Optimal Online Generalized Linear Regression with Stochastic Noise and Its Application to Heteroscedastic Bandits

Heyang Zhao [1] **Dongruo Zhou** [1] **Jiafan He** [1] **Quanquan Gu** [1]

Abstract

We study the problem of online generalized linear regression in the stochastic setting, where the label is generated from a generalized linear model with possibly unbounded additive noise. We provide a sharp analysis of the classical *follow-the-regularized-leader* (FTRL) algorithm to cope with the label noise. More specifically, for σ-sub-Gaussian label noise, our analysis provides an regret upper bound of $O(\sigma^2 d \log T) + o(\log T)$, where d is the dimension of the input vector, T is the total number of rounds. We also prove a $\Omega(\sigma^2 d \log(T/d))$ lower bound for stochastic online linear regression, which indicates that our upper bound is nearly optimal. In addition, we extend our analysis to a more refined Bernstein noise condition. As an application, we study generalized linear bandits with heteroscedastic noise and propose an algorithm based on FTRL to achieve the first variance-aware regret bound.

1. Introduction

Online learning (Cesa-Bianchi & Lugosi, 2006) plays a crucial role in modern data analytics and machine learning, where a learner progressively interacts with an environment, interactively updates its prediction utilizing sequential data. As a fundamental problem in online learning, online linear regression has been well studied in the adversarial setting (Littlestone et al., 1991; Azoury & Warmuth, 2001; Bartlett et al., 2015).

In the classic adversarial setting of online linear regression with square loss, the adversary initially generates a sequence of feature vectors $\{\mathbf{x}_t\}_{t\geq 1}$ in $\mathbb{R}^d$ with a sequence of labels (i.e., responses) $\{y_t\}_{t\geq 1}$ in $\mathbb{R}$. At each round $t \geq 1$, $\mathbf{x}_t$ is revealed to the learner and the learner then makes a prediction $\widehat{y}_t \in \mathbb{R}$ on $\mathbf{x}_t$. Afterward, the adversary reveals y_t, penalizes the learner by the square loss $(y_t - \widehat{y}_t)^2$, and enters the next round. The goal of the learner is to minimize the total loss of the first T rounds, which is measured by the adversarial regret defined as follows (Bartlett et al., 2015):

$$\mathcal{R}^{\text{adv}}(T) := \sum_{t=1}^{T} (\widehat{y}_t - y_t)^2 - \inf_{\boldsymbol{\mu}\in\mathbb{R}^d} \sum_{t=1}^{T} (\langle \mathbf{x}_t, \boldsymbol{\mu}\rangle - y_t)^2,$$

The adversarial regret indicates how far the current predictor is away from the best linear predictor in hindsight. Since the labels $\{y_t\}_{t\geq 1}$ are arbitrarily chosen by the adversary in the adversarial setting, existing results on regret upper bound usually require $\{y_t\}$ to be uniformly bounded, i.e., $y_t \in [-Y, Y]$ for all $t \geq 1$ and sometimes $\{\mathbf{x}_t\}_{t\geq 1}$ is assumed to be known to the learner at the beginning of the learning process (e.g., Bartlett et al., 2015). For adversarial setting, it has been shown that the minimax-optimal regret is of $O(dY^2 \log T)$ (Azoury & Warmuth, 2001), which gives a complete understanding about the statistical complexity.

It is also interesting to consider a stochastic variant of the classic online linear regression problem where y_t is generated from an underlying linear model with possibly unbounded noise ϵ_t. Under this setting, the stochastic regret, which will be formally introduced in Section 3, is defined more intuitively as the 'gap' between the predicted label and the underlying linear function $f_{\boldsymbol{\mu}^*}(\mathbf{x}_t) = \langle \mathbf{x}_t, \boldsymbol{\mu}^*\rangle$. It is worth noting that this stochastic setting is first studied by Ouhamma et al. (2021), for which they studied σ-sub-Gaussian noise and attained an $\widetilde{O}(\sigma^2 d^2)$ high-probability regret bound. However, whether such a bound is tight or improvable remains unknown. Thus, a natural question arises: *what is the optimal regret bound for stochastic online linear regression?*

