smart "predict, then optimize". *Management Science* .

[30] Estes A (2021) Slow rates of convergence in optimization with side information. *Available at SSRN 3803427* .

[31] Estes A, Richard JP (2019) Objective-aligned regression for two-stage linear programs. *Available at SSRN 3469897* .

[32] Filippi S, Cappé O, Garivier A (2010) Optimism in reinforcement learning and kullback-leibler divergence. *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 115–122 (IEEE).

[33] Filippi S, Cappe O, Garivier A, Szepesvári C (2010) Parametric bandits: The generalized linear case. *Advances in Neural Information Processing Systems*, 586–594.

[34] Foster DJ, Agarwal A, Dudík M, Luo H, Schapire RE (2018) Practical contextual bandits with regression oracles. *arXiv preprint arXiv:1803.01088* .

[35] Foster DJ, Rakhlin A (2020) Beyond ucb: Optimal and efficient contextual bandits with regression oracles. *arXiv preprint arXiv:2002.04926* .

[36] Gotoh Jy, Kim MJ, Lim A (2021) A data-driven approach to beating saa out-of-sample. *Available at SSRN 3853493* .

[37] Gross D (2008) *Fundamentals of queueing theory* (John Wiley & Sons).

[38] Ho-Nguyen N, Kılınç-Karzan F (2020) Risk guarantees for end-to-end prediction and optimization processes. *arXiv preprint arXiv:2012.15046* .

[39] Hu Y, Kallus N, Mao X (2020) Fast rates for contextual linear optimization. *arXiv preprint arXiv:2011.03030* .

[40] Hu Y, Kallus N, Mao X (2020) Smooth contextual bandits: Bridging the parametric and non-differentiable regret regimes. *Conference on Learning Theory*, 2007–2010.

[41] Javanmard A, Nazerzadeh H (2019) Dynamic pricing in high-dimensions. *The Journal of Machine Learning Research* 20(1):315–363.

[42] Jiang N, Xie W (2021) Distributionally favorable optimization: A generic framework for data-driven decision-making with outliers. *Virginia Tech ISE Tech Report* .

[43] Jun KS, Willett R, Wright S, Nowak R (2019) Bilinear bandits with low-rank structure. *International Conference on Machine Learning*, 3163–3172 (PMLR).

[44] Kallus N, Udell M (2020) Dynamic assortment personalization in high dimensions. *Operations Research* .

[45] Kannan R, Bayraksan G, Luedtke JR (2020) Data-driven sample average approximation with covariate information. *Optimization Online* .

[46] Kannan R, Bayraksan G, Luedtke JR (2020) Residuals-based distributionally robust optimization with covariate information. *arXiv preprint arXiv:2012.01088* .

[47] Karimi H, Nutini J, Schmidt M (2016) Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 795–811 (Springer).

[48] Katariya S, Kveton B, Szepesvári C, Vernade C, Wen Z (2017) Bernoulli rank-1 bandits for click feedback. *arXiv preprint arXiv:1703.06513* .

[49] Katariya S, Kveton B, Szepesvari C, Vernade C, Wen Z (2017) Stochastic rank-1 bandits. *Artificial Intelligence and Statistics*, 392–401 (PMLR).

[50] Keskin NB, Zeevi A (2017) Chasing demand: Learning and earning in a changing environment. *Mathematics of Operations Research* 42(2):277–307.

[51] Kocsis L, Szepesvári C (2006) Bandit based monte-carlo planning. *European conference on machine learning*, 282–293 (Springer).

[52] Kveton B, Szepesvári C, Rao A, Wen Z, Abbasi-Yadkori Y, Muthukrishnan S (2017) Stochastic low-rank bandits. *arXiv preprint arXiv:1712.04644* .

[53] Kveton B, Zaheer M, Szepesvari C, Li L, Ghavamzadeh M, Boutilier C (2020) Randomized exploration in generalized linear bandits. *International Conference on Artificial Intelligence and Statistics*, 2066–2076.

[54] Lai TL, Robbins H (1985) Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics* 6(1):4–22.

[55] Lai TL, Robbins H, Wei CZ (1979) Strong consistency of least squares estimates in multiple regression ii. *Journal of multivariate analysis* 9(3):343–361.

[56] Lai TL, Wei CZ (1982) Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics* 10(1):154–166.

[57] Lattimore T, Szepesvári C (2020) *Bandit algorithms* (Cambridge University Press).

[58] Li L, Chu W, Langford J, Schapire RE (2010) A contextual-bandit approach to personalized news article recommendation. *Proceedings of the 19th international conference on World wide web*, 661–670.

[59] Li L, Lu Y, Zhou D (2017) Provably optimal algorithms for generalized linear contextual bandits. *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2071–2080 (JMLR.org).

[60] Li Y, Xie W, Deng X, Freeman L (2021) Efficient sequential experiment design for generalized linear models. *Virginia Tech ISE Tech Report* .

[61] Liu S, He L, Max Shen ZJ (2021) On-time last-mile delivery: Order assignment with travel-time predictors. *Management Science* 67(7):4095–4119.

[62] Liyanage LH, Shanthikumar JG (2005) A practical inventory control policy using operational statistics. *Operations Research Letters* 33(4):341–348.

[63] Loke GG, Tang Q, Xiao Y (2021) Decision-driven regularization: A blended model for predict-then-optimize. *Available at SSRN 3623006* .

[64] Lu M, Shanthikumar JG, Shen ZJM (2015) Technical note–operational statistics: Properties and the risk-averse case. *Naval Research Logistics (NRL)* 62(3):206–214.

[65] Lu X, Wen Z, Kveton B (2018) Efficient online recommendation via low-rank ensemble sampling. *Proceedings of the 12th ACM Conference on Recommender Systems*, 460–464.

[66] Lu Y, Meisami A, Tewari A (2021) Low-rank generalized linear bandit problems. *International Conference on Artificial Intelligence and Statistics*, 460–468 (PMLR).

[67] Munos R, et al. (2014) From bandits to monte-carlo tree search: The optimistic principle applied to optimization and planning. *Foundations and Trends® in Machine Learning* 7(1):1–129.

[68] Neu G, Pike-Burke C (2020) A unifying view of optimism in episodic reinforcement learning. *Advances in Neural Information Processing Systems* 33.

[69] Nguyen VA, Shafieezadeh-Abadeh S, Yue MC, Kuhn D, Wiesemann W (2019) Calculating optimistic likelihoods using (geodesically) convex optimization. *arXiv preprint arXiv:1910.07817* .

[70] Nguyen VA, Shafieezadeh-Abadeh S, Yue MC, Kuhn D, Wiesemann W (2019) Optimistic distributionally robust optimization for nonparametric likelihood approximation. *NeurIPS*.

[71] Nguyen VA, Si N, Blanchet J (2020) Robust bayesian classification using an optimistic score ratio. *International Conference on Machine Learning*, 7327–7337 (PMLR).

[72] Norton M, Takeda A, Mafusalov A (2017) Optimistic robust optimization with applications to machine learning. *arXiv preprint arXiv:1711.07511* .

[73] Oh Mh, Iyengar G (2021) Multinomial logit contextual bandits: Provable optimality and practicality. *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 9205–9213.

[74] Peña VH, Lai TL, Shao QM (2008) *Self-normalized processes: Limit theory and Statistical Applications* (Springer Science & Business Media).

[75] Perchet V, Rigollet P, et al. (2013) The multi-armed bandit problem with covariates. *The Annals of Statistics* 41(2):693–721.

[76] Petersen KB, Pedersen MS, et al. (2008) The matrix cookbook. *Technical University of Denmark* 7(15):510.

[77] Qiang S, Bayati M (2016) Dynamic pricing with demand covariates. *Available at SSRN 2765257* .

[78] Rigollet P, Zeevi A (2010) Nonparametric bandits with covariates. *arXiv preprint arXiv:1003.1630* .

[79] Rusmevichientong P, Shen ZJM, Shmoys DB (2010) Dynamic assortment optimization with a multinomial logit choice model and capacity constraint. *Operations research* 58(6):1666–1680.

[80] Rusmevichientong P, Tsitsiklis JN (2010) Linearly parameterized bandits. *Mathematics of Operations Research* 35(2):395–411.

[81] Russo D, Van Roy B (2014) Learning to optimize via posterior sampling. *Mathematics of Operations Research* 39(4):1221–1243.

[82] Sedhain S, Menon A, Sanner S, Xie L, Braziunas D (2017) Low-rank linear cold-start recommendation from social data. *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

[83] Sen S, Deng Y (2018) Learning enabled optimization: Towards a fusion of statistical learning and stochastic programming. *Working paper* .

[84] Sim M, Tang Q, Zhou M, Zhu T (2021) The analytics of robust satisficing. *Available at SSRN 3829562* .

[85] Simchi-Levi D, Xu Y (2020) Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *Available at SSRN* .

[86] Slivkins A (2011) Contextual bandits with similarity information. *Proceedings of the 24th annual Conference On Learning Theory*, 679–702 (JMLR Workshop and Conference Proceedings).

[87] Song J, Zhao C (2020) Optimistic distributionally robust policy optimization. *arXiv preprint arXiv:2006.07815* .

[88] Sutter T, Van Parys BP, Kuhn D (2020) A general framework for optimal data-driven optimization. *arXiv preprint arXiv:2010.06606* .

[89] Tropp JA (2011) User-friendly tail bounds for matrix martingales. Technical report, California Institute of Technology.

[90] Tulabandhula T, Rudin C (2013) Machine learning with operational costs. *Journal of Machine Learning Research* 14:1989–2028.

[91] Valko M, Korda N, Munos R, Flaounas I, Cristianini N (2013) Finite-time analysis of kernelised contextual bandits. *arXiv preprint arXiv:1309.6869* .

[92] Van Parys BP, Esfahani PM, Kuhn D (2020) From data to decisions: Distributionally robust optimization is optimal. *Management Science* .

[93] Vershynin R (2010) Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027* .

[94] Xu Y, Zeevi A (2020) Upper counterfactual confidence bounds: a new optimism principle for contextual bandits. *arXiv preprint arXiv:2007.07876* .

[95] Zahavy T, Mannor S (2019) Deep neural linear bandits: Overcoming catastrophic forgetting through likelihood matching. *arXiv preprint arXiv:1901.08612* .

[96] Zhang H, Yin W (2013) Gradient methods for convex minimization: better rates under weaker conditions. *arXiv preprint arXiv:1303.4645* .

[97] Zhen J, Kuhn D, Wiesemann W (2021) Mathematical foundations of robust and distributionally robust optimization. *arXiv preprint arXiv:2105.00760* .

[98] Zhou D, Li L, Gu Q (2020) Neural contextual bandits with ucb-based exploration. *International Conference on Machine Learning*, 11492–11502 (PMLR).

[99] Zhu T, Xie J, Sim M (2021) Joint estimation and robustness optimization. *Management Science* .

Proofs of Statements

Appendix EC.1: Proofs for Section 2

Proof of Lemma 1. We only consider the maximization problem since the result for the minimization problem can be shown similarly by switching the sign of the reward function. Using Lagrangian multipliers $\nu \geq 0, \varphi \in \mathbb{R}^d$, we have

$$\begin{aligned}
&\max_{\theta\in\mathbb{R}^d} \max_{\varepsilon\in\mathbb{R}^{n-1}} \left\{ r_n(X_n,a;\theta) : \sum_{i=1}^{n-1} \psi(Y_i,X_i,A_i,\varepsilon_i;\theta) = 0,\ \sum_{i=1}^{n-1} \varepsilon_i^2/\omega_{n,i}^2 \leq \rho_n^2 \right\} \\
&= \max_{\theta} \left\{ r_n(X_n,a;\theta) + \max_{\varepsilon\in\mathbb{R}^{n-1}} \min_{\nu\geq 0,\varphi\in\mathbb{R}^d} \left\{ \nu\rho_n^2 + \varphi^\top \Big(\sum_{i=1}^{n-1} \psi(Y_i,X_i,A_i,\varepsilon_i;\theta) \Big) - \nu\varepsilon_i^2/\omega_{n,i}^2 \right\} \right\} \\
&\leq \max_{\theta} \left\{ r_n(X_n,a;\theta) + \max_{\varepsilon\in\mathbb{R}^{n-1}} \min_{\nu\geq 0,\varphi\in\mathbb{R}^d} \left\{ \nu\rho_n^2 + \varphi^\top \Big(\sum_{i=1}^{n-1} \psi(Y_i,X_i,A_i,0;\theta) - \psi'(Y_i,X_i,A_i,0;\theta)\varepsilon_i \Big) \|\varphi\|_2 \hbar_\ell \varepsilon_i^2 - \nu\varepsilon_i^2/\omega_{n,i}^2 \right\} \right\}.
\end{aligned}$$

By convex programming duality, the inner max-min problem on the right side equals

$$\begin{aligned}
&\min_{\nu\geq 0,\varphi\in\mathbb{R}^d} \max_{\varepsilon\in\mathbb{R}^{n-1}} \left\{ \nu\rho_n^2 + \varphi^\top \Big(\sum_{i=1}^{n-1} \psi(Y_i,X_i,A_i,0;\theta) - \psi'(Y_i,X_i,A_i,0;\theta)\varepsilon_i \Big) + \|\varphi\|_2 \hbar_\ell \varepsilon_i^2 - \nu\varepsilon_i^2/\omega_{n,i}^2 \right\} \\
&= \min_{\nu\geq 0,\varphi\in\mathbb{R}^d} \left\{ \nu\rho_n^2 + \sum_{i=1}^{n-1} \max_{\varepsilon_i\in\mathbb{R}} \left\{ \varphi^\top \psi(Y_i,X_i,A_i,0;\theta) - \varphi^\top \psi'(Y_i,X_i,A_i,0;\theta)\varepsilon_i - \Big(\nu/\omega_{n,i}^2 - \|\varphi\|_2\hbar_\ell \Big) \varepsilon_i^2 \right\} \right\} \\
&= \min_{\nu\geq 0,\varphi\in\mathbb{R}^d} \left\{ \nu\rho_n^2 + \sum_{i=1}^{n-1} \varphi^\top \psi(Y_i,X_i,A_i,0;\theta) + \sum_{i=1}^{n-1} \max_{\varepsilon_i\in\mathbb{R}} \left\{ \varepsilon_i \varphi^\top \left(\psi'(Y_i,X_i,A_i,0;\theta) \right) - \Big(\nu/\omega_{n,i}^2 - \|\varphi\|_2\hbar_\ell \Big) \varepsilon_i^2 \right\} \right\} \\
&= \min_{\nu\geq 0,\varphi\in\mathbb{R}^d} \left\{ \nu\rho_n^2 + \sum_{i=1}^{n-1} \varphi^\top \psi(Y_i,X_i,A_i,0;\theta) + \frac{1}{4} \sum_{i=1}^{n-1} \frac{\left(\varphi^\top \psi'(Y_i,X_i,A_i,0;\theta) \right)^2}{\nu/\omega_{n,i}^2 - \|\varphi\|_2\hbar_\ell} : \nu/\omega_{n,i}^2 - \|\varphi\|_2\hbar_\ell > 0 \right\}.
\end{aligned}$$

We now prove that optimal φ is bounded by a constant. Note that $\nu/\omega_{n,i}^2 > \|\varphi\|_2\hbar_\ell$, otherwise the optimal value does not exist. Let $\overline{\omega} = \max_{n,i} \omega_{n,i}$. Then the above equation is bounded by

$$\begin{aligned}
0 &\geq \min_{\nu\geq\overline{\omega},\varphi\in\mathbb{R}^d} \left\{ \nu\rho_n^2 + \sum_{i=1}^{n-1} \varphi^\top \psi(Y_i,X_i,A_i,0;\theta) + \frac{1}{4} \sum_{i=1}^{n-1} \frac{\left(\varphi^\top \psi'(Y_i,X_i,A_i,0;\theta) \right)^2}{\nu/\overline{\omega}^2 - \|\varphi\|_2\hbar_\ell} : \nu/\omega_{n,i}^2 - \|\varphi\|_2\hbar_\ell > 0 \right\} \\
&= \min_{\nu\geq\overline{\omega},\varphi\in\mathbb{R}^d} \left\{ \overline{\omega}^2\rho_n^2(\nu/\overline{\omega}^2 - \|\varphi\|_2\hbar_\ell) + \sum_{i=1}^{n-1} \varphi^\top \psi(Y_i,X_i,A_i,0;\theta) + \frac{1}{4} \sum_{i=1}^{n-1} \frac{\left(\varphi^\top \psi'(Y_i,X_i,A_i,0;\theta) \right)^2}{\nu/\overline{\omega}^2 - \|\varphi\|_2\hbar_\ell} \right. \\
&\qquad \left. + \overline{\omega}^2\rho_n^2\|\varphi\|_2\hbar_\ell : \nu/\omega_{n,i}^2 - \|\varphi\|_2\hbar_\ell > 0 \right\} \\
&= \min_{\varphi\in\mathbb{R}^d} \left\{ \overline{\omega}\rho_n\|\varphi\|_{V_n} + \sum_{i=1}^{n-1} \varphi^\top \psi(Y_i,X_i,A_i,0;\theta) + \overline{\omega}^2\rho_n^2\|\varphi\|_2\hbar_\ell \right\},
\end{aligned}$$

where $V_n = \sum_{i=1}^{n-1} \psi'(Y_i, X_i, A_i, 0; \theta)\psi'(Y_i, X_i, A_i, 0; \theta)^\top$. Thus, it is a quadratic function of φ so its minimizer is bounded, say by a constant B_φ. Then the optimization problem is bounded by

$$
\begin{aligned}
&\min_{\nu \geq 0, \varphi \in \mathbb{R}^d} \max_{\varepsilon \in \mathbb{R}^{n-1}} \left\{ \nu \rho_n^2 + \varphi^\top \Big(\sum_{i=1}^{n-1} \psi(Y_i, X_i, A_i, 0; \theta) - \psi'(Y_i, X_i, A_i, 0; \theta)\varepsilon_i \Big) + \|\phi\|_2 \omega_{n,i}^2 \hbar_\ell \varepsilon_i^2 / \omega_{n,i}^2 - \nu \varepsilon_i^2 / \omega_{n,i}^2 \right\} \\
\leq &\min_{\nu \geq B_\varphi \hbar_\ell, \varphi \in \mathbb{R}^d} \max_{\varepsilon \in \mathbb{R}^{n-1}} \left\{ \nu \rho_n^2 + \varphi^\top \Big(\sum_{i=1}^{n-1} \psi(Y_i, X_i, A_i, 0; \theta) - \psi'(Y_i, X_i, A_i, 0; \theta)\varepsilon_i \Big) - \left(\nu - B_\varphi \hbar_\ell\right) \varepsilon_i^2 / \omega_{n,i}^2 \right\} \\
= &\min_{\nu \geq B_\varphi \hbar_\ell, \varphi \in \mathbb{R}^d} \left\{ \nu \rho_n^2 + \sum_{i=1}^{n-1} \varphi^\top \psi(Y_i, X_i, A_i, 0; \theta) + \max_{\varepsilon \in \mathbb{R}^{n-1}} \left\{ \sum_{i=1}^{n-1} \varepsilon_i \varphi^\top \left(\psi'(Y_i, X_i, A_i, 0; \theta)\right) - \left(\nu - B_\varphi \hbar_\ell\right) \varepsilon_i^2 / \omega_{n,i}^2 \right\} \right\} \\
= &\min_{\nu \geq B_\varphi \hbar_\ell, \varphi \in \mathbb{R}^d} \left\{ \nu \rho_n^2 + \sum_{i=1}^{n-1} \varphi^\top \psi(Y_i, X_i, A_i, 0; \theta) + \frac{1}{4} \sum_{i=1}^{n-1} \frac{\omega_{n,i}^2 \left(\varphi^\top \left(\psi'(Y_i, X_i, A_i, 0; \theta)\right)\right)^2}{\nu - B_\varphi \hbar_\ell} \right\} \\
= &\min_{\nu \geq B_\varphi \hbar_\ell, \varphi \in \mathbb{R}^d} \left\{ \nu \rho_n^2 + \varphi^\top \zeta_n(\theta) + \frac{\varphi^\top V_n(\theta) \varphi}{4(\nu - B_\varphi \hbar_\ell)} \right\},
\end{aligned}
$$

where

$$
\zeta_n(\theta) = \sum_{i=1}^{n-1} \psi(Y_i, X_i, A_i, 0; \theta) \quad \text{and} \quad V_n(\theta) = \sum_{i=1}^{n-1} \omega_{n,i}^2 \psi'(Y_i, X_i, A_i, 0; \theta)\psi'(Y_i, X_i, A_i, 0; \theta)^\top.
$$

Since $\zeta_n(\theta) \perp \ker(V_n(\theta))$ because $V_n(\theta)$ is invertible, the minimum over $\varphi \in \mathbb{R}^d$ is finite and equals $-(\nu - B_\varphi \hbar_\ell)\|\zeta_n(\theta)\|^2_{V_n(\theta)^{-1}}$. As a result, the original problem is equivalent to

$$
\begin{aligned}
&\max_{\theta \in \mathbb{R}^d} \left\{ r(X_n, a; \theta) + \min_{\nu \geq B_\varphi \hbar_\ell} \left\{ \nu \rho_n^2 - (\nu - B_\varphi \hbar_\ell) \|\xi_n(\theta)\|^2_{V_n(\theta)^{-1}} \right\} \right\} \\
= \quad &\max_{\theta \in \mathbb{R}^d} \left\{ r(X_n, a; \theta) + \min_{\nu \geq B_\varphi \hbar_\ell} \left\{ (\nu - B_\varphi \hbar_\ell) \rho_n^2 - (\nu - B_\varphi \hbar_\ell) \|\xi_n(\theta)\|^2_{V_n(\theta)^{-1}} \right\} \right\} + B_\varphi \hbar_\ell \rho_n^2 \\
= \quad &\max_{\theta \in \mathbb{R}^d} \left\{ r(X_n, a; \theta) : \ \|\zeta_n(\theta)\|^2_{V_n(\theta)^{-1}} \leq \rho_n^2 \right\} + B_\varphi \hbar_\ell \rho_n^2.
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
&\min_{\nu \geq 0, \varphi \in \mathbb{R}^d} \max_{\varepsilon \in \mathbb{R}^{n-1}} \left\{ \nu \rho_n^2 + \varphi^\top \Big(\sum_{i=1}^{n-1} \psi(Y_i, X_i, A_i, 0; \theta) - \psi'(Y_i, X_i, A_i, 0; \theta)\varepsilon_i + \Delta_i \omega_{n,i}^2 \varepsilon_i^2 / \omega_{n,i}^2 \Big) - \nu \varepsilon_i^2 / \omega_{n,i}^2 \right\} \\
\geq \quad &\max_{\theta \in \mathbb{R}^d} \left\{ r(X_n, a; \theta) : \ \|\zeta_n(\theta)\|^2_{V_n(\theta)^{-1}} \leq \rho_n^2 \right\} - B_\varphi \hbar_\ell \rho_n^2.
\end{aligned}
$$

□

Proof of Lemma 2. Consider the optimization problem

$$
\max_{\theta \in \Theta_n} \ r_n(X_n, a; \hat{\theta}_n) + (\theta - \hat{\theta}_n)^\top \nabla r_n(X_n, a; \hat{\theta}_n).
$$

Observe that for any $\theta \in \Theta_n$, Cauchy–Schwarz inequality gives that

$$
(\theta - \hat{\theta}_n)^\top \nabla r_n(X_n, a; \hat{\theta}_n) \leq \|\theta - \hat{\theta}_n\|_{\bar{V}_n} \|\nabla r_n(X_n, a; \hat{\theta}_n)\|_{\bar{V}_n^{-1}} \leq \rho_n \cdot \|\nabla r_n(X_n, a; \hat{\theta}_n)\|_{\bar{V}_n^{-1}},
$$

where the equality is attained at

$$
\theta - \hat{\theta}_n = \rho_n \cdot \frac{V_n^{-1} \nabla r_n(X_n, a; \hat{\theta}_n)}{\|\nabla r_n(X_n, a; \hat{\theta}_n)\|_{\bar{V}_n^{-1}}}.
$$

It implies that

$$\max_{\theta\in\Theta_n} r_n(X_n,a;\hat\theta_n)+(\theta-\hat\theta_n)^\top\nabla r_n(X_n,a;\hat\theta_n)= r_n(X_n,a;\hat\theta_n)+\rho_n\cdot\|\nabla r_n(X_n,a;\hat\theta_n)\|_{\bar V_n^{-1}}=\widehat{U}_n(a).$$

On the other hand, by the mean value theorem, for any $\theta\in\Theta_n$, there exists $\bar\theta_n\in\Theta_n$ that is a convex combination of θ and $\hat\theta_n$ such that

$$\begin{aligned}
&|r_n(X_n,a;\theta)-(r_n(X_n,a;\hat\theta_n)-(\theta-\hat\theta_n)^\top\nabla r_n(X_n,a;\hat\theta_n))|\\
\le&|r_n(X_n,a;\hat\theta_n)+(\theta-\hat\theta_n)^\top\nabla r_n(X_n,a;\hat\theta_n))+(\theta-\hat\theta_n)^\top\nabla^2 r_n(X_n,a;\bar\theta_n))(\theta-\hat\theta_n)\\
&-(r_n(X_n,a;\hat\theta_n)-(\theta-\hat\theta_n)^\top\nabla r_n(X_n,a;\hat\theta_n))|\\
\le&\hbar_r\|\theta-\hat\theta_n\|_2^2.
\end{aligned}$$

Thus, for every $a\in\mathcal{A}$, the optimal values of the minimization/maximization problems in (UQ) and ($\widehat{\text{UQ}}$) differ by at most $\hbar_r\|\tilde\theta_n-\hat\theta_n\|_2^2$. □

Appendix EC.2: Proofs for Section 3

EC.2.1. Proofs for Section 3.1

Proof of Lemma 3. Let $\mathbb{B}^d_{\kappa_f}$ be a ball with radius κ_f. Then by Vershynin [93, Lemma 5.2 and Lemma 5.3], there exists a $1/2$-covering set $\mathcal{C}_{1/2}$ of $\mathbb{B}^d_{\kappa_f}$ with cardinality bounded by $|\mathcal{C}_{1/2}|\le(5\kappa_f)^d$, such that

$$\sup_{x\in\mathbb{B}^d_{\kappa_f}}\left\langle x,\hat V_n^{-1/2}\hat\xi_n\right\rangle\le 2\max_{x\in\mathcal{C}_{1/2}}\left\langle x,\hat V_n^{-1/2}\hat\xi_n\right\rangle.$$

Since

$$\|\hat\xi_n\|_{\hat V_n^{-1}}=\|\hat V_n^{-1/2}\hat\xi_n\|_2=\sup_{\|x\|_2\le1}\left\langle x,\hat V_n^{-1/2}\hat\xi_n\right\rangle.$$

By union bound, it follows that

$$\mathbb{P}\left\{\|\hat\xi_n\|_{\hat V_n^{-1}}>c\right\}\le\mathbb{P}\left\{\max_{x\in\mathcal{C}_{1/2}}\left\langle x,\hat V_n^{-1/2}\hat\xi_n\right\rangle>c/2\right\}\le\sum_{x\in\mathcal{C}_{1/2}}\mathbb{P}\left\{\left\langle x,\hat V_n^{-1/2}\hat\xi_n\right\rangle>c/2\right\}.$$

Recall that $\hat\xi_n=\sum_{i=1}^{n-1}\epsilon_i\nabla f_{\hat\theta_n}(X_i,A_i)$. Using Hoeffding's inequality, we have

$$\mathbb{P}\left\{\left\langle x,\hat V_n^{-1/2}\hat\xi_n\right\rangle>c/2\right\}\le\exp\left(-\frac{c^2}{8\sigma^2\sum_{i=1}^{n-1}\|x^\top\hat V_n^{-1/2}\nabla f(X_i,A_i)\|_2^2}\right).$$

Since for any $x\in\mathbb{B}_{\kappa_f}$,

$$\sum_{i=1}^{n-1}\|x^\top\hat V_n^{-1/2}\nabla f(X_i,A_i)\|_2^2=\sum_{i=1}^{n-1}x^\top\hat V_n^{-1/2}\nabla f(X_i,A_i)\nabla f(X_i,A_i)^\top\hat V_n^{-1/2}x=\|x^\top\hat V_n^{-1/2}\hat V_n\hat V_n^{-1/2}\|=\|x\|_2^2.$$

Therefore, we have that

$$\mathbb{P}\left\{\|\hat\xi_n\|_{\hat V_n^{-1}}>c\right\}\le(5\kappa_f)^d\exp\left(-\frac{c^2}{8\sigma^2\|x\|_2^2}\right)\le\exp\left(-\frac{c^2}{8\kappa_f^2\sigma^2}+d\log(5\kappa_f)\right).$$

Taking $c=4\kappa_f\sigma\sqrt{\log(5\kappa_f)(d+\log(1/\delta))}$ yields

$$\mathbb{P}\left\{\|\hat\xi_n\|_{\hat V_n^{-1}}>4\kappa_f\sigma\sqrt{\log(5\kappa_f)(d+\log(1/\delta))}\right\}\le\exp\left(-\frac{16\sigma^2\kappa_f^2\log(5\kappa_f)(d+\log(1/\delta))}{8\sigma^2\kappa_f^2}+d\log(5\kappa_f)\right)<\delta.$$

□

EC.2.2. Sufficient Condition for Assumption 3

LEMMA EC.1. *If for every $x \in \mathcal{X}$, $a \in \mathcal{A}$ and $y \in \mathcal{Y}$, ℓ satisfies the restricted secant inequality at θ^*, then Assumption 3 holds.*

Proof of Lemma EC.1. Let us first verify that

$$\nabla \ell(y, \phi \circ \theta(x,a)) = (\phi \circ f_\theta(x,a) - y)\nabla f_\theta(x,a).$$

For the square loss $\ell(y, f_\theta(x,a)) = \frac{1}{2}(y - f_\theta(x,a))^2$ with ϕ being the identity map, we have $\nabla \ell(y, \phi \circ \theta(x,a)) = (f_\theta(x,a) - y)\nabla f_\theta(x,a)$. For the log-likelihood $\ell(y, f_\theta) = -y \log(\phi \circ f_\theta(x,a)) - (1-y)\log(1 - \phi \circ f_\theta(x,a))$, where $\phi(z) = \frac{\exp(z)}{1+\exp(z)}$, we have

$$\nabla \ell(y, f_\theta) = -\frac{y\phi' \nabla f_\theta(x,a)}{\phi \circ f_\theta(x,a)} + \frac{(1-y)\phi' \nabla f_\theta(x,a)}{1 - \phi \circ f_\theta(x,a)}.$$

Since

$$-\frac{y\phi'}{\phi} + \frac{(1-y)\phi'}{1-\phi} = \frac{-y\phi(1-\phi)}{\phi} + \frac{(1-y)\phi(1-\phi)}{1-\phi} = -y + \phi \circ f_\theta(x,a),$$

we calculate that $\nabla \ell(y, f_\theta(x,a)) = (\phi \circ f_\theta(x,a) - y)\nabla f_\theta(x,a)$. Suppose ℓ is $\tilde{\alpha}$-strongly convex in θ, in both cases, one can easily verify that

$$\begin{aligned}
\big(\phi \circ f_\theta(x,a) - \phi \circ f_{\theta^*}(x,a)\big)\nabla f_\theta(x,a)^\top(\theta - \theta^*) &= \big(\nabla \ell(y, f_\theta) - \nabla \ell(y, f_{\theta^*})\big)^\top(\theta - \theta^*)\,|_{y = \phi \circ f^*_\theta(x,a)} \\
&\geq \tilde{\alpha}\|\theta - \theta^*\|_2^2 \\
&\geq \tilde{\alpha}/\kappa_f^2 (\theta - \theta^*)^\top \nabla f_\theta(x,a) \nabla f_\theta(x,a)^\top (\theta - \theta^*).
\end{aligned}$$

□

EC.2.3. Proof of Theorem 1

Proof. Define $\Delta_n = \hat{\theta}_n - \theta^*$. It suffices to show that $\|\hat{\xi}_n\|^2_{\hat{V}_n^{-1}} \geq C\|\Delta_n\|^2_{\hat{V}_N}$ for some properly chosen constant C.