Beyond the optimality of the existing regret bound, another concern is *whether the analysis for sub-Gaussian noise can be extended to other types of zero-mean noise.* Previous analyses for *online-ridge-regression* and *forward* algorithm provided by Ouhamma et al. (2021) highly rely on the *self-normalized concentration inequality for vector-valued martingale* (Abbasi-Yadkori et al., 2011, Theorem 1), which is for sub-Gaussian random variables.

[1]Department of Computer Science, University of California, Los Angeles, CA 90095, USA. Correspondence to: Quanquan Gu <qgu@cs.ucla.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

In this paper, we simultaneously address the aforementioned questions for stochastic online generalized linear regression, which admits online linear regression as a special case. We provide a sharp analysis for FTRL and a nearly matching lower bound in the stochastic setting.

1.1. Our Contributions

In this paper, we make a first attempt on achieving a nearly minimax-optimal regret for *stochastic* online linear regression by a fine-grained analysis of *follow-the-regularized-leader* (FTRL). To show the universality of our analysis, we consider a slightly larger function class, generalized linear class, and our main result on online linear regression is given in the form of a corollary.

Our contributions are summarized as follows:

- We propose a novel analysis on FTRL for online generalized linear regression with stochastic noise, which provides an $\widetilde{O}(\sigma^2 d) + o(\log T)$ regret bound under σ-sub-Gaussian noise, where d is the dimension of feature vectors, T is the number of rounds. Moreover, for general noise with a variance of σ^2 (not necessarily to be sub-Gaussian), we prove a more fine-grained upper bound of order $\widetilde{O}(A^2B^2 + d\sigma^2) + o(\log T)$, where A is the maximum Euclidean norm of feature vectors, B is the maximum Euclidean norm of $\boldsymbol{\mu}^*$.

- We provide a matching regret lower bound for online linear regression, indicating that our analysis of FTRL is sharp and attains the nearly optimal regret bound in the stochastic setting. To the best of our knowledge, this is the first regret lower bound for online (generalized) linear regression in the stochastic setting.

- As an application of our tighter result of FTRL for stochastic online regression, we consider generalized linear bandits with heteroscedastic noise (e.g., Zhou et al., 2021; Zhang et al., 2021; Dai et al., 2022). We propose a novel algorithm MOR-UCB based on FTRL, which achieves an $\widetilde{O}(d\sqrt{\sum_{t\in[T]} \mathrm{Var}\ \epsilon_t} + \sqrt{dT})$ regret. This is the first variance-aware regret for generalized linear bandits.

Notation. We denote by $[n]$ the set $\{1, \ldots, n\}$. For a vector $\mathbf{x} \in \mathbb{R}^d$ and matrix $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$, a positive semi-definite matrix, we denote by $\|\mathbf{x}\|_2$ the vector's Euclidean norm and define $\|\mathbf{x}\|_{\mathbf{\Sigma}} = \sqrt{\mathbf{x}^\top \mathbf{\Sigma}\mathbf{x}}$. For two positive sequences $\{a_n\}$ and $\{b_n\}$ with $n = 1, 2, \ldots$, we write $a_n = O(b_n)$ if there exists an absolute constant $C > 0$ such that $a_n \leq Cb_n$ holds for all $n \geq 1$ and write $a_n = \Omega(b_n)$ if there exists an absolute constant $C > 0$ such that $a_n \geq Cb_n$ holds for all $n \geq 1$. $\widetilde{O}(\cdot)$ is introduced to further hide the polylogarithmic factors. For a random event $\mathcal{E}$, we denote its indicator by $\mathbb{1}(\mathcal{E})$.