Using the mean value theorem, there exists $\bar{\theta}_n$ which is a convex combination of $\hat{\theta}_n$ and θ^* such that

$$\phi \circ f_{\hat{\theta}_n}(X_n, A_n) - \phi \circ f_{\theta^*}(X_n, A_n) = g_{\bar{\theta}_n}(X_n, A_n)^\top(\hat{\theta}_n - \theta^*),$$

where g is defined in (10). Using the definition of $\hat{\xi}_n$ and the first-order condition (8) that $\hat{\theta}_n$ satisfies, it follows that

$$\begin{aligned}
\hat{\xi}_n &= \sum_{i=1}^{n-1} \epsilon_i \nabla f_{\hat{\theta}_n}(X_i, A_i) \\
&= \sum_{i=1}^{n-1} (Y_i - \phi \circ f_{\theta^*}(X_i, A_i)) \nabla f_{\hat{\theta}_n}(X_i, A_i) \\
&= \sum_{i=1}^{n-1} (\phi \circ f_{\hat{\theta}_n}(X_i, A_i) - \phi \circ f_{\theta^*}(X_i, A_i)) \nabla f_{\hat{\theta}_n}(X_i, A_i) \\
&= (\hat{\theta}_n - \theta^*)^\top \sum_{i=1}^{n-1} g_{\bar{\theta}_i}(X_i, A_i) \nabla f_{\hat{\theta}_n}(X_i, A_i)^\top.
\end{aligned}$$

Define a matrix B as

$$B := \sum_{i=1}^{n-1} g_{\bar{\theta}_i}(X_i, A_i) \nabla f^\top_{\hat{\theta}_n}(X_i, A_i) - \sum_{i=1}^{n-1} \underline{\kappa}_\phi \nabla f_{\hat{\theta}_n}(X_i, A_i) \nabla f_{\hat{\theta}_n}(X_i, A_i)^\top.$$

It follows that

$$\begin{aligned}\|\hat{\xi}_n\|^2_{\hat{V}_n^{-1}} &= \Delta_n^\top \left(\underline{\kappa}_\phi \sum_{i=1}^{n-1} \nabla f_{\hat{\theta}_n}(X_i,A_i)\nabla f_{\hat{\theta}_n}(X_i,A_i)^\top + B\right)\hat{V}_n^{-1}\left(\underline{\kappa}_\phi \sum_{i=1}^{n-1} \nabla f_{\hat{\theta}_n}(X_i,A_i)\nabla f_{\hat{\theta}_n}(X_i,A_i)^\top + B^\top\right)\Delta_n \\ &= \underline{\kappa}_\phi^2 \|\Delta_n\|^2_{\hat{V}_n} + 2\underline{\kappa}_\phi \Delta_n^\top B\Delta_n + \Delta_n^\top B\hat{V}_n^{-1}B^\top\Delta_n.\end{aligned} \tag{EC.1}$$

If $\Delta_n^\top B\Delta_n \geq 0$, since $\Delta_n^\top B\hat{V}_n^{-1}B^\top\Delta_n \geq 0$, we conclude that

$$\|\hat{\xi}_n\|^2_{\hat{V}_n^{-1}} \geq \underline{\kappa}_\phi^2 \|\Delta_n\|^2_{\hat{V}_n},$$

thus we reach the conclusion.

Otherwise if $\Delta_n^\top B\Delta_n < 0$, using Assumption 3 we have

$$(\phi\circ f_{\hat{\theta}_n}(X_i,A_i) - \phi\circ f_{\theta^*}(X_i,A_i))\nabla f_{\hat{\theta}_n}(X_i,A_i)^\top\Delta_n \geq \alpha(\Delta_n^\top \nabla f_{\hat{\theta}_n}(X_i,A_i)^2.$$

Summing s from 1 to $n-1$ yields

$$\sum_{i=1}^{n-1}(\phi\circ f_{\hat{\theta}_n}(X_i,A_i) - \phi\circ f_{\theta^*}(X_i,A_i))\nabla f_{\hat{\theta}_n}(X_i,A_i)^\top\Delta_n \geq \alpha\sum_{i=1}^{n-1}(\nabla f_{\hat{\theta}_n}(X_i,A_i)^\top\Delta_n)^2.$$

It follows that

$$\Delta_n^\top B\Delta_n = \sum_{i=1}^{n-1}(\phi\circ f_{\hat{\theta}_n}(X_i,A_i) - \phi\circ f_{\theta^*}(X_i,A_i))\nabla f_{\hat{\theta}_n}^\top\Delta_n - \underline{\kappa}_\phi\|\Delta_n\|^2_{\hat{V}_n} \geq (\alpha-\underline{\kappa}_\phi)\|\Delta_n\|^2_{\hat{V}_n}.$$

By Cauchy-Schwarz inequality, we have that

$$(\Delta_n^\top B\hat{V}_n^{-1}B^\top\Delta_n)^{\frac{1}{2}} = \|B\Delta_n\|_{\hat{V}_n^{-1}} \geq \frac{|\Delta_n^\top B\Delta_n|}{\|\Delta_n\|_{\hat{V}_n}} \geq (\alpha-\underline{\kappa}_\phi)\|\Delta_n\|_{\hat{V}_n}.$$

Combined with (EC.1), we conclude that

$$\|\hat{\xi}_n\|^2_{\hat{V}_n^{-1}} \geq \underline{\kappa}_\phi^2\|\Delta_n\|^2_{\hat{V}_n} + 2\underline{\kappa}_\phi\Delta_n^\top B\Delta_n + \frac{(\Delta_n^\top B\Delta_n)^2}{\|\Delta_n\|^2_{\hat{V}_n}} = \left(\underline{\kappa}_\phi\|\Delta_n\|_{\hat{V}_n} + \frac{\Delta_n^\top B\Delta_n}{\|\Delta_n\|_{\hat{V}_n}}\right)^2 \geq \alpha^2\|\Delta_n\|^2_{\hat{V}_n}.$$

Thus by Lemma 3, it holds with probability at least $1-\delta$ that

$$\|\Delta_n\|^2_{\hat{V}_n} \leq \frac{1}{\min(\underline{\kappa}_\phi^2,\alpha^2)}\|\hat{\xi}_n\|^2_{\hat{V}_n^{-1}} \leq \frac{16\sigma\log(5\kappa_f)(d+\log(1/\delta))}{\min(\underline{\kappa}_\phi^2,\alpha^2)}.$$

□

EC.2.4. Tightness of the Bound in Theorem 1

Consider a linear model $Y = X^\top\theta^* + \epsilon$. Given a sample $\{(X_i,Y_i)\}_{i=1}^n$, the least-square estimator $\hat{\theta}_n$ equals

$$\hat{\theta}_n - \theta^* = (\sum_{i=1}^n X_iX_i^\top)^{-1}(\sum_{i=1}^n Y_iX_i) - \theta^* = \hat{V}_n^{-1}\left(\sum_{i=1}^n(\epsilon_i + X_i^\top\theta^*)X_i\right) - \theta^* = \hat{V}_n^{-1}\left(\sum_{i=1}^n \epsilon_iX_i\right).$$

Let $W = [X_1,\cdots,X_n]$, and $E = [\epsilon_1;\cdots;\epsilon_n]$. It follows that

$$\|\hat{\theta}_n - \theta^*\|^2_{\hat{V}_n} = \|\hat{V}_n^{-1}W^\top E\|^2_{\hat{V}_n} = E^\top W\hat{V}_n^{-1}W^\top E = \mathrm{tr}(W\hat{V}_n^{-1}W^\top EE^\top).$$

Therefore, using properties of the trace operator, we have

$$\mathbb{E}[\|\hat{\theta}_n - \theta^*\|^2_{\hat{V}_n}] = \mathbb{E}[\mathrm{tr}(W\hat{V}_n^{-1}WEE^\top)] = \mathrm{tr}(W\hat{V}_n^{-1}W\mathbb{E}[EE^\top]) = \mathrm{tr}(W\hat{V}_n^{-1}W^\top\sigma^2I) = \mathrm{tr}(\sigma^2I) = \sigma^2d.$$

We prove the tightness of the bound in Theorem 1 by contradiction. If Theorem 1 gives a tighter bound that $\|\hat{\theta}_n - \theta^*\|^2_{\hat{V}_n} \leq C\cdot d^q$ with high probability for some $q<1$, then $\mathbb{E}[\|\hat{\theta}_n - \theta^*\|^2_{\hat{V}_n}] = O(d^q)$, which is contradictory to the fact that $\mathbb{E}[\|\hat{\theta}_n - \theta^*\|^2_{\hat{V}_n}] = \sigma^2d$.

EC.2.5. Proofs for Section 3.2

EC.2.5.1. Proof of Corollary 3

Proof. Recall that $\tau_1 = \max\{\lceil \frac{\lambda_0}{\underline{\lambda}} \rceil, \frac{2\bar{\lambda}\log(d/\delta)}{\underline{\lambda}\log(e/2)}\}$. Applying in Tropp [89, Theorem 3.1], we can bound $\lambda_{\min}(V_n(\theta))$ for any $\theta \in \Theta$ by

$$\begin{aligned}
&\mathbb{P}\left\{\lambda_{\min}\left(V_{\tau_1}(\theta)\right) \le \frac{1}{2}\underline{\lambda}\tau_0\right\} \\
\le&\mathbb{P}\left\{\lambda_{\min}\left\{V_{\tau_1}(\theta)\right\} \le \frac{1}{2}\underline{\lambda}\tau \text{ and } \lambda_{\min}\left(\sum_{i=1}^{N}\mathbb{E}[\nabla f_\theta(X_i,A_i)\nabla f_\theta(X_i,A_i)^\top|\mathcal{F}_{i-1}]\right) \ge \underline{\lambda}\tau\right\} \\
\le& d\left(\frac{e}{2}\right)^{-\underline{\lambda}\tau_1/(2\bar{\lambda})} \le \delta.
\end{aligned}$$

□

EC.2.5.2. Proof of Lemma 4

The proof of Lemma 4(I) relies on the following result for self-normalized random variable.

LEMMA EC.2 **(COROLLARY 2.2 IN [25])**. *Suppose random variables Z_1 and Z_2 satisfies for all $u \in \mathbb{R}$,*

$$\mathbb{E}\left[\exp\left(uZ_1 - \frac{u^2}{2}Z_2^2\right)\right] \le 1.$$

Then for any $c \ge \sqrt{2}$ and $\epsilon > 0$,

$$\mathbb{P}\left\{|Z_1| \ge c\sqrt{(Z_2^2+\epsilon)\left(1+\frac{1}{2}\log\left(\frac{Z_2^2}{\epsilon}+1\right)\right)}\right\} \le \exp(-c^2/2).$$

Proof of Lemma 4(I). Fix θ. We start by introducing some notations. Let $x \in \mathbb{R}^d$ whose value will be specified later. Define

$$\begin{aligned}
Z_1 &:= x^\top \xi_n(\theta), \\
Z_2 &:= \sqrt{2}\sigma\|x\|_{V_n(\theta)}.
\end{aligned}$$

Recall $\xi_n(\theta) = \sum_{i=1}^{n-1}\epsilon_i\nabla f_\theta(X_i,A_i)$ and $V_n(\theta) = \sum_{i=1}^{n-1}\nabla f_\theta(X_i,A_i)\nabla f_\theta(X_i,A_i)^\top$. For any $u \ge 0$, we have

$$\begin{aligned}
uZ_1 - \frac{u^2}{2}Z_2^2 =& ux^\top\xi_n(\theta) - \frac{u^2x^\top V_n(\theta)x}{2} \\
=& \sum_{i=1}^{n-1} u\epsilon_i\nabla f_\theta(X_i,A_i)^\top x - \frac{u^2}{2}\cdot 2\sigma^2\cdot(\nabla f_\theta(X_i,A_i)^\top x)^2.
\end{aligned}$$

Define $D_i = u\epsilon_i\nabla f_\theta(X_i,A_i)^\top x - u^2\sigma^2(\nabla f_\theta(X_i,A_i)^\top x)^2$. Note that conditioning on $\mathcal{H}_{i-1}$ and θ, the randomness of D_i comes from ϵ_i only. Thus,

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(u\epsilon_i\nabla f_\theta(X_i,A_i)^\top x - u^2\sigma^2(\nabla f_\theta(X_i,A_i)^\top x)^2\right) \mid \mathcal{H}_{i-1}\right] \\
\le& \mathbb{E}\left[\exp\left(u\epsilon_i\nabla f_\theta(X_i,A_i)^\top x - u^2\sigma^2(\nabla f_\theta(X_i,A_i)^\top x)^2\right) \mid \mathcal{H}_{i-1}\right] \\
\le& \max\left\{1, \exp\left(u^2\sigma^2(\nabla f_\theta(X_i,A_i)^\top x)^2 - u^2\sigma^2(\nabla f_\theta(X_i,A_i)^\top x)^2\right)\right\} = 1.
\end{aligned}$$

It follows that

$$\mathbb{E}[\exp(D_i) \mid \mathcal{H}_{i-1}] \le 1.$$

Using the tower property of conditional expectations and the inequality above, we obtain that

$$\mathbb{E}\Big[\prod_{i=1}^{n}\exp(D_i)\Big]=\mathbb{E}\Big[\prod_{i=1}^{n-1}\exp(D_i)\mathbb{E}[\exp(D_{n-1}\mid\mathcal{H}_{n-1}]\Big]\le\mathbb{E}\Big[\prod_{i=1}^{n-1}\exp(D_i)\Big].$$

Applying this inequality recursively yields

$$\mathbb{E}\big[\exp(uZ_1-\frac{u^2}{2}Z_2^2)\big]\le\mathbb{E}\Big[\prod_{i=1}^{n-1}\exp(D_i)\Big]\le\cdots\le\mathbb{E}[\exp(D_1)]\le 1.$$

The derivation above verifies the condition required by Lemma EC.2. Set $u=2\sigma^2\lambda_{\min}(V_\tau(\theta))\|x\|_2^2$ in Lemma EC.2, then for any $0<\delta\le 1/e$ and $n\ge 1$, with probability $1-\delta$,

$$|x^\top\xi_n(\theta)|\le\sqrt{2}\sqrt{(2\sigma^2\|x\|_{V_n(\theta)}^2+2\sigma^2\lambda_{\min}(V_\tau(\theta))\|x\|_2^2)\Big(1+\tfrac{1}{2}\log\big(1+\tfrac{\|x\|_{V_n(\theta)}^2}{\lambda_{\min}(V_\tau(\theta))\|x\|_2^2}\big)\Big)}\sqrt{2\log(1/\delta)}. \tag{EC.2}$$

Note that for $n\ge\max(d,2)$, $\lambda_{\min}(V_\tau(\theta))\|x\|_2^2\le\|x\|_{V_n(\theta)}^2\le n\|x\|_2^2\beta_g^2$, we have $\|x\|_{V_n(\theta)}^2+\lambda_{\min}(V_\tau(\theta))\|x\|_2^2\le 2\|x\|_{V_n(\theta)}^2$ and $1+\frac{1}{2}\log\Big(1+\frac{\|x\|_{V_n(\theta)}^2}{\lambda_{\min}(V_\tau(\theta))}\Big)\le 1+\frac{1}{2}\log(1+\frac{n\beta_g^2}{\lambda_{\min}(V_\tau(\theta))})\le\eta^2\log(n)/2$ where $\eta=\sqrt{3+2\log(1+2\kappa_f^2/\lambda_{\min}(V_\tau(\theta)))}$. Therefore,

$$\begin{aligned}|x^\top\xi_n(\theta)|&\le 2\sigma\sqrt{2\log(1/\delta)\cdot 2\|x\|_{V_n(\theta)}^2\Big(1+\tfrac{1}{2}\log\big(1+\tfrac{n\kappa_f^2}{\lambda_{\min}(V_\tau(\theta))}\big)\Big)}\\&\le 4\sigma\eta\|x\|_{V_n(\theta)}\sqrt{\log\tfrac{1}{\delta}\log n}.\end{aligned} \tag{EC.3}$$

Now we specify the value of x. Let $x=V_n(\theta)^{-1/2}e_j$. Observe that

$$\begin{aligned}\|\xi_n(\theta)\|_{V_n(\theta)^{-1}}^2&=\xi_n(\theta)^\top V_n(\theta)^{-1}\xi_n(\theta)=\xi_n(\theta)^\top V_n(\theta)^{-1/2}IV_n(\theta)^{-1/2}\xi_n(\theta)\\&=\sum_{j=1}^{d}\xi_n(\theta)^\top V_n(\theta)^{-1/2}e_je_j^\top V_n(\theta)^{-1/2}\xi_n(\theta),\end{aligned}$$

where $\{e_j\}_{j=1}^d$ denotes the standard orthonormal basis in $\mathbb{R}^d$. Thus, for any constant $c>0$ it holds that

$$\begin{aligned}\mathbb{P}\left\{\|\xi_n(\theta)\|_{V_n(\theta)^{-1}}^2\ge dc^2\right\}&=\mathbb{P}\left\{\sum_{j=1}^{d}\xi_n(\theta)^\top V_n(\theta)^{-1/2}e_je_j^\top V_n(\theta)^{-1/2}\xi_n(\theta)\ge dc^2\right\}\\&\le\sum_{j=1}^{d}\mathbb{P}\left\{\xi_n(\theta)^\top V_n(\theta)^{-1/2}e_je_j^\top V_n(\theta)^{-1/2}\xi_n(\theta)\ge c^2\right\}\\&\le\sum_{j=1}^{d}\mathbb{P}\left\{|\xi_n(\theta)^\top V_n(\theta)^{-1/2}e_j|\ge c\right\}.\end{aligned}$$

Set $c=4\eta\sigma\sqrt{\log n\log(d/\delta)}$, $j=1,\dots,d$, in the above inequality, we conclude that

$$\begin{aligned}&\mathbb{P}\big\{\|\xi_n(\theta)\|_{V_n(\theta)^{-1}}^2\ge 16d\eta^2\sigma^2\log(n)\log(d/\delta)\big\}\\&\le\mathbb{E}[\mathbf{1}(\|\xi_n(\theta)\|_{V_n(\theta)^{-1}}^2\ge 16d\eta^2\sigma^2\log(n)\log(d/\delta))]\\&\le\sum_{j=1}^{d}\mathbb{E}\left[\mathbf{1}\left(|\xi_n(\theta)^\top V_n(\theta)^{-1/2}e_j|\ge c\right)\right]\\&=\sum_{j=1}^{d}\mathbb{P}\left\{|\xi_n(\theta)^\top V_n(\theta)^{-1/2}e_j|\ge c\right\}\le\delta.\end{aligned}$$

□

LEMMA EC.3 **(MAXIMUM OF SUB-GAUSSIAN VARIABLES).** *Let $\{\epsilon_i\}_{1\le i\le n}$ be a sequence of zero-mean sub-Gaussian random variables with parameter σ. Define $\epsilon_{\max} = \max_{1\le i\le n}|\epsilon_i|$. Then for every $u \ge 0$, it holds that $\mathbb{P}\{\epsilon_{\max} \ge 2\sqrt{\sigma^2 \log n + u}\} \le 2\exp(-\frac{u^2}{2\sigma^2})$.*

Proof. Let $\epsilon_i = -\epsilon_{i-n}$ for $s = t+1, \cdots, 2n$. Now we have $2n$ Gaussian random variables in total. Since ε_i's are sub-Gaussian, we have

$$\mathbb{P}\left\{\epsilon_{\max} \ge 2\sqrt{\sigma^2\log n} + u\right\} \le \sum_{j=1}^{2n} \mathbb{P}\left\{\epsilon_j \ge 2\sqrt{\sigma^2\log n} + u\right\} \le 2n\exp\left(-\frac{4\sigma^2\log n + u^2}{2\sigma^2}\right) \le 2e^{-u^2/2\sigma^2}.$$

□

Proof of Lemma 4 (II). We use covering number argument to prove this result. Define $H(\theta) = \|\xi_n(\theta)\|^2_{V_n(\theta)^{-1}}$, set $\bar{\epsilon} = 2\sqrt{\sigma^2\log n} + \sqrt{2\sigma^2\log(2/\delta)}$, and define events $\mathcal{B}_n = \{\max_{1\le i\le n}|\epsilon_i| \le \bar{\epsilon}\}$. Lemma EC.3 states that event $\mathcal{B}_n$ holds with probability at least $1-\delta$. We first construct $\omega-$net where $\omega = \log n$. For any θ_1 and θ_2,

$$\begin{aligned}
|H(\theta_1) - H(\theta_2)| &= \left|\|\xi_n(\theta_1)\|^2_{V_n(\theta_1)^{-1}} - \|\xi_n(\theta_2)\|^2_{V_n(\theta_2)^{-1}}\right| \\
&= \left|\|\xi_n(\theta_1)\|^2_{V_n(\theta_1)^{-1}} - \|\xi_n(\theta_1)\|^2_{V_n(\theta_2)^{-1}} + \|\xi_n(\theta_1)\|^2_{V_n(\theta_2)^{-1}} - \|\xi_n(\theta_2)\|^2_{V_n(\theta_2)^{-1}}\right| \\
&\le \left|\|\xi_n(\theta_1)\|^2_{V_n(\theta_1)^{-1}} - \|\xi_n(\theta_1)\|^2_{V_n(\theta_2)^{-1}}\right| + \left|\|\xi_n(\theta_1)\|^2_{V_n(\theta_2)^{-1}} - \|\xi_n(\theta_2)\|^2_{V_n(\theta_2)^{-1}}\right|.
\end{aligned}$$

For ease of notation, let $u_1 = \xi_n(\theta_1)$, $u_2 = \xi_n(\theta_2)$, $V_1 = V_n(\theta_1)$, and $V_2 = V_n(\theta_2)$. We first note that

$$\begin{aligned}
&\left|\|u_1\|^2_{V_2^{-1}} - \|u_2\|^2_{V_2^{-1}}\right| \mathbf{1}(\mathcal{B}_n) \\
=&(u_1-u_2)^\top V_2^{-1}(u_1+u_2)\mathbf{1}(\mathcal{B}_n) \le \|u_1-u_2\|_2 \mathbf{1}(\mathcal{B}_n) \max_{\theta\in\Theta} 2\|V_2^{-1}\xi_n(\theta)\|_2 \\
\le&\hbar_f n\bar{\epsilon}\|\theta_1-\theta_2\|_2 \cdot \frac{2n\kappa_f\bar{\epsilon}}{\lambda_0} = \frac{2\kappa_f\hbar_f(n\bar{\epsilon})^2}{\lambda_0}\|\theta_1-\theta_2\|_2,
\end{aligned}$$

where the last inequality holds because there exists $\bar{\theta}$ which is a convex combination of θ_1 and θ_2 such that $\|u_1-u_2\|_2\mathbf{1}(\mathcal{B}_n) = \|\nabla\xi_n(\bar{\theta})^\top(\theta_1-\theta_2)\|_2\mathbf{1}(\mathcal{B}_n) \le n\hbar_f\bar{\epsilon}\|\theta_1-\theta_2\|_2$ and $\|\xi_n(\theta)\|\mathbf{1}(\mathcal{B}_n) \le n\kappa_f\bar{\epsilon}$. Now we analyze $\left|\|u_1\|^2_{V_1^{-1}} - \|u_1\|^2_{V_2^{-1}}\right|$. Let $W = V_n(\theta)^{-1}$. By chain rule, we have

$$\begin{aligned}
\partial_{\theta_j}(\|u_1\|^2_{V_n(\theta)^{-1}})\mathbf{1}(\mathcal{B}_n) &= \mathrm{tr}\left(\nabla_W(\|u_1\|^2_W)\,|_{W=V_n(\theta)^{-1}} \cdot \partial_{\theta_j}(V_n(\theta)^{-1})\right)\mathbf{1}(\mathcal{B}_n) \\
&= \mathrm{tr}\left(u_1u_1^\top\partial_{\theta_j}(V_n(\theta)^{-1})\right)\mathbf{1}(\mathcal{B}_n),
\end{aligned}$$

where the first equality holds due to Equation (137) in [76]. According to Equation (59) in [76] that

$$\partial_{\theta_j}V_n(\theta)^{-1} = -V_n(\theta)^{-1}\partial_{\theta_j}V_n(\theta)V_n(\theta)^{-1}, \quad j = 1,\ldots,d,$$

we have

$$\begin{aligned}
\left|\partial_{\theta_j}(\|u_1\|^2_{V_n(\theta)^{-1}})\right|\mathbf{1}(\mathcal{B}_n) &= \left|\mathrm{tr}\left(u_1u_1^\top\partial_{\theta_j}(V_n(\theta)^{-1})\right)\right|\mathbf{1}(\mathcal{B}_n) = \left|\mathrm{tr}\left(u_1u_1^\top V_n(\theta)^{-1}\partial_{\theta_j}V_n(\theta)V_n(\theta)^{-1}\right)\right|\mathbf{1}(\mathcal{B}_n) \\
&\le \|u_1\|_2^2\||V_n(\theta)^{-1}\|_2^2\left\|\partial_{\theta_j}\Big(\sum_{i=1}^{n-1}\nabla f_\theta(X_i,A_i)\nabla f_\theta(X_i,A_i)^\top\Big)\right\|_2\mathbf{1}(\mathcal{B}_n) \\
&\le 2\frac{\|u_1\|_2^2}{\lambda_{\min}(V_n(\theta))^2}\cdot\kappa_f\hbar_f n\left\|\partial_{\theta_j}\Big(\sum_{i=1}^{n-1}\nabla f_\theta(X_i,A_i)\nabla f_\theta(X_i,A_i)^\top\Big)\right\|_2.
\end{aligned}$$

It implies that

$$\|\nabla_\theta(\|u_1\|^2_{V_n(\theta)^{-1}})\|_2 \mathbf{1}(\mathcal{B}_n) \le 2\frac{\|u_1\|_2^2}{\lambda_{\min}(V_n(\theta))^2}\cdot \kappa_f \hbar_f n \left\|\nabla_\theta\Big((\sum_{i=1}^{n-1}\nabla f_\theta(X_i,A_i)\nabla f_\theta(X_i,A_i)^\top)\Big)\right\|_2,$$

where $\nabla_\theta(V_n(\theta)^{-1})$ is a $d\times d\times d$ tensor and the expansion along the last dimension has the form $\nabla_\theta(V_n(\theta)^{-1}) = (\partial_{\theta_1}(V_n(\theta)^{-1}),\dots,\partial_{\theta_d}(V_n(\theta)^{-1}))$. Let $\nabla f_\theta(X_i,A_i) = [\nabla f_\theta(X_i,A_i)[1],\cdots,\nabla f_\theta(X_i,A_i)[d]]$. Note that

$$\begin{aligned}&\nabla_\theta(\nabla f_\theta(X_i,A_i)\nabla f_\theta(X_i,A_i)^\top)_{k,l} = \nabla_\theta(\nabla f_\theta(X_i,A_i)[k]\nabla f_\theta(X_i,A_i)[l])\\ =&\nabla^2 f_\theta(X_i,A_i)[k]\nabla f_\theta(X_i,A_i)[l] + \nabla f_\theta(X_i,A_i)[k]\nabla^2 f_\theta(X_i,A_i)[l],\end{aligned}$$

which implies that

$$\left\|\nabla_\theta\Big((\sum_{i=1}^{n-1}\nabla f_\theta(X_i,A_i)\nabla f_\theta(X_i,A_i)^\top)\Big)\right\|_2 \le 2nd^2\kappa_f\hbar_f.$$

Thus, we can bound $\left\|\nabla_\theta(\|u_1\|^2_{V_n(\theta)^{-1}})\right\|_2$ by

$$\left\|\nabla_\theta(\|u_1\|^2_{V_n(\theta)^{-1}})\right\|_2 \mathbf{1}(\mathcal{B}_n) \le 2\frac{(n\bar\epsilon\kappa_f)^2\cdot\kappa_f\hbar_f n d^2}{\lambda_0^2} = \frac{2\bar\epsilon^2\kappa_f^3\hbar_f n^3 d^2}{\lambda_0^2}.$$

Then,

$$\begin{aligned}|H(\theta_1)-H(\theta_2)|\mathbf{1}(\mathcal{B}_n) &\le \left(\frac{2\kappa_f\hbar_f n^2\bar\epsilon^2}{\lambda_0} + \frac{2\bar\epsilon^2\hbar_f^3\kappa_f n^3 d^2}{\lambda_0^2}\right)\|\theta_1-\theta_2\|_2\\ &\le L\|\theta_1-\theta_2\|_2,\end{aligned}$$

where $L = \frac{2\kappa_f\hbar_f n^2\bar\epsilon^2}{\lambda_0} + \frac{2\bar\epsilon^2\hbar_f^3\kappa_f n^3 d^2}{\lambda_0^2} = \frac{2\bar\epsilon^2\kappa_f\hbar_f n^3 d^2}{\lambda_0}(1+\frac{\hbar_f}{\lambda_0})$.

From Lemma 4(I), we conclude that for any $\gamma>0$,

$$\mathbb{P}\{H(\theta)\mathbf{1}(\mathcal{B}_n) > dc\gamma\} \le d\exp(-\gamma),$$

where $c = 16\eta^2\sigma^2\log(n)$. Define $\gamma = 2d\log(L\beta_\Theta/\omega) + \log(d/\delta)$. Let Θ_ω be the $\omega-$covering set for Θ regarding function $H(\cdot)$, i.e., for any $\theta\in\Theta$, there exists $\theta'\in\Theta_\omega$ such that $\|H(\theta)-H(\theta')\|_2\le\omega$. Define $N(\omega;H,\|\cdot\|_2)$ as the $\omega-$covering number. Therefore,

$$\begin{aligned}\mathbb{P}\{H(\hat\theta_n)\mathbf{1}(\mathcal{B}_n) > 2dc\gamma\} &\le \mathbb{P}(\exists\theta, H(\theta)\mathbf{1}(\mathcal{B}_n) > 2dc\gamma)\\ &\le \mathbb{P}(\exists\theta\in\Theta_\omega, H(\theta)\mathbf{1}(\mathcal{B}_n) > dc\gamma)\\ &\le 2N(\epsilon;H,\|\cdot\|_2)d\exp(-\gamma)\\ &\le 2\left(\frac{L\beta_\Theta}{\omega}\right)^d d\exp(-\gamma)\\ &= 2\exp(d\log(L\beta_\Theta/\omega)+\log d-\gamma)\\ &\le 2\exp(-\log(1/\delta)) \le 2\delta.\end{aligned}$$

Since $\mathcal{B}_n$ holds with probability at least $1-\delta$, we conclude that

$$\mathbb{P}\{H(\hat\theta_n) > 2dc\gamma\} \le 3\delta.$$

It implies that

$$
\begin{aligned}
3\delta &\geq \mathbb{P}\{H(\hat{\theta}_n) > 2dc\gamma\} \\
&= \mathbb{P}\left\{H(\hat{\theta}_n) > 32\eta^2\sigma^2\log(n)d\left(2d\log\left(\frac{2\bar{\epsilon}^2\kappa_f\hbar_f n^3 d^2\beta_\Theta}{\omega\lambda_0}\cdot(1+\frac{\hbar_f}{\lambda_0})\right)+\log(d/\delta)\right)\right\} \\
&= \mathbb{P}\left\{H(\hat{\theta}_n) > 32\eta^2\sigma^2\log(t)d\left(36d\log\left(\frac{\beta_\Theta\left(2\sigma+\sigma\sqrt{2\log(2/\delta)}\right)\kappa_f\hbar_f nd}{\lambda_0}\cdot(1+\frac{\hbar_f}{\lambda_0})\right)+\log(d/\delta)\right)\right\} \\
&\geq \mathbb{P}\left\{H(\hat{\theta}_n) > 32\eta^2\sigma^2\log(n)d\left(36d\log\left(\frac{4\beta_\Theta\sigma\kappa_f\hbar_f nd\cdot 2\sqrt{\log(2/\delta)}}{\lambda_0\delta}\cdot(1+\frac{\hbar_f}{\lambda_0})\right)+\log(d/\delta)\right)\right\} \\
&\geq \mathbb{P}\left\{H(\hat{\theta}_n) > 32\eta^2\sigma^2\log(n)d\left(37d\log\left(\frac{8\beta_\Theta\sigma\kappa_f\hbar_f nd\log(2/\delta)}{\lambda_0\delta}\cdot(1+\frac{\hbar_f}{\lambda_0})\right)\right)\right\} \\
&\geq \mathbb{P}\left\{H(\hat{\theta}_n) > (35\eta)^2\sigma^2 d^2\log(n)\log(nd\varsigma\log(2/\delta)/\delta)\right\},
\end{aligned}
$$

where $\varsigma = \frac{8\beta_\Theta\sigma\kappa_f\hbar_f}{\lambda_0}\cdot(1+\frac{\hbar_f}{\lambda_0})$. □

EC.2.5.3. Proof of Theorem 2

Proof of Theorem 2. Similarly to the Proof of Theorem 1, combined with Lemma 4, it holds with probability at least $1-3\delta$ that

$$
\|\Delta_n\|^2_{\hat{V}_n} \leq \frac{1}{\min(\underline{\kappa}^2_\phi,\alpha^2)}\|\hat{\xi}_n\|^2_{\hat{V}_n^{-1}} \leq \frac{(35\eta)^2\sigma^2 d^2\log(n)\log(nd\varsigma\log(2/\delta)/\delta)}{\min(\underline{\kappa}^2_\phi,\alpha^2)}.
$$

□

Appendix EC.3: Proofs for Section 4.2

EC.3.1. Proof of Lemma 5

Define events $\mathcal{E}_n = \{\|\hat{\theta}_n - \theta^*\|_2 \leq \gamma_n\}$.