2. Related Work

Online linear regression in adversarial setting. Online linear regression has long been studied in the setting where the response variables (or labels) are bounded and chosen by an adversary (Foster, 1991; Littlestone et al., 1991; Cesa-Bianchi et al., 1996; Kivinen & Warmuth, 1997; Vovk, 1997; Bartlett et al., 2015; Malek & Bartlett, 2018). This problem is initiated by Foster (1991), where binary labels and ℓ_1 constrained parameters are considered. Cesa-Bianchi et al. (1996) proposed a gradient-descent based algorithm, which gives a regret bound of order $O(\sqrt{T})$ when the hidden vector is ℓ_2 constrained. Vovk (1997) proposed Aggregating Algorithm, achieving $O(Y^2 d \log T)$ regret where Y is the scale of labels and d is the dimension of the feature vectors. Bartlett et al. (2015) considered the case where the feature vectors are known to the learner at the start of the game and proposed an exact minimax regret for the problem. Afterwards, Malek & Bartlett (2018) generalized the results of Bartlett et al. (2015) to the cases where the labels and covariates can be chosen adaptively by the environment. Later on, Gaillard et al. (2019) proposed Forward algorithm, showing that Foward algorithm without regularization can achieve optimal asymptotic regret bound uniform over bounded observations.

Stochastic online linear regression. Recently, Ouhamma et al. (2021) considered the stochastic setting where the response variables are unbounded and revealed by the environment with additional random noise on the true labels. Ouhamma et al. (2021) discussed the limitations of online learning algorithms in the adversarial setting and further advocated for the need of complementary analyses for existing algorithms under the stochastic unbounded setting. In their paper, new analyses for online ridge regression and Forward algorithm are proposed, achieving asymptotic $O(\sigma^2 d^2 \log T)$ regret bound. [1]

Learning heteroscedastic bandits. Heteroscedastic noise has been studied in many settings such as active learning (Antos et al., 2010), regression (Aitken, 1936; Goldberg et al., 1997; Chaudhuri et al., 2017; Kersting et al., 2007), principal component analysis (Hong et al., 2016; 2018) and Bayesian optimization (Assael et al., 2014). However, only

[1] We noticed that Ouhamma et al. (2021) also mentioned a way to acquire a tighter regret bound by applying confident sets with an $O\left(\sqrt{d\log\log T + \log(1/\delta)}\right)$ radius (Tirinzoni et al., 2020), where the previous $O\left(\sqrt{d\log(T/\delta)}\right)$ one proposed by Abbasi-Yadkori et al. (2011) leads to an $\widetilde{O}(\sigma^2 d^2)$. In this way, the T dependence in their bound can be further improved. However, the d dependence is still quadratic.

a few works have considered heteroscedastic noise in bandit settings. Cowan et al. (2015) considered a variant of multi-armed bandits where the noise at each round is a Gaussian random variable with unknown variance. Kirschner & Krause (2018) is the first to formally introduce the concept of stochastic bandits with heteroscedastic noise. In their model, the variance of the noise at each round t is a function of the evaluation point x_t, $\rho_t = \rho(x_t)$, and they further assume that the noise is ρ_t-sub-Gaussian. ρ_t can either be observed at time t or either be estimated from the observations. Zhou et al. (2021) and Zhou & Gu (2022) considered linear bandits with heteroscedastic noise and generalized the heteroscedastic noise setting in Kirschner & Krause (2018) in the sense that they no longer assume the noise to be ρ_t-sub-Gaussian, but only requires the variance of noise to be upper bounded by ρ_t^2 and the variances are arbitrarily decided by the environment, which is not necessarily a function of the evaluation point. In the same setting as in Zhou et al. (2021), Zhang et al. (2021) further considered a strictly harder setting where the noise has unknown variance. They proposed an algorithm that can deal with unknown variance through a computationally inefficient clip technique. Our work basically considers the noise setting proposed by Zhou et al. (2021) and further generalizes their setting to bandits with general function classes. We will consider extending it to the harder setting as Zhang et al. (2021) as future work.

3. Preliminaries

We will introduce our problem setting and some basic concepts in this section.

3.1. Problem Setup

Stochastic online regression. Let T be the number of rounds. At each round $t \in [T]$, the learner observes a feature vector $\mathbf{x}_t \in \mathbb{R}^d$ which is arbitrarily generated by the environment with $\|\mathbf{x}_t\|_2 \leq A$. The environment also generates the true label based on underlying function $f_{\boldsymbol{\mu}^*}$ parameterized by $\boldsymbol{\mu}^*$ with $\|\boldsymbol{\mu}^*\|_2 \leq B$ and a stochastic noise ϵ_t, i.e. true label $y_t := f_{\boldsymbol{\mu}^*}(\mathbf{x}_t) + \epsilon_t$. After observing $\mathbf{x}_t$, the learner should output a prediction $\widehat{y}_t = f_{\widehat{\boldsymbol{\mu}}_t}(\mathbf{x}_t) \in \mathbb{R}$ where $\widehat{\boldsymbol{\mu}}_t$ stands for its estimation of $\boldsymbol{\mu}^*$. y_t is subsequently revealed to the learner at the end of the t-th round.

Generalized linear function class. In this work, we assume that $f_{\boldsymbol{\mu}^*}$ belongs to the generalized linear function class $\mathcal{G}$ with known activation function $\phi : \mathbb{R} \to \mathbb{R}$ such that

$$\mathcal{G} = \left\{ f_{\boldsymbol{\mu}} | \boldsymbol{\mu} \in \mathbb{R}^d, f_{\boldsymbol{\mu}}(\mathbf{x}) = \phi\left(\langle \boldsymbol{\mu}, \mathbf{x} \rangle\right) \text{ for } \forall \mathbf{x} \in \mathbb{R}^d \right\} \quad (1)$$

To make the regression problem tractable, we require the following assumption on activation function ϕ, which is a common assumption in literature (Filippi et al., 2010)

Assumption 3.1. The activation function $\phi(\cdot)$ is an increasing differentiable function on $[-B, B]$ and there exists $\kappa, K \in \mathbb{R}$ such that $0 < \kappa \leq \phi'(z) \leq K$ for all $z \in [-B, B]$.

Assunmption on noise. Two types of noise are considered in this work:

Condition 3.2 (sub-Gaussianity of noise)**.** Suppose the noise sequence $\{\epsilon_t\}_{t\in[T]}$ is a sequence of i.i.d. zero-mean sub-Gaussian random variables:

$$\forall t \geq 1, s \in \mathbb{R}, \quad \mathbb{E}[\exp(s\epsilon_t)] \leq \exp\left(\frac{\sigma^2 s^2}{2}\right).$$

Condition 3.3 (Bernstein's condition)**.** The noise sequence $\{\epsilon_t\}_{t\in[T]}$ is a sequence of independent zero-mean random variables such that

$$\forall t \geq 1, \quad \mathbb{P}\left(|\epsilon_t| \leq R\right) = 1, \mathbb{E}[\epsilon_t^2] \leq \sigma^2.$$

Remark 3.4. Noise with Bernstein condition naturally implies sub-Gaussianity with a variance parameter R. However, we want to emphasize that we are more interested in the separate dependence of R and σ defined in Condition 3.3 in the complexity bound we will derive. Simplely regarding Bernstein condition noise as a sub-Gaussian noise will omit the refined dependence we want to have.