Proof. Observe that

$$
\begin{aligned}
&\|\vartheta\|^2_{\hat{V}_n}\mathbf{1}(\mathcal{E}_n) \\
&= \vartheta^\top\left(\sum_{i=1}^{n-1}\nabla f_{\hat{\theta}_n}(X_i,A_i)\nabla f_{\hat{\theta}_n}(X_i,A_i)^\top\right)\vartheta\mathbf{1}(\mathcal{E}_n) \\
&= \vartheta^\top\left(\sum_{i=1}^{n-1}\left(\nabla f_{\theta^*}(X_i,A_i)+(\nabla f_{\hat{\theta}_n}(X_i,A_i)-\nabla f_{\theta^*}(X_i,A_i))\right)\left(\nabla f_{\theta^*}(X_i,A_i)+(\nabla f_{\hat{\theta}_n}(X_i,A_i)-\nabla f_{\theta^*}(X_i,A_i))\right)^\top\right) \\
&\quad\cdot\vartheta\mathbf{1}(\mathcal{E}_n) \\
&= \|\vartheta\|^2_{V_{n,*}}\mathbf{1}(\mathcal{E}_n) + 2\vartheta^\top\sum_{i=1}^{n-1}\nabla f_{\theta^*}(X_i,A_i)(\nabla f_{\hat{\theta}_n}(X_i,A_i)-\nabla f_{\theta^*}(X_i,A_i))^\top\vartheta\mathbf{1}(\mathcal{E}_n) \\
&\quad+\vartheta^\top\sum_{i=1}^{n-1}(\nabla f_{\hat{\theta}_n}(X_i,A_i)-\nabla f_{\theta^*}(X_i,A_i))(\nabla f_{\hat{\theta}_n}(X_i,A_i)-\nabla f_{\theta^*}(X_i,A_i))^\top\vartheta\mathbf{1}(\mathcal{E}_n) \\
&\geq \|\vartheta\|^2_{V_{n,*}}\mathbf{1}(\mathcal{E}_n) + 2\vartheta^\top\sum_{i=1}^{n-1}\nabla f_{\theta^*}(X_i,A_i)(\nabla f_{\hat{\theta}_n}(X_i,A_i)-\nabla f_{\theta^*}(X_i,A_i))^\top\vartheta\mathbf{1}(\mathcal{E}_n) \\
&\geq \|\vartheta\|^2_{V_{n,*}}\mathbf{1}(\mathcal{E}_n) - 2\sqrt{\vartheta^\top\sum_{i=1}^{n-1}\nabla f_{\theta^*}(X_i,A_i)\nabla f_{\theta^*}(X_i,A_i)^\top\vartheta}\sqrt{\sum_{i=1}^{n-1}((\nabla f_{\hat{\theta}_n}(X_i,A_i)-\nabla f_{\theta^*}(X_i,A_i))^\top\vartheta)^2\mathbf{1}(\mathcal{E}_n)},
\end{aligned}
$$

where the last inequality holds due to Cauchy-Schwarz inequality. Therefore, by Assumption 5, we have that

$$\|\vartheta\|_{\hat{V}_n}^2 \mathbf{1}\{\|\hat{\theta}_n - \theta^*\|_2 \le \gamma_n\} \ge (1-2\alpha_f)\|\vartheta\|_{V_{n,*}}^2 \mathbf{1}\{\|\hat{\theta}_n - \theta^*\|_2 \le \gamma_n\}.$$

□

EC.3.2. Proof of Lemma 6

Proof. From the definition of $V_{n,*}$, we have

$$\begin{aligned} V_{n+1,*} &= V_{n,*} + \nabla f_{\theta^*}(X_n, A_n)\nabla f_{\theta^*}(X_n, A_n)^\top \\ &= V_{n,*}^{1/2}\Big(I + V_{n,*}^{-1/2}\big(\nabla f_{\theta^*}(X_n, A_n)\nabla f_{\theta^*}(X_n, A_n)^\top\big)V_{n,*}^{-1/2}\Big)V_{n,*}^{1/2}, \end{aligned}$$

which implies that

$$\begin{aligned} \det(V_{n+1,*}) &= \det(V_{n,*})\det\Big(I + V_{n,*}^{-1/2}\big(\nabla f_{\theta^*}(X_n, A_n)\nabla f_{\theta^*}(X_n, A_n)^\top\big)V_{n,*}^{-1/2}\Big) \\ &= \det(V_{n,*})\big(1 + \|\nabla f_{\theta^*}(X_n, A_n)\|_{V_{n,*}^{-1}}^2\big). \end{aligned}$$

Note that

$$\det(V_{n,*}) = \prod_{i=1}^d \lambda_i \le \left(\frac{1}{d}\mathrm{tr}(V_{n,*})\right)^d \le \left(\frac{n\kappa_f^2}{d}\right)^d.$$

Thus using $1 \wedge u \le \log(1+u)$, we can bound the sum by

$$\begin{aligned} &\sum_{n=\tau}^N 1 \wedge \Big(1 + \|\nabla f_{\theta^*}(X_n, A_n)\|_{V_{n,*}^{-1}}^2\Big) \le \sum_{n=\tau}^N 1 \wedge \Big(1 + \|\nabla f_{\theta^*}(X_n, A_n)\|_{V_{n,*}^{-1}}^2\Big) \\ &\le 2d\log\left(\frac{N\kappa_f^2}{d\det(V_{\tau,*})^{1/d}}\right) \le 2d\log\left(\frac{N\kappa_f^2}{d\lambda_{\min}(V_{\tau,*})}\right). \end{aligned}$$

□

EC.3.3. Proof of Theorem 3

LEMMA EC.4 (**CONTROL ON $\|\theta - \theta^*\|_2$**). *When $\lambda_{\min}(\hat{V}_n) \ge \rho_n^2/\gamma_n^2$, for any θ satisfying $\|\theta - \theta^*\|_{\hat{V}_n} \le \rho_n$, it holds that $\|\theta - \theta^*\|_2 \le \gamma_n$.*

Proof. Since $\|\theta - \theta^*\|_{\hat{V}_n}^2 \ge \lambda_{\min}(\hat{V}_n)\|\theta - \theta^*\|_2^2$, we conclude that when $\lambda_{\min}(\hat{V}_n) \ge \rho_n^2/\gamma_n^2$, we have

$$\|\theta - \theta^*\|_2^2 \le \frac{\|\theta - \theta^*\|_{\hat{V}_n}^2}{\lambda_{\min}(\hat{V}_n)} \le \rho_n^2/\lambda_{\min}(\hat{V}_n) \le \gamma_n^2.$$

□

Proof of Theorem 3. Define events

$$\mathcal{E}_n = \{\lambda_{\min}(\hat{V}_n) \ge \rho_n^2/\gamma_n^2\},\ \mathcal{E}_n^V = \{\|\hat{\theta}_n - \theta^*\|_{\hat{V}_n} \le \rho_n\},\ \mathcal{E}_n^\theta = \{\|\hat{\theta}_n - \theta^*\|_2 \le \gamma_n\}.$$

By Lemma 4, $\mathcal{E}_n^V$ holds with probability at least $1 - 3\delta/N$. Let $\tilde{\theta}_n$ be the most optimistic θ obtained from (DOO). Then whenever event $\mathcal{E}_n^V$ holds, we have

$$r(X_n, A_n; \tilde{\theta}_n) = U_n(A_n) \ge U_n(A_n^*) \ge r(X_n, A_n^*; \theta^*) \ge r(X_n, A_n; \theta^*),$$

where the second inequality holds because θ^* satisfies the constraint that $\|\hat{\theta}_n - \theta^*\|_{\hat{V}_n} \le \rho_n$ when $\mathcal{E}_n^V$ holds. As a result, whenever $\mathcal{G}_n$ holds, the one-step regret satisfies

$$r(X_n, A_n^*; \theta^*) - r(X_n, A_n; \theta^*) \le \beta_r \wedge \Big(r(X_n, A_n; \tilde{\theta}_n) - r(X_n, A_n; \theta^*)\Big),$$

where the first inequality holds because the total reward is less than or equal to β_r and the second inequality holds according to Assumption 4. From Lemma 2 that $|\widetilde{U}_n(a) - \widehat{U}_n(a)| \le \hbar_r \|\tilde{\theta}_n - \hat{\theta}_n\|_2^2$, thus if arm A_n is selected by (DOO), then the one-step regret satisfies

$$r(X_n, A_n^*; \theta^*) - r(X_n, A_n; \theta^*) \le \beta_r \wedge \Big(r(X_n, A_n; \tilde{\theta}_n) + \hbar_r \|\tilde{\theta}_n - \hat{\theta}_n\|_2^2 - r(X_n, A_n; \theta^*)\Big) \tag{EC.4}$$

Therefore, we can unify the bounds by

$$\begin{aligned} r(X_n, A_n^*; \theta^*) - r(X_n, A_n; \theta^*) &\le \beta_r \wedge \Big(r(X_n, A_n; \tilde{\theta}_n) - r(X_n, A_n; \theta^*)\Big) \\ &\le \beta_r \wedge \Big(\mu |f_{\tilde{\theta}_n}(X_n, A_n) - f_{\theta^*}(X_n, A_n)| + \hbar \|\tilde{\theta}_n - \theta^*\|_2^2 + \hbar_r \|\tilde{\theta}_n - \hat{\theta}_n\|_2^2\Big) \\ &\le \beta_r \wedge (\mu |f_{\tilde{\theta}_n}(X_n, A_n) - f_{\theta^*}(X_n, A_n)|) + \hbar \|\tilde{\theta}_n - \theta^*\|_2^2 + \hbar_r \|\tilde{\theta}_n - \hat{\theta}_n\|_2^2 \end{aligned} \tag{EC.5}$$

Define an event

$$\tilde{\mathcal{E}}_n^\theta = \{\|\tilde{\theta}_n - \theta^*\|_2 \le \gamma_n\}.$$

By Assumption 5 we have

$$|(\nabla f_{\tilde{\theta}_n}(X_n, A_n) - \nabla f_{\theta^*}(X_n, A_n)))^\top (\tilde{\theta}_n - \theta^*)| \mathbf{1}(\tilde{\mathcal{E}}_n^\theta) \le \alpha_f \sqrt{\frac{1}{n-1} \|\tilde{\Delta}_n\|_{V_{n,*}}^2},$$

where $\tilde{\Delta}_n = \tilde{\theta}_n - \theta^*$. Applying mean value theorem, there exists $\bar{\theta}_n$, which is a convex combination of $\tilde{\theta}_n$ and θ^* such that

$$\begin{aligned} f_{\tilde{\theta}_n}(X_n, A_n) &= f_{\theta^*}(X_n, A_n) + \nabla f_{\bar{\theta}_n}(X_n, A_n))^\top (\tilde{\theta}_n - \theta^*) \\ &= f_{\theta^*}(X_n, A_n) + (\nabla f_{\theta^*}(X_n, A_n) + (\nabla f_{\bar{\theta}_n}(X_n, A_n) - \nabla f_{\theta^*}(X_n, A_n)))^\top (\tilde{\theta}_n - \theta^*). \end{aligned}$$

Summing over n, it follows that

$$\begin{aligned} &\sum_{n=\tau}^N \left| f_{\tilde{\theta}_n}(X_n, A_n) - f_{\theta^*}(X_n, A_n) \right| \mathbf{1}(\tilde{\mathcal{E}}_n^\theta) \\ &\le \sum_{n=\tau}^N \left| \nabla f_{\theta^*}(X_n, A_n)^\top (\tilde{\theta}_n - \theta^*) + \Big(\nabla f_{\bar{\theta}_n}(X_n, A_n) - \nabla f_{\theta^*}(X_n, A_n)\Big)^\top (\tilde{\theta}_n - \theta^*) \right| \mathbf{1}(\tilde{\mathcal{E}}_n^\theta) \\ &\le \sum_{n=\tau}^N \|\nabla f_{\theta^*}(X_n, A_n)\|_{V_{n,*}^{-1}} \|\tilde{\Delta}_n\|_{V_{n,*}} + \sum_{n=\tau}^N \frac{\alpha_f}{\sqrt{n-1}} \|\tilde{\Delta}_n\|_{V_{n,*}}, \end{aligned}$$

where the last inequality holds according to Assumption 5 with $\theta = \bar{\theta}_n$ and $\vartheta = (\tilde{\theta}_n - \theta^*)$. Therefore, using (18) we have

$$\sum_{n=\tau}^N \Big(f_{\tilde{\theta}_n}(X_n, A_n) - f_{\theta^*}(X_n, A_n)\Big) \mathbf{1}(\mathcal{E}_n^V) \mathbf{1}(\tilde{\mathcal{E}}_n^\theta) \mathbf{1}(\mathcal{E}_n^\theta) \le \sum_{n=\tau}^N \frac{2}{1-2\alpha_f} \rho_n \left(\|\nabla f_{\theta^*}(X_n, A_n)\|_{V_{n,*}^{-1}} + \frac{\alpha_f}{\sqrt{n-1}} \right).$$

Thus, we obtain that

$$\begin{aligned} &\sum_{n=\tau}^N \Big(\beta_r \wedge \Big(\mu\big(f_{\tilde{\theta}_n}(X_n, A_n) - f_{\theta^*}(X_n, A_n)\big)\Big)\Big) \mathbf{1}(\mathcal{E}_n^V) \mathbf{1}(\tilde{\mathcal{E}}_n^\theta) \mathbf{1}(\mathcal{E}_n^\theta) \\ &\le \sum_{n=\tau}^N \beta_r \wedge \left(\frac{2}{1-2\alpha_f} \rho_n \mu \left(\|\nabla f_{\theta^*}(X_n, A_n)\|_{V_{n,*}^{-1}} + \frac{\alpha_f}{\sqrt{n-1}} \right) \right) \\ &\le (\beta_r \vee \tfrac{2}{1-2\alpha_f} \mu) \rho_N \sum_{n=\tau}^N \Big(1 \wedge \|\nabla f_{\theta^*}(X_n, A_n)\|_{V_{n,*}^{-1}}\Big) + \frac{2\alpha_f \mu \sqrt{N}}{1-2\alpha_f} \rho_N. \end{aligned} \tag{EC.6}$$

Note that when events $\mathcal{E}_n^V$ and $\mathcal{E}_n$ hold, then it holds that $\|\hat{\theta}_n - \theta^*\|_{\hat{V}_n} \le \rho_n$ and $\|\tilde{\theta}_n - \theta^*\|_{\hat{V}_n} \le \rho_n$. According to Lemma EC.4, event $\mathcal{E}_n \cap \mathcal{E}_n^V$ implies $\tilde{\mathcal{E}}_n^\theta$ and $\mathcal{E}_n^\theta$. Hence, we have

$$\begin{aligned}
\mathsf{Regret}(T)\mathbf{1}(\cap_{n=\tau}^N(\mathcal{E}_n \cap \mathcal{E}_n^V)) &= \beta_r\tau + \sum_{n=\tau+1}^N \left(r(X_n, A_n^*; \theta^*) - r(X_n, A_n; \theta^*)\right)\mathbf{1}(\mathcal{E}_n)\mathbf{1}\left(\mathcal{E}_n^V\right) \\
&\le \beta_r\tau + \sum_{n=\tau+1}^N \left(r(X_n, A_n; \tilde{\theta}_n) - r(X_n, A_n; \theta^*)\right)\mathbf{1}(\mathcal{E}_n)\mathbf{1}\left(\mathcal{E}_n^V\right) \\
&\overset{\text{Lemma EC.4}}{\le} \beta_r\tau + \sum_{n=\tau+1}^N \left(r(X_n, A_n; \tilde{\theta}_n) - r(X_n, A_n; \theta^*)\right)\mathbf{1}(\tilde{\mathcal{E}}_n^\theta)\mathbf{1}(\mathcal{E}_n^\theta)\mathbf{1}\left(\mathcal{E}_n^V\right).
\end{aligned}$$