Loss and regret. Following Jun et al. (2017), we define the loss function ℓ_t at each round $t \in [T]$ as follows .

$$\ell_t(\boldsymbol{\mu}) := -\mathbf{x}_t^\top \boldsymbol{\mu} y_t + \int_0^{\mathbf{x}_t^\top \boldsymbol{\mu}} \phi(z) \mathrm{d}z. \quad (2)$$

For linear case where ϕ is the identical mapping, the loss function ℓ_t is equivalent to the square loss between y_t and $\widehat{y}_t$, i.e. they only differ by a constant:

$$\ell_t(\boldsymbol{\mu}) = -\mathbf{x}_t^\top \boldsymbol{\mu} y_t + \int_0^{\mathbf{x}_t^\top \boldsymbol{\mu}} z \mathrm{d}z = \frac{1}{2}\left(\mathbf{x}_t^\top \boldsymbol{\mu} - y_t\right)^2 - \frac{1}{2} y_t^2.$$

Remark 3.5. Intuitively, our definition of loss functions $\{\ell_t\}_{t=1}^T$ is a sequence of negative log likelihood when the distribution of y belongs to the exponential family (Nelder & Wedderburn, 1972), i.e.

$$\mathbb{P}(y_t | \mathbf{x}_t^\top \boldsymbol{\mu} = z) = h(y, z) \exp\left(\frac{y \cdot z - a(z)}{b}\right)$$

where $a'(z) = \phi(z)$.

Aligned with the definition proposed by (Ouhamma et al., 2021), the stochastic regret is defined as the relative cumulative loss over $f_{\boldsymbol{\mu}^*}$:

$$\mathcal{R}^{\text{stoc}}(T) := \sum_{t\in[T]} \ell_t(\widehat{\boldsymbol{\mu}}_t) - \sum_{t\in[T]} \ell_t(\boldsymbol{\mu}^*). \quad (3)$$

Given number of rounds T, vector sequence $\{\mathbf{x}_t\}_{t\in T}$. For $t = 1, 2, \cdots, T$:

- At the beginning of round t, learner outputs a prediction $\widehat{y}_t$.
- The adversary reveals $y_t \in [-Y, Y] \subset \mathcal{R}$ to the learner.
- The learner observes y_t and incurs loss $(y_t - \widehat{y}_t)^2$.

Figure 1. Protocol of adversarial online linear regression

Remark 3.6. Consistent with our stochastic setting, the stochastic regret is defined in a more natural way, representing the gap of cumulative loss between the learner and the underlying function $f_{\boldsymbol{\mu}^*}$. In comparison, the adversarial regret is defined as

$$\mathcal{R}^{\text{adv}}(T) := \sum_{t\in[T]} \ell_t(\widehat{\boldsymbol{\mu}}^t) - \inf_{\boldsymbol{\mu}\in\mathbb{R}^d} \sum_{t\in[T]} \ell_t(\boldsymbol{\mu}).$$

These two definitions of regret are highly similar as what have discussed in previous work (Ouhamma et al., 2021, Theorem 3.1). The similarity has been well-studied for the adversarial bandit and stochastic bandit setting (Lattimore & Szepesvári, 2020). Actually, the two regrets are closed to each other in stochastic online linear regression. As shown later in Section 4, our bound for stochastic regret also leads to an upper bound for $\mathcal{R}^{\text{adv}}$ of the same order.

3.2. Comparison between Adversarial Setting and Stochastic Setting

In this subsection, we discuss existing results on the adversarial regret bounds for online linear regression. The minimax regret in the adversarial setting is first derived by Bartlett et al. (2015), in the case where all the feature vectors are known to the learner at the beginning of the first round.

The protocol of adversarial online regression can be summarized in the following Figure 1.

Exploiting the adversarial nature of the sequential data, Bartlett et al. (2015) directly solved the following minimax regret

$$\min_{\widehat{y}_1} \max_{y_1} \cdots \min_{\widehat{y}_T} \max_{y_T} \mathcal{R}^{\text{adv}}(T).$$

The optimal minimax regret is $O(Y^2 d \log T)$, while the optimal strategy for the adversary is to set the label y_t according to the following distribution:

$$y_t = \begin{cases} Y & w.p.\ \frac{1}{2} + \mathbf{x}_t \mathbf{P}_t \left(\sum_{\tau=1}^{t-1} y_\tau \mathbf{x}_\tau\right)/(2B) \\ -Y & w.p.\ \frac{1}{2} - \mathbf{x}_t \mathbf{P}_t \left(\sum_{\tau=1}^{t-1} y_\tau \mathbf{x}_\tau\right)/(2B) \end{cases} \tag{4}$$