Therefore, using (EC.5)(EC.6), we can bound Regret_N by

$$\begin{aligned}
&\mathsf{Regret}_N\mathbf{1}(\cap_{n=\tau}^N(\mathcal{E}_n \cap \mathcal{E}_n^V)) \\
&\overset{\text{(EC.5)}}{\le} \beta_r\tau + \sum_{n=\tau+1}^N \left(\beta_r \wedge |\mu(f_{\tilde{\theta}_n}(X_n, A_n) - f_{\theta^*}(X_n, A_n))|\right)\mathbf{1}(\tilde{\mathcal{E}}_n^\theta)\mathbf{1}(\mathcal{E}_n^\theta)\mathbf{1}\left(\mathcal{E}_n^V\right) \\
&\quad + (\hbar + \hbar_r)\sum_{n=\tau+1}^N \left(\|\tilde{\theta}_n - \theta^*\|_2^2 + \|\tilde{\theta}_n - \hat{\theta}_n\|_2^2\right)\mathbf{1}(\tilde{\mathcal{E}}_n^\theta)\mathbf{1}\left(\mathcal{E}_n^V\right) \\
&\overset{\text{(EC.6)}}{\le} \beta_r\tau + (\beta_r \vee \tfrac{2}{1-2\alpha_f}\mu)\rho_N \sum_{n=\tau}^N \left(1 \wedge \|\nabla f_{\theta^*}(X_n, A_n)\|_{V_{n,*}^{-1}}\right)\mathbf{1}(\mathcal{E}_n^V) + \frac{2\alpha_f\mu}{1-2\alpha_f}\rho_N\sqrt{N} \\
&\quad + (\hbar + \hbar_r)\sum_{n=\tau+1}^N \frac{3\rho_n^2}{\lambda_{\min}(\hat{V}_n)}\mathbf{1}(\tilde{\mathcal{E}}_n^\theta)\mathbf{1}\left(\mathcal{E}_n^V\right) \\
&\le \beta_r\tau + (\beta_r \vee \tfrac{2}{1-2\alpha_f}\mu)\rho_N\sqrt{N\sum_{n=\tau}^N \left(1 \wedge \|\nabla f_{\theta^*}(X_n, A_n)\|_{V_{n,*}^{-1}}^2\right)}\mathbf{1}(\mathcal{E}_n^V) + \frac{2\alpha_f\mu}{1-2\alpha_f}\rho_N\sqrt{N} + 3(\hbar + \hbar_r)\sum_{n=\tau+1}^N \gamma_n^2 \\
&\le \beta_r\tau + (\beta_r \vee \tfrac{2}{1-2\alpha_f}\mu)\rho_N\sqrt{Nd\log\left(\frac{N\kappa_f^2}{d\det(V_{\tau,*})^{1/d}}\right)}\mathbf{1}(\mathcal{E}_n^V) + \frac{2\alpha_f\mu}{1-2\alpha_f}\rho_N\sqrt{N} + 3(\hbar + \hbar_r)c\sqrt{N} \\
&\le \beta_r\tau + \frac{3}{1-2\alpha_f}\rho_N(\beta_r + \mu + \hbar c + \hbar_r c)\sqrt{Nd\log\left(\frac{N\kappa_f^2}{d\lambda_0}\right)}.
\end{aligned}$$

Thus, we conclude that

$$\begin{aligned}
&\mathbb{P}\left\{\mathsf{Regret}_N\mathbf{1}(\cap_{n=\tau}^N\mathcal{E}_n) \ge \beta_r\tau + \frac{3}{1-2\alpha_f}\rho_N(\beta_r + \mu + \hbar\gamma_0 + \hbar_r\gamma_0)\sqrt{Nd\log\left(\frac{N\kappa_f^2}{d\lambda_0}\right)}\right\} \\
&\le \mathbb{E}\left[\mathbf{1}\left(\mathsf{Regret}_N\mathbf{1}(\cap_{n=\tau}^N\mathcal{E}_n) \ge \beta_r\tau + \frac{3}{1-2\alpha_f}\rho_N(\beta_r + \mu + \hbar\gamma_0 + \hbar_r\gamma_0)\sqrt{Nd\log\left(\frac{N\kappa_f^2}{d\lambda_0}\right)}\right)\mathbf{1}\left(\bigcap_{n=\tau}^N\mathcal{E}_n^V\right)\right] \\
&\quad + \sum_{n=\tau}^N \mathbb{P}\{(\mathcal{E}_n^V)^c\} \le 3\delta,
\end{aligned}$$

where the last inequality holds according to Theorem 2. □

EC.3.4. Choice of τ_3

Proof. Recall that $\tau_3 = \max\left\{\frac{2\rho_N^2}{\gamma_N^2 \underline{\lambda}}, \frac{2\bar{\lambda}\log(dN/\delta)}{\underline{\lambda}\log(e/2)}\right\}$. Applying in Tropp [89, Theorem 3.1], we can bound $\lambda_{\min}(V_n(\theta))$ for any $\theta \in \Theta$ by

$$\begin{aligned}
\mathbb{P}\left\{\lambda_{\min}\left(V_N(\theta)\right) \le \frac{\rho_N^2}{\gamma_N^2}\right\} &\le \mathbb{P}\left\{\lambda_{\min}\left(V_n(\theta)\right) \le \frac{1}{2}\underline{\lambda}\tau_3\right\} \\
&\le \mathbb{P}\left\{\lambda_{\min}\left(V_N(\theta)\right) \le \frac{1}{2}\underline{\lambda}\tau_3 \text{ and } \lambda_{\min}\left(\sum_{i=1}^{N}\mathbb{E}[\nabla f_\theta(X_i,A_i)\nabla f_\theta(X_i,A_i)^\top|\mathcal{F}_{i-1}]\right) \ge \underline{\lambda}\tau_3\right\} \\
&\le d\left(\frac{e}{2}\right)^{-\underline{\lambda}\tau_3/(2\bar{\lambda})} \le \frac{\delta}{N}.
\end{aligned}$$

□

Appendix EC.4: Proofs for Section 5.1

EC.4.1. Mean and Variance of the Service Time

In the two-stage service system, we let Q_1 and Q_2 be the service time for the first-stage and second-stage, respectively. We define I as the indicator of the presence of the second-stage service. Then the total service time Q is

$$Q = Q_1 + I \cdot Q_2.$$

It implies that the expected service time is

$$\mathbb{E}[Q] = \mathbb{E}[Q_1] + \mathbb{E}[I]\cdot\mathbb{E}[Q_2] = 1/u_1 + \mathbb{E}[\phi({\theta^*}^\top X)]/u_2,$$

and the variance of the service time is

$$\begin{aligned}
\mathbb{V}\text{ar}[Q] &= \mathbb{V}\text{ar}[Q_1] + \mathbb{V}\text{ar}[I\cdot Q_2] = \mathbb{V}\text{ar}[Q_1] + \mathbb{E}[Q_2]^2\mathbb{V}\text{ar}[I] + \mathbb{V}\text{ar}[Q_2]\mathbb{E}[I^2] \\
&= \frac{1}{u_1^2} + \frac{1}{u_2^2}\mathbb{E}[\phi({\theta^*}^\top X)(2-\phi({\theta^*}^\top X))].
\end{aligned}$$

When the number of arrivals is large, the empirical counterparts provide an good estimate.

EC.4.2. Verification of Assumptions

To verify Assumption 3, we have that

$$\begin{aligned}
&\left(\phi(x^\top\theta) - \phi(x^\top\theta^*)\right)x^\top(\theta-\theta^*) \\
=&\left(\frac{\exp(x^\top\theta)}{1+\exp(x^\top\theta)} - \frac{\exp(x^\top\theta^*)}{1+\exp(x^\top\theta^*)}\right)(x^\top(\theta-\theta^*)) \\
=&\phi'(x^\top\bar{\theta})(x^\top\theta - x^\top\theta^*)\cdot(x^\top\theta - x^\top\theta^*) \\
\ge&\min_{x\in\mathcal{X},\theta\in\Theta}\ \phi'(x^\top\theta)\cdot\left(x^\top(\theta-\theta^*)\right)^2.
\end{aligned}$$

Recall that $m(\theta^*) = \frac{1}{u_1} + \frac{1}{u_2}\frac{1}{M_n}\sum_{j=1}^{M_n}\phi(\theta^{*\top}X_{nj})$ and $v(\theta^*) = \frac{1}{u_1^2} + \frac{1}{u_2^2}\frac{1}{M_n}\sum_{j=1}^{M_n}\phi(\theta^{*\top}X_{nj})(2-\phi(\theta^{*\top}X_{nj}))$. To verify Assumption 4, we have that

$$\begin{aligned}
&|\mathbb{E}[r(p;\theta)|X_n] - \mathbb{E}[r(p;\theta^*)|X_n]| = cM_n|W(p;\theta) - W(p;\theta^*)| \\
=&\frac{c}{2}M_n\left|\frac{\Lambda(p)v(\theta)}{1-\Lambda(p)m(\theta)} - \frac{\Lambda(p)v(\theta^*)}{1-\Lambda(p)m(\theta^*)}\right| \\
=&\frac{cM_n\Lambda(p)}{2}\left|\frac{1}{1-\Lambda(p)m(\theta^*)}(v(\theta)-v(\theta^*)) + \left(\frac{1}{1-\Lambda(p)m(\theta)} - \frac{1}{1-\Lambda(p)m(\theta^*)}\right)v(\theta)\right| \\
\le&\frac{cM_n\Lambda(p)}{2}\left(\frac{1}{\min_{\theta\in\Theta}\ 1-\Lambda(p)m(\theta)}|v(\theta)-v(\theta^*)|\right. \\
&\left. + \max_{\theta\in\Theta}\left\{\frac{v(\theta)\Lambda(p)}{(1-\Lambda(p)m(\theta))(1-\Lambda(p)m(\theta^*))}\right\}|m(\theta)-m(\theta^*)|\right).
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&|\mathbb{E}[r(p;\theta)|X_n] - \mathbb{E}[r(p;\theta^*)|X_n]| \\
\le&\frac{cM_n\Lambda(p)}{2u_2^2}\frac{1}{\min_{\theta\in\Theta}\ 1-\Lambda(p)m(\theta)}\left(\frac{2}{M_n}\left|\sum_{j=1}^{M_n}(\phi(X_{nj}^\top\theta)-\phi(X_{nj}^\top\theta^*))\right| + \frac{1}{M_n}\left|\sum_{j=1}^{M_n}(\phi(X_{nj}^\top\theta)^2-\phi(X_{nj}^\top\theta^*)^2)\right|\right) \\
&+\frac{cM_n\Lambda(p)^2}{2u_2^2}\max_{\theta\in\Theta}\frac{v(\theta)}{(1-\Lambda(p)m(\theta))(1-\Lambda(p)m(\theta^*))}\frac{1}{M_n}\left|\sum_{j=1}^{M_n}(\phi(X_{nj}^\top\theta)^2-\phi(X_{nj}^\top\theta^*)^2)\right| \\
=&\frac{c\Lambda(p)}{2u_2^2}\frac{1}{\min_{\theta\in\Theta}\ 1-\Lambda(p)m(\theta)}\left(4\sum_{j=1}^{M_n}|\phi(X_{nj}^\top\theta)-\phi(X_{nj}^\top\theta^*)|\right) \\
&+\frac{c\Lambda(p)^2}{2u_2^2}\max_{\theta\in\Theta}\frac{v(\theta)}{(1-\lambda(p)m(\theta))(1-\Lambda(p)m(\theta^*))}\left(2\sum_{j=1}^{M_n}|\phi(X_{nj}^\top\theta)-\phi(X_{nj}^\top\theta^*)|\right) \\
\le&\mu\sum_{j=1}^{M_n}|\phi(X_{nj}^\top\theta)-\phi(X_{nj}^\top\theta^*)|,
\end{aligned}$$

where $\mu = \max_{p\in[\underline{p},\overline{p}]}\frac{2c\Lambda(p)}{u_2^2\min_{\theta\in\Theta}\ 1-\Lambda(p)m(\theta)} + \max_{\theta\in\Theta,p\in[\underline{p},\overline{p}]}\frac{c\Lambda(p)^2}{u_2^2}\frac{v(\theta)}{(1-\Lambda(p)m(\theta))(1-\Lambda(p)m(\theta^*))}$.

EC.4.3. Regret

The regret analysis follows similarly from the result in Appendix A. Since there are multiple observations per time period, we have $\hat{V}_n = \sum_{i=1}^{n-1}\sum_{j=1}^{M_i}\phi'(X_{ij}^\top\hat{\theta}_n)X_{ij}X_{ij}^\top$ and $\hat{\xi}_n = \sum_{i=1}^{n-1}\sum_{j=1}^{M_i}\epsilon_{ij}X_{ij}$. Lemma 7 holds for $\|\hat{\theta}_n-\theta^*\|_{\hat{V}_N}^2$. Regarding Lemma 8, we have

$$V_{n+1,*} = V_{n,*} + \sum_{j=1}^{M_n}X_{nj}X_{nj}^\top = V_{n,*}^{1/2}\left(I + \sum_{j=1}^{M_n}V_{n,*}^{-1/2}\left(X_{nj}X_{nj}^\top\right)V_{n,*}^{-1/2}\right)V_{n,*}^{1/2},$$

which implies that

$$\det(V_{n+1,*}) = \det(V_{n,*})\det\left(I + V_{n,*}^{-1/2}\Big(\sum_{j=1}^{M_n}X_{nj}X_{nj}^\top\Big)V_{n,*}^{-1/2}\right) \ge \det(V_{n,*})\left(1+\sum_{j=1}^{M_n}\|X_{nj}\|_{V_{n,*}^{-1}}^2\right).$$

Then we can bound the sum by

$$\sum_{n=\tau}^{N}\sum_{j=1}^{M_n}1\wedge\left(1+\|X_{nj}\|_{V_{n,*}^{-1}}^2\right) \le \sum_{n=\tau}^{N}\sum_{j=1}^{M_n}1\wedge\left(1+\|X_{nj}\|_{V_{n,*}^{-1}}^2\right) \le 2d\log\left(\frac{\kappa_f^2\sum_{n=1}^{N}M_n}{d\det(V_{\tau,*})^{1/d}}\right) \le 2d\log\left(\frac{\kappa_f^2\sum_{n=1}^{N}M_n}{d\lambda_{\min}(V_{\tau,*})}\right), \tag{EC.7}$$

which gives an extended result on multiple observations.

Note that

$$\mathbb{E}[|r_n(p;\theta)-r_n(p;\theta^*)|X_n] \le \mu \sum_{j=1}^{M_n} |\phi(X_{nj}^\top\theta)-\phi(X_{nj}^\top\theta^*)|,$$

Then inequality (EC.5) is adapted to

$$\begin{aligned}\mathbb{E}[r(p^*;\theta^*)-r(p_n;\theta^*)|X_n] &\le \beta_r \wedge \left(\mathbb{E}[r(p_n;\tilde{\theta}_n)|X_n]-\mathbb{E}[r(p_n;\theta^*)|X_n]\right)\\ &\le \beta_r \wedge \left(\mu \sum_{j=1}^{M_n} |\phi(X_{nj}^\top\tilde{\theta}_n)-\phi(X_{nj}^\top\theta^*)|\right).\end{aligned}$$

Then inequality (EC.6) holds that

$$\begin{aligned}&\sum_{n=\tau}^{N}\left(\beta_r \wedge \left(\mu\sum_{j=1}^{M_n}\left|\phi(\tilde{\theta}_n^\top X_{nj})-\phi(X_{nj}^\top\theta^*)\right|\right)\right)\mathbf{1}(\mathcal{E}_n^V)\mathbf{1}(\tilde{\mathcal{E}}_n^\theta)\mathbf{1}(\mathcal{E}_n^\theta)\\ \le &\sum_{n=\tau}^{N}\beta_r \wedge \left(\frac{2}{1-2\alpha_f}\rho_N\mu\left(\sum_{j=1}^{M_n}\|X_{nj}\|_{V_{n,*}^{-1}}\right)\right)\\ \le &(\beta_r \vee \tfrac{2}{1-2\alpha_f}\mu)\rho_N\sum_{n=\tau}^{N}\sum_{j=1}^{M_n}\left(1\wedge\|X_{nj}\|_{V_{n,*}^{-1}}\right).\end{aligned}$$

Using (EC.7), and the rest of the proof follows similarly from Theorem 3. Therefore, we reach the conclusion that the regret is upper bounded by $O(d\sqrt{N})$ up to logarithmic terms.