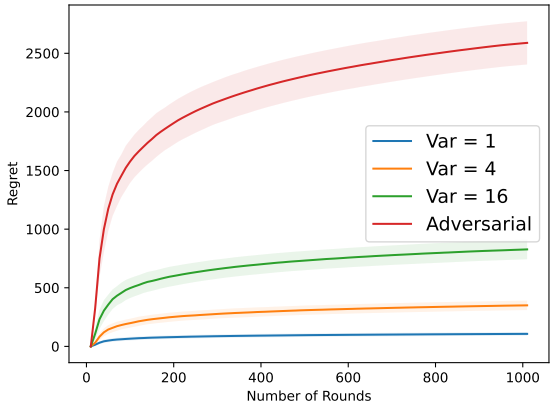


Figure 2. Regret of FTRL under different labels.

where $\mathbf{P}_t$ is defined by

$$\mathbf{P}_t^{-1} = \sum_{\tau=1}^{t} \mathbf{x}_\tau \mathbf{x}_\tau^\top + \sum_{\tau=t+1}^{T} \frac{\mathbf{x}_\tau^\top \mathbf{P}_\tau \mathbf{x}_\tau}{1 + \mathbf{x}_\tau^\top \mathbf{P}_\tau \mathbf{x}_\tau} \mathbf{x}_\tau \mathbf{x}_\tau^\top.$$

It is not hard to see that such a regret bound still holds in the stochastic setting if we set Y to be sufficiently large since $\inf_{\boldsymbol{\mu}\in\mathbb{R}^d} \sum_{t=1}^T (\langle \mathbf{x}_t, \boldsymbol{\mu}\rangle - y_t)^2 \le \sum_{t=1}^T (\langle \mathbf{x}_t, \boldsymbol{\mu}^*\rangle - y_t)^2$.

A natural question that arises here is whether it is significant to have a new regret bound for the stochastic setting.

To answer this question, we carry out experiments for *follow-the-regularized-leader* (FTRL) with both bounded stochastic label and bounded adversarial label defined in (4).

We conduct this experiment under different types of noise, including clipped Gaussian noise with standard variance 1, 2, 4 and adversarial noise according to Bartlett et al. (2015). The stochastic noise is clipped to ensure that all the labels are in the same bounded interval $[-Y, Y]$. We plot the regret of FTRL under different types of labels in Figure 2.

We can see that the regret of FTRL in the stochastic setting is remarkably smaller than that in the adversarial setting, thus we can not directly use the regret bound for the adversarial setting as a valid estimation of the regret in the stochastic setting. Besides, our experimental results indicate that the regret of FTRL highly depends on the noise variance, even though the ranges of labels are the same.

4. Optimal Stochastic Online Generalized Linear Regression

We provide analyses on *follow-the-regularized-leader* (FTRL) in this section. We apply a quadratic regularization in FTRL, as shown in Algorithm 1. At each round, FTRL aims to predict the unknown $\boldsymbol{\mu}^*$ defined in Section 3.1. To achieve this goal, FTRL computes the minimizer of the regularized cumulative loss over all previously observed

Algorithm 1 Follow The Regularized Leader (FTRL)

Input: $\mathcal{G}, \lambda$.
Initialize: $\widehat{\boldsymbol{\mu}}_1 \leftarrow \mathbf{0}$.
for $t \geq 1$ **do**
  Observe $\mathbf{x}_t$.
  Output $\widehat{y}_t = f_{\widehat{\boldsymbol{\mu}}_t}(\mathbf{x}_t)$.
  Observe y_t.
  Update $\widehat{\boldsymbol{\mu}}_{t+1} \leftarrow \operatorname{argmin}_{\boldsymbol{\mu} \in \mathbb{R}^d} \lambda \|\boldsymbol{\mu}\|_2^2 + \sum_{\tau=1}^{t} \ell_\tau(\boldsymbol{\mu})$,where ℓ_τ is defined in (2).
end for

context $\mathbf{x}_t$ and label y_t. This is slightly different from what we want to do for the adversarial setting, where the goal of FTRL is to predict the best predictor over T rounds.

4.1. Regret Upper Bound

We first propose the stochastic regret upper bound for FTRL under the stochastic online regression setting.

Theorem 4.1 (Regret of FTRL)**.** *Set $\lambda = 4A^2K^2/\kappa$ and assume that the noise ϵ_t satisfies Condition 3.2 at all rounds $t \in [T]$, then with probability at least $1 - 2\delta$, the regret of Algorithm 1 for the first T rounds is bounded as follows:*

$$\mathcal{R}^{stoc}(T) \leq 34\kappa^{-1} \cdot \sigma^2 d \log \frac{4dK^2 + T\kappa^2}{4dK^2} + 8\frac{K^2A^2B^2}{\kappa} + 2\left(\frac{2}{\kappa} + \frac{3\kappa}{K^2}\right)\sigma^2 \log(1/\delta).$$

Directly applying this theorem to online linear regression, we have the following result.

Corollary 4.2 (Regret of FTRL in stochastic online linear regression)**.** *Suppose that ϕ is the identical mapping. Set $\lambda = 4A^2$ and assume that Condition 3.2 holds. With probability at least $1 - 2\delta$, the regret of Algorithm 1 for the first T rounds is bounded as:*

$$\mathcal{R}^{stoc}(T) \leq O\left(\sigma^2 d \log T + A^2B^2\right).$$

Remark 4.3. Corollary 4.2 suggests a regret upper bound for stochastic online linear regression with square loss. Recent work by Ouhamma et al. (2021) studied this stochastic setting and managed to get rid of the $O(A^2B^2d\log T)$ term in classic result for online linear regression considering adversarial setting. Ouhamma et al. (2021) derived a high probability regret bound of $\widetilde{O}(\sigma^2d^2)$ after omitting the $o(\log(T)^2)$ terms (Theorem 3.3, Ouhamma et al. 2021). Unlike their result, our result does not suffer from the quadratic dependence on d. As for the $O(A^2B^2)$ term in our result, it is not hard to see that this part of loss is inevitable, since at the first round, the algorithm has no prior knowledge of $\boldsymbol{\theta}^*$. We defer the detailed analysis on the lower bound of the problem to the next section.

Remark 4.4. It is worth noting that there is also a line of works considering non-stationary online linear regression with noisy observations where $\boldsymbol{\theta}^*$ is time-varying and the goal is to minimize the regret with respect to the best linear predictor at each round (Besbes et al., 2015; Herbster & Warmuth, 2001; Zinkevich, 2003; Zhang et al., 2018; Baby & Wang, 2019; Raj et al., 2020). However, the regret upper bounds in this harder setting have polynomial dependence on T, which cannot be directly compared with our result.

4.2. Experimental Results

In this subsection, we provide experimental evidence supporting that the high-probability regret of FTRL grows linearly on the dimension of the feature vectors. We plot the stochastic regret with respect to the dimension of the feature vectors in Figure 3. More specifically, we compute the stochastic regret over different numbers of rounds, i.e., $\mathcal{R}^{\text{stoc}}(1000), \mathcal{R}^{\text{stoc}}(2000), \mathcal{R}^{\text{stoc}}(5000)$, under different Gaussian noise with $\sigma^2 = 1, 4, 16$. In each trial we sample $\boldsymbol{\mu}^*$ uniformly from $[-1/\sqrt{d}, 1/\sqrt{d}]^d$. At each round, we sample a feature vector uniformly from the unit sphere centered at the origin.