We remark that if Λ is multiplied by a constant M, then from our analysis the regret scales by M because $\beta_r \le \max_{p\in\mathcal{A}}\Lambda(p)p$. The setting with multiple observations per round is similar to parallel contextual bandits [21], where a fixed number M of decisions and the associated rewards are observed in parallel per time period. It is shown that the regret bound scales in $\sqrt{M}$ for parallel linear bandits. Using the doubling-round routine (Theorem 1 and Lemma 1 in [21]), it can be shown similarly that our regret bound scales by $\sqrt{M}$ when Λ is multiplied by M.

Appendix EC.5: Proofs for Section 5.2

We verify the non-convexity of ℓ as a function of θ. We have that

$$\nabla_\theta \ell(x,a) = -2(y-x^\top\theta\theta^\top a)(x^\top\theta a + a^\top\theta x)$$

and

$$\nabla_\theta^2\ell(x,a) = 2((x^\top\theta)^2 aa^\top + (2\theta^\top a x^\top\theta - y)(xa^\top + ax^\top) + (\theta^\top a)^2 xx^\top),$$

thus $\nabla_\theta^2\ell(x,a)$ can be negative definite for some $y\in\mathbb{R}$ on $[\underline{\theta},\overline{\theta}]^d$.

Appendix EC.6: Proofs for Appendix A

Throughout this section, we recall and define several notations

$$\begin{aligned}\mathcal{E}_n^V &= \{\|\hat{\theta}_n-\theta^*\|_{\hat{V}_n}\le\rho_n\},\\ \Delta_n &= \hat{\theta}_n-\theta^*,\\ \xi_n &= \sum_{i=1}^{n-1}\epsilon_i A_i,\\ V_{n,*} &= \sum_{i=1}^{n-1}\phi'(A_i^\top\theta^*)A_iA_i^\top.\end{aligned} \tag{EC.8}$$

EC.6.1. Proof of Lemma 7

Before proving Lemma 7, we first prove the following Lemma.

LEMMA EC.5. *Assume Assumptions 1' and 2" are in force. Let* $\delta \in (0,1)$, $\lambda_0 = \lambda_{\min}(V_{d,*}) > 0$, *and* $\eta = \sqrt{3 + 2\log(1 + 2\bar{\kappa}_\phi^2/\lambda_0)}$. *For all* $n \geq d$, *with probability at least* $1-\delta$, *it holds that*

$$\|\xi_n\|^2_{V_{n,*}^{-1}} \leq \frac{16 d\eta^2\sigma^2}{\underline{\kappa}_\phi}\log(n)\log(d/\delta).$$

Proof of Lemma EC.5. This proof is similar to that of Lemma 4. Let $x \in \mathbb{R}^d$ whose value will be specified later. Define

$$Z_1 := x^\top \xi_n,$$
$$Z_2 := \sqrt{2}\frac{\sigma}{\sqrt{\underline{\kappa}_\phi}}\|x\|_{V_{n,*}}.$$

For any $u \geq 0$, we have

$$\begin{aligned} uZ_1 - \frac{u^2}{2}Z_2^2 &= ux^\top\xi_n - \frac{u^2 x^\top V_{n,*}x}{2}\cdot 2\frac{\sigma^2}{\underline{\kappa}_\phi} \\ &= \sum_{i=1}^{n-1} u\epsilon_i A_i^\top x - \frac{u^2}{2}\cdot 2\frac{\sigma^2}{\underline{\kappa}_\phi}\cdot(\sqrt{\phi'(A_i^\top\theta^*)}A_i^\top x)^2 \\ &\leq \sum_{i=1}^{n-1} u\epsilon_i A_i^\top x - u^2\sigma^2\cdot(A_i^\top x)^2. \end{aligned}$$

Similar to the proof of Lemma 4 with $\theta = \theta^*$, define $D_i = u\epsilon_i A_i^\top x - u^2\sigma^2(A_i^\top x)^2$. We can prove that

$$\mathbb{E}\Big[\exp(uZ_1 - \frac{u^2}{2}Z_2^2)\Big] \leq \mathbb{E}\Big[\prod_{i=1}^{n-1}\exp(D_i)\Big] \leq \cdots \leq \mathbb{E}[\exp(D_1)] \leq 1,$$

and conclude that

$$\mathbb{P}\big\{\|\xi_n\|^2_{V_{n,*}^{-1}} \geq \frac{16 d\eta^2\sigma^2}{\underline{\kappa}_\phi}\log(n)\log(d/\delta)\big\} \leq \delta,$$

where $\eta = \sqrt{3 + 2\log(1 + 2\bar{\kappa}_\phi\beta_{\mathcal{A}}^2/\lambda_0)}$. □

Proof for Lemma 7. Applying the mean value theorem, for each i, there exists $\bar{\theta}_i$ such that

$$\phi(A_i^\top\theta^*) = \phi(A_i^\top\hat{\theta}_n) + \phi'(A_i^\top\hat{\theta}_n)A_i^\top\Delta_n + \phi''(A_i^\top\bar{\theta}_i)(A_i^\top\Delta_n)^2.$$

Thus

$$\begin{aligned} \|\xi_n\|^2_{V_{n,*}^{-1}} &= \left(\sum_{i=1}^{n-1}(\phi(A_i^\top\hat{\theta}_n) - \phi(A_i^\top\theta^*))A_i\right)^\top\left(\sum_{i=1}^{n-1}\phi'(A_i^\top\theta^*)A_iA_i^\top\right)^{-1}\left(\sum_{i=1}^{n-1}(\phi(A_i^\top\hat{\theta}_n) - \phi(A_i^\top\theta^*))A_i\right) \\ &= \Delta_n^\top\left(\sum_{i=1}^{n-1}\phi'(A_i^\top\bar{\theta}_i)A_iA_i^\top\right)\left(\sum_{i=1}^{n-1}\phi'(A_i^\top\theta^*)A_iA_i^\top\right)^{-1}\left(\sum_{i=1}^{n-1}\phi'(A_i^\top\bar{\theta}_i)A_iA_i^\top\right)\Delta_n \\ &\geq \frac{\underline{\kappa}_\phi^2}{\bar{\kappa}_\phi^2}\Delta_n^\top V_{n,*}V_{n,*}^{-1}V_{n,*}\Delta_n = \frac{\underline{\kappa}_\phi^2}{\bar{\kappa}_\phi^2}\|\Delta_n\|^2_{V_{n,*}}, \end{aligned}$$

where the inequality follows from Assumption 2". It follows that

$$\|\Delta_n\|_2^2 \le \frac{\|\Delta_n\|_{V_{n,*}}^2}{\lambda_{\min}(V_{n,*})} \le \frac{\bar{\kappa}_\phi^2 \|\xi_n\|_{V_{n,*}^{-1}}^2}{\underline{\kappa}_\phi^2 \lambda_{\min}(V_{n,*})}.$$

Define $U_n = \sum_{i=1}^{n-1} A_i A_i^\top$. Therefore, when $\lambda_{\min}(U_n) \ge \bar{\kappa}_\phi^2 256 d \left(\frac{\eta^2\sigma^2\bar{\kappa}_\phi\hbar_\phi\beta_{\mathcal{A}}}{\underline{\kappa}_\phi^3}\right)^2 \log(n)\log(dN/\delta)$, we have

$$\begin{aligned}\lambda_{\min}(V_{n,*}) \ge& \frac{1}{\bar{\kappa}_\phi^2}\lambda_{\min}(U_n) \ge 256 d \left(\frac{\eta^2\sigma^2\bar{\kappa}_\phi\hbar_\phi\beta_{\mathcal{A}}}{\underline{\kappa}_\phi^3}\right)^2 \log(n)\log(d/\delta)\\ =& 16 d \frac{\eta^2\sigma^2}{\underline{\kappa}_\phi^2}\log(n)\log(dN/\delta)\cdot\left(\frac{4\beta_{\mathcal{A}}\hbar_\phi}{\underline{\kappa}_\phi}\right)^2 \frac{\bar{\kappa}_\phi^2}{\underline{\kappa}_\phi^2}.\end{aligned}$$

Let event $\mathcal{J}_n = \{\|\xi_n\|_{V_{n,*}^{-1}}^2 \ge \frac{16 d\eta^2\sigma^2}{\underline{\kappa}_\phi}\log(n)\log(d/\delta)\}$. Combined with Lemma EC.5, it also holds that

$$\frac{\beta_{\mathcal{A}}\hbar_\phi}{\underline{\kappa}_\phi}\|\Delta_n\|_2 \mathbf{1}(\mathcal{E}_n)\mathbf{1}(\mathcal{J}_n) \le \frac{\beta_{\mathcal{A}}\hbar_\phi}{\underline{\kappa}_\phi}\frac{\bar{\kappa}_\phi\|\xi_n\|_{V_{n,*}^{-1}}}{\underline{\kappa}_\phi\sqrt{16 d\frac{\eta^2\sigma^2}{\underline{\kappa}_\phi^2}\log(n)\log(dN/\delta)\cdot\left(\frac{4\beta_{\mathcal{A}}\hbar_\phi}{\underline{\kappa}_\phi}\right)^2\frac{\bar{\kappa}_\phi^2}{\underline{\kappa}_\phi^2}}} \le \frac{1}{4}. \tag{EC.9}$$

Similarly, it holds that

$$\frac{\beta_{\mathcal{A}}\hbar_\phi}{\underline{\kappa}_\phi}\|\Delta_n\|_2 \mathbf{1}(\mathcal{E}_n)\mathbf{1}(\mathcal{E}_n^V) \le \frac{\beta_{\mathcal{A}}\hbar_\phi}{\underline{\kappa}_\phi}\frac{\|\Delta_n\|_{V_{n,*}}^2}{\lambda_{\min}(V_{n,*})}\mathbf{1}(\mathcal{E}_n)\mathbf{1}(\mathcal{E}_n^V) \le \frac{1}{4}. \tag{EC.10}$$

Therefore, using the mean value theorem again with $\tilde{\theta}_i$ being the intermediate value and Assumption 2", we have that

$$\begin{aligned}\|\xi_n\|_{V_{n,*}^{-1}}^2 =& \Delta_n^\top\left(\sum_{i=1}^{n-1}\phi'(A_i^\top\theta^*)A_iA_i^\top + \phi''(A_i^\top\tilde{\theta}_i)A_i^\top\Delta_n A_iA_i^\top\right)\left(\sum_{i=1}^{n-1}\phi'(A_i^\top\theta^*)A_iA_i^\top\right)^{-1}\\ &\left(\sum_{i=1}^{n-1}\phi'(A_i^\top\theta^*)A_iA_i^\top + \phi''(A_i^\top\tilde{\theta}_i)A_i^\top\Delta_n A_iA_i^\top\right)\Delta_n\\ \ge& \|\Delta_n\|_{V_{n,*}}^2 - 2\frac{\beta_{\mathcal{A}}\hbar_\phi}{\underline{\kappa}_\phi}\|\Delta_n\|_2\|\Delta_n\|_{V_{n,*}}^2 - \frac{\beta_{\mathcal{A}}^2\hbar_\phi^2}{\underline{\kappa}_\phi^2}\|\Delta_n\|_2^2\|\Delta_n\|_{V_{n,*}}^2\\ =& \|\Delta_n\|_{V_{n,*}}^2\left(1 - 2\frac{\beta_{\mathcal{A}}\hbar_\phi}{\underline{\kappa}_\phi}\|\Delta_n\|_2 - \frac{\beta_{\mathcal{A}}^2\hbar_\phi^2}{\underline{\kappa}_\phi^2}\|\Delta_n\|_2^2\right).\end{aligned}$$

Combining the two inequalities above, we conclude that

$$\|\Delta_n\|_{V_{n,*}}^2\mathbf{1}(\mathcal{E}_n)\mathbf{1}(\mathcal{J}_n) \le \frac{\|\xi_n\|_{V_{n,*}^{-1}}^2}{1 - 2\frac{\beta_{\mathcal{A}}\hbar_\phi}{\underline{\kappa}_\phi}\|\Delta_n\|_2 - \frac{\beta_{\mathcal{A}}^2\hbar_\phi^2}{\underline{\kappa}_\phi^2}\|\Delta_n\|_2^2}\mathbf{1}(\mathcal{E}_n)\mathbf{1}(\mathcal{J}_n) \overset{\text{(EC.9)}}{\le} \frac{256}{7\underline{\kappa}_\phi} d\eta^2\sigma^2\log(n)\log(dN/\delta).$$

By the mean value theorem, for each i there exists $\check{\theta}_i$ in the line segment connecting $\hat{\theta}_n$ and θ^* such that

$$\phi'(A_i^\top\hat{\theta}_n) = \phi'(A_i^\top\theta^*) + \phi''(A_i^\top\check{\theta}_i)^\top(A_i^\top\hat{\theta}_n - A_i^\top\theta^*).$$

It follows that

$$\begin{aligned}
\|\Delta_n\|_{\hat{V}_n}^2 \mathbf{1}(\mathcal{E}_n)\mathbf{1}(\mathcal{J}_n) &= \Delta_n^\top \left(\sum_{i=1}^{n-1} \phi'(A_i^\top \hat{\theta}_n) A_i A_i^\top\right) \Delta_n \mathbf{1}(\mathcal{E}_n)\mathbf{1}(\mathcal{J}_n) \\
&= \Delta_n^\top \left(\sum_{i=1}^{n-1} \frac{\phi'(A_i^\top \hat{\theta}_n)}{\phi'(A_i^\top \theta^*)} \phi'(A_i^\top \theta^*) A_i A_i^\top\right) \Delta_n \mathbf{1}(\mathcal{E}_n)\mathbf{1}(\mathcal{J}_n) \\
&= \Delta_n^\top \left(\sum_{i=1}^{n-1} \frac{\phi'(A_i^\top \theta^*) + \phi''(A_i^\top \check{\theta}_i) A_i^\top \Delta_n}{\phi'(A_i^\top \theta^*)} \phi'(A_i^\top \theta^*) A_i A_i^\top\right) \Delta_n \mathbf{1}(\mathcal{E}_n)\mathbf{1}(\mathcal{J}_n) \\
&\leq \left(1 + \frac{\beta_{\mathcal{A}} \hbar_\phi}{\underline{\kappa}_\phi} \|\Delta_n\|_2\right) \|\Delta_n\|_{V_{n,*}}^2 \mathbf{1}(\mathcal{E}_n)\mathbf{1}(\mathcal{J}_n) \overset{\text{(EC.9)}}{\leq} \frac{5}{4} \|\Delta_n\|_{V_{n,*}}^2 \mathbf{1}(\mathcal{E}_n)\mathbf{1}(\mathcal{J}_n).
\end{aligned}$$

Thus combined with Lemma EC.5, we conclude that with probability at least $1 - \delta/N$, it holds that $\|\Delta_n\|_{\hat{V}_n} \mathbf{1}(\mathcal{E}_n) \leq \rho_n$. □

EC.6.2. Proof of Lemma 8

To prove Lemma 8, we first prove the following lemma, which is a counterpart result of Lemma 6 for (UQ_o) with $\omega_{n,i}$ not equal to 1.