In Figure 3, we observe that under a fixed number of rounds, the value of stochastic regret has a linear dependence on d, which corroborates our theoretical analysis in Section 4.

4.3. Extension to Bernstein's Condition

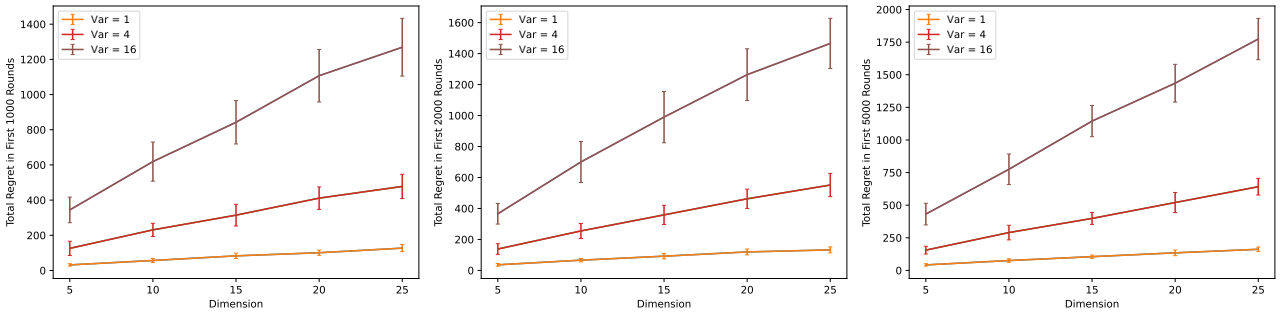
In some real-world scenarios, however, the sub-Gaussianity condition (Condition 3.2) may be too strong for general zero-mean noise.

In this subsection, we consider another assumption that the variance of ϵ_t is not larger than σ^2, which is formally stated in Condition 3.3. To remove the sub-Gaussianity condition, we introduce a new parameter R in Condition 3.3, serving as a large uniform upper bound on the noise $\{\epsilon_t\}_{t\in[T]}$.

Theorem 4.5 (Regret of FTRL)**.** *Set $\lambda = 4A^2K^2/\kappa$ and assume that the noise ϵ_t satisfy Condition 3.3 at all rounds $t \in [T]$, then with probability at least $1 - 2\delta$, the regret of Algorithm 1 for the first T rounds is bounded as follows:*

$$\mathcal{R}^{stoc}(T) \leq 6\kappa^{-1} \cdot \sigma^2 d \log \frac{4dK^2 + T\kappa^2}{4dK^2} + 8\frac{K^2A^2B^2}{\kappa} + 2\left(\frac{2}{\kappa} + \frac{5\kappa}{24K^2}\right)R^2 \log(1/\delta).$$

Remark 4.6. When T is sufficiently large, this bound becomes $O(\sigma^2 d \log T)$ in stochastic online linear regression, where σ is a uniform bound on $\mathbb{E}[\epsilon_t^2]$ for $t \in [T]$. Compared with Theorem 4.1, this theorem deals with a wider class of noise types. We defer the proof of this theorem to Subsection B.1.



(a) $T = 1000$ (b) $T = 2000$ (c) $T = 5000$

Figure 3. Dimension-dependence of regret of FTRL.

4.4. Proof Outline of Theorem 4.1

We provide the proof sketch for Theorem 4.1 here. We introduce the following concepts before presenting our results and analyses. First, we define the cumulative loss as follows:

$$\mathcal{L}_t(\boldsymbol{\mu}) = \sum_{\tau=1}^{t} \ell_\tau(\boldsymbol{\mu}) + \lambda\|\boldsymbol{\mu}\|_2^2. \tag{5}$$

For simplicity, we denote the Hessian matrix of $\mathcal{L}_t$ by $\mathbf{H}_t$ for each $t \geq 1$, i.e.,

$$\mathbf{H}_t(\boldsymbol{\mu}) := \mathbf{H}_{\mathcal{L}_t}(\boldsymbol{\mu