Lemma EC.6.

$$\sum_{n=\tau}^{N} \left(1 \wedge \|\sqrt{\phi'(A_n^\top \theta^*)} A_n\|_{V_{n,*}^{-1}}^2\right) \leq 2d \log\left(\frac{N \bar{\kappa}_\phi^2 \beta_{\mathcal{A}}^2}{d \det(V_{\tau,*})^{1/d}}\right)$$

Proof of Lemma EC.6. From the definition of $V_{n,*}$ (EC.8), we have

$$V_{n+1,*} = V_{n,*} + \phi'(A_n^\top \theta^*) A_n A_n^\top = V_{n,*}^{1/2} (I + V_{n,*}^{-1/2} \phi'(A_n^\top \theta^*) A_n A_n^\top V_{n,*}^{-1/2}) V_{n,*}^{1/2}$$

which implies that

$$\begin{aligned}
\det(V_{n+1,*}) &= \det(V_{n,*}) \det(I + V_{n,*}^{-1/2} \phi'(A_n^\top \theta^*) A_n A_n^\top V_{n,*}^{-1/2}) \\
&= \det(V_{n,*})(1 + \|\sqrt{\phi'(A_n^\top \theta^*)} A_n\|_{V_{n,*}^{-1}}^2).
\end{aligned}$$

Note that

$$\det(V_{n,*}) = \prod_{i=1}^{d} \lambda_i(V_{n,*}) \leq \left(\frac{1}{d} \mathrm{tr}(V_{n,*})\right)^d \leq \left(\frac{\mathrm{tr}(V_{\tau,*}) + n\bar{\kappa}_\phi \beta_{\mathcal{A}}^2}{d}\right)^d,$$

where $\lambda_i(V_{n,*})$ is the i^{th} eigenvalue of matrix $V_{n,*}$. Therefore,

$$\sum_{n=\tau}^{N} \left(1 \wedge \|\sqrt{\phi'(A_n^\top \theta^*)} A_n\|_{V_{n,*}^{-1}}^2\right) \leq 2 \log\left(\frac{\det V_{n,*}}{\det V_{\tau,*}}\right) \leq 2d \log\left(\frac{n\bar{\kappa}_\phi \beta_{\mathcal{A}}^2}{d \det(V_{\tau,*})^{1/d}}\right).$$

□

Proof for Lemma 8. First we note that

$$\|\sqrt{\phi'(a^\top \hat{\theta}_n)} a\|_{\hat{V}_n^{-1}}^2 = \phi'(a^\top \hat{\theta}_n) a^\top \hat{V}_n^{-1} a = a^\top \left(\sum_{i=1}^{n-1} \frac{\phi'(A_i^\top \hat{\theta}_n)}{\phi'(a^\top \hat{\theta}_n)} A_i \Theta n^\top\right)^{-1} a,$$

and

$$\|\sqrt{\phi'(a^\top \theta^*)} a\|_{V_{n,*}^{-1}}^2 = \phi'^2(a^\top \theta^*) a^\top V_{n,*}^{-1} a = a^\top \left(\sum_{i=1}^{n-1} \frac{\phi'(A_i^\top \theta^*)}{\phi'(a^\top \theta^*)} A_i A_i^\top\right)^{-1} a.$$

Define

$$\hat{c}_{ni} = \frac{\phi'(A_i^\top \hat{\theta}_n)}{\phi'(a^\top \hat{\theta}_n)} \text{ and } c_{ni} = \frac{\phi'(A_i^\top \theta^*)}{\phi'(a^\top \theta^*)}.$$

Since there exist some $\bar{\theta}_{ni}$ and $\tilde{\theta}_n$ such that

$$\phi'(A_i^\top \hat{\theta}_n) = \phi'(A_i^\top \theta^*) + \phi''(A_i^\top \bar{\theta}_{ni}) A_i^\top \Delta_n$$

and

$$\phi'(a^\top \hat{\theta}_n) = \phi'(a^\top \theta^*) + \phi''(a^\top \tilde{\theta}_n) a^\top \Delta_n,$$

we have

$$\begin{aligned}\frac{\hat{c}_{ni}}{c_{ni}} &= \frac{\phi'(A_i^\top \hat{\theta}_n) + \phi''(A_i^\top \bar{\theta}_{ni}) A_i^\top \Delta_n}{\phi'(A_i^\top \theta^*)} \frac{\phi'(a^\top \theta^*)}{\phi'(a^\top \theta^*) + \phi''(a^\top \tilde{\theta}_n) a^\top \Delta_n} \\ &\geq \left(1 - \frac{\beta_{\mathcal{A}} \hbar_\phi}{\underline{\kappa}_\phi} \|\Delta_n\|_2\right) \Big/ \left(1 + \frac{\beta_{\mathcal{A}} \hbar_\phi}{\underline{\kappa}_\phi} \|\Delta_n\|_2\right) \geq 1 - 2\frac{\beta_{\mathcal{A}} \hbar_\phi}{\underline{\kappa}_\phi} \|\Delta_n\|_2.\end{aligned}$$

Similarly, we have

$$\frac{c_{ni}}{\hat{c}_{ni}} \leq 1 + 2\frac{\beta_{\mathcal{A}} \hbar_\phi}{\underline{\kappa}_\phi} \|\Delta_n\|_2.$$

It implies that

$$\begin{aligned}\|\sqrt{\phi'(a^\top \hat{\theta}_n)} a\|^2_{\hat{V}_n^{-1}} &= a^\top \left(\sum_{i=1}^{n-1} \frac{\phi'(A_i^\top \hat{\theta}_n)}{\phi'(a^\top \hat{\theta}_n)} A_i A_i^\top\right)^{-1} a \\ &\leq \left(1 + 2\frac{\hbar_\phi \beta_{\mathcal{A}}}{\underline{\kappa}_\phi} \|\Delta_n\|_2\right) a^\top \left(\sum_{i=1}^{n-1} \frac{\phi'(A_i^\top \theta^*)}{\phi'(a^\top \theta^*)} A_i A_i^\top\right)^{-1} a \\ &\leq \left(1 + 2\frac{\hbar_\phi \beta_{\mathcal{A}}}{\underline{\kappa}_\phi} \|\Delta_n\|_2\right) \|\sqrt{\phi'(a^\top \theta^*)} a\|^2_{V_n^{-1}}.\end{aligned}$$

Therefore when the events $\mathcal{E}_n^V$ and $\mathcal{E}_n$ hold, it holds that

$$\sum_{n=\tau}^{N} \left(1 \wedge \|\sqrt{\phi'(A_n^\top \hat{\theta}_n)} A_n\|^2_{\hat{V}_n^{-1}}\right) \overset{\text{(EC.10)}}{\leq} \frac{3}{2} \sum_{n=\tau}^{N} \left(1 \wedge \|\sqrt{\phi'(A_n^\top \theta^*)} A_n\|^2_{V_{n,*}^{-1}}\right).$$

Combined with Lemma EC.6, we conclude that

$$\sum_{n=\tau}^{N} \left(1 \wedge \|\sqrt{\phi'(A_n^\top \hat{\theta}_n)} A_n\|^2_{\hat{V}_n^{-1}}\right) \mathbf{1}(\mathcal{E}_n^V \cap \mathcal{E}_n) \leq 3d \log\left(\frac{N \bar{\kappa}_\phi \beta_{\mathcal{A}}^2}{d \det(V_{\tau,*})^{1/d}}\right).$$

□

EC.6.3. Proof of Proposition 1

Lemma EC.7. *For any $a \in \mathcal{A}_n$ and θ satisfying $\|\theta - \hat{\theta}_n\|_{\hat{V}_n^{-1}} \leq \rho_n$, it holds that*

$$(\phi(a^\top \theta^*) - \phi(a^\top \theta)) \mathbf{1}(\mathcal{E}_n \cap \mathcal{E}_n^V) \leq \frac{5}{2} \rho_n \|\phi'(a^\top \hat{\theta}_n) a\|_{\hat{V}_n^{-1}}.$$

Proof of Lemma EC.7. For any $a \in \mathcal{A}_n$, by the mean value theorem, there exists $\bar{\theta}$ which is a convex combination of θ^* and θ such that

$$\phi(a^\top \theta) = \phi(a^\top \theta^*) + \phi'(a^\top \bar{\theta}) a^\top (\theta - \theta^*).$$

Therefore we have

$$\begin{aligned}\phi(a^\top \theta) &\le \phi(a^\top \theta^*) + \|\theta^* - \theta\|_{\hat{V}_n} \|\phi'(a^\top \bar{\theta}) a\|_{\hat{V}_n^{-1}} \\ &= \phi(a^\top \theta^*) + \|\theta^* - \theta\|_{\hat{V}_n} \frac{\phi'(a^\top \bar{\theta})}{\phi'(a^\top \hat{\theta}_n)} \|\phi'(a^\top \hat{\theta}_n) a\|_{\hat{V}_n^{-1}}.\end{aligned}$$

Since there exists some $\tilde{\theta}_n$ which is a convex combination of $\bar{\theta}$ and $\hat{\theta}_n$ such that

$$\phi'(a^\top \bar{\theta}) = \phi'(a^\top \hat{\theta}_n) + \phi''(a^\top \tilde{\theta}_n)(\bar{\theta} - \hat{\theta}_n)^\top a \le \phi'(a^\top \theta) + \beta_{\mathcal{A}} \hbar_\phi \|\bar{\theta} - \hat{\theta}_n\|_2,$$

we have

$$\phi(a^\top \theta) \le \phi(a^\top \theta^*) + \|\theta^* - \theta\|_{\hat{V}_n} \left(1 + \frac{\beta_{\mathcal{A}} \hbar_\phi}{\underline{\kappa}_\phi} \|\bar{\theta} - \hat{\theta}_n\|_2\right) \|\phi'(a^\top \hat{\theta}_n) a\|_{\hat{V}_n^{-1}}.$$

Note that the two-norm distance can be bounded by

$$\|\bar{\theta} - \theta^*\|_2^2 \le \frac{\|\bar{\theta} - \theta^*\|_{\hat{V}_n}^2}{\lambda_{\min}(\hat{V}_n)} \le \frac{\rho_n^2}{\underline{\kappa}_\phi^2 \lambda_{\min}(U_n)} \le \left(\frac{1}{4} \frac{\underline{\kappa}_\phi}{\beta_{\mathcal{A}} \hbar_\phi}\right)^2.$$

Therefore, it holds that

$$\phi(a^\top \theta) \le \phi(a^\top \theta^*) + \frac{5}{2} \rho_n \|\phi'(a^\top \hat{\theta}_n) a\|_{\hat{V}_n^{-1}},$$

where we reach the conclusion. □

LEMMA EC.8. *$\mathcal{E}_n$ holds with probability at least $1 - \delta/N$ for any $n \ge \tau$.*

Proof. Recall that

$$\bar{\lambda}_n = \max\left\{\left(\frac{16\eta^2 \sigma^2 \bar{\kappa}_\phi^3 \hbar_\phi \beta_{\mathcal{A}}}{\underline{\kappa}_\phi^3}\right)^2 d \log(n) \log(dN/\delta), \left(\frac{4\beta_{\mathcal{A}} \hbar_\phi}{\underline{\kappa}_\phi^2}\right)^2 \rho_n^2\right\}.$$

Since $\tau \ge \frac{2\bar{\lambda}_n}{\underline{\lambda}}$, applying Tropp [89, Theorem 3.1], we can bound $\lambda_{\min}(U_n)$ by

$$\begin{aligned}\mathbb{P}\left\{\lambda_{\min}(U_n) \le \bar{\lambda}_n\right\} &\le \mathbb{P}\left\{\lambda_{\min}(U_n) \le \frac{1}{2}\underline{\lambda}\tau\right\} \\ &\le \mathbb{P}\left\{\lambda_{\min}\left(\sum_{l=1}^{\tau} A_i A_i'\right) \le \frac{1}{2}\underline{\lambda}\tau \text{ and } \lambda_{\min}\left(\sum_{i=1}^{\tau} \mathbb{E}[A_i A_i' | \mathcal{F}_{i-1}]\right) \ge \underline{\lambda}\tau\right\} \\ &\le d\left(\frac{e}{2}\right)^{-\underline{\lambda}\tau/(2\bar{\lambda})} \le \frac{\delta}{N}.\end{aligned}$$

□

Proof of Proposition 1. By Lemma 4, $\mathcal{E}_n^V$ holds with probability at least $1 - \delta/N$. Let $\tilde{\theta}_n$ be the most optimistic θ obtained from (DOO). Then whenever event $\mathcal{E}_n^V$ holds, we have

$$\phi(A_n^\top \tilde{\theta}_n) = U_n(A_n) \ge U_n(A_n^*) \ge \phi(A_n^{*\top} \theta^*) \ge \phi(A_n^\top \theta^*),$$

where the second inequality holds because θ^* satisfies the constraint that $\|\hat{\theta}_n - \theta^*\|_{\hat{V}_n} \le \rho_n$ when $\mathcal{E}_n^V$ holds. As a result, whenever $\mathcal{G}_n$ holds, the one-step regret satisfies

$$\left(\phi(A_n^{*\top}\theta^*) - \phi(A_n^\top\theta^*)\right) \le \left(\beta_r \wedge \left(\phi(A_n^\top\tilde{\theta}_n) - \phi(A_n^\top\theta^*)\right)\right)$$

where the last inequality holds because the total reward is less than or equal to β_r. From Lemma EC.7 we have

$$(\phi(A_n^\top\tilde{\theta}_n) - \phi(A_n^\top\theta^*))\mathbf{1}(\mathcal{E}_n^V \cap \mathcal{E}_n) \le \frac{5}{2}\rho_n\sqrt{\bar{\kappa}_\phi}\|\sqrt{\phi'(a^\top\hat{\theta}_n)}a\|_{\hat{V}_n^{-1}}.$$

Therefore, we can bound the regret by

$$\begin{aligned}
\mathsf{Regret}_N \mathbf{1}(\cap_{n=\tau}^N \mathcal{E}_n)\mathbf{1}\left(\cap_{n=\tau}^N \mathcal{E}_n^V\right) &= \beta_r\tau + \sum_{n=\tau+1}^N \left(\phi(A_n^{*\top}\theta^*) - \phi(A_n^\top\theta^*)\right)\mathbf{1}(\mathcal{E}_n)\mathbf{1}\left(\mathcal{E}_n^V\right) \\
&\le \beta_r\tau + \sum_{n=\tau+1}^N \beta_r \wedge \left(\phi(A_n^\top\tilde{\theta}_n) - \phi(A_n^\top\theta^*)\right)\mathbf{1}(\mathcal{E}_n)\mathbf{1}\left(\mathcal{E}_n^V\right) \\
&\le \beta_r\tau + \frac{5}{2}\rho_N\beta_r\sqrt{\bar{\kappa}_\phi}\sum_{n=\tau}^N \left(1 \wedge \|\sqrt{\phi'(a^\top\hat{\theta}_n)}a\|_{\hat{V}_n^{-1}}\right) \\
&\le \beta_r\tau + \frac{5}{2}\rho_N\beta_r\sqrt{\bar{\kappa}_\phi}\sqrt{N\sum_{n=\tau}^N \left(1 \wedge \|\sqrt{\phi'(a^\top\hat{\theta}_n)}a\|_{\hat{V}_n^{-1}}^2\right)} \\
&\le \beta_r\tau + \frac{5}{2}\rho_N\beta_r\sqrt{\bar{\kappa}_\phi}\sqrt{3Nd\log\left(\frac{N\bar{\kappa}_\phi\beta_{\mathcal{A}}^2}{d\det(V_{\tau,*})^{1/d}}\right)}.
\end{aligned}$$

It implies that

$$\begin{aligned}
&\mathbb{P}\left\{\mathsf{Regret}_N \ge \beta_r\tau + \frac{5\sqrt{3}}{2}\rho_N\beta_r\sqrt{Nd\log\left(\frac{N\bar{\kappa}_\phi^2\beta_{\mathcal{A}}^2}{d\underline{\kappa}_\phi^2\lambda_{\min}(V_{\tau,*})}\right)}\right\} \\
\le&\mathbb{P}\left\{\mathsf{Regret}_N \mathbf{1}(\cap_{n=\tau}^N \mathcal{E}_n)\mathbf{1}\left(\cap_{n=\tau}^N \mathcal{E}_n^V\right) \ge \beta_r\tau + \frac{5\sqrt{3}}{2}\rho_N\beta_r\sqrt{Nd\log\left(\frac{N\bar{\kappa}_\phi^2\beta_{\mathcal{A}}^2}{d\underline{\kappa}_\phi^2\lambda_{\min}(V_{\tau,*})}\right)}\right\} \\
&+\sum_{n=1}^N \mathbb{P}\{(\mathcal{E}_n)^c\} + \sum_{n=1}^N \mathbb{P}\{(\mathcal{E}_n^V)^c\} \\
\le&2\delta.
\end{aligned}$$

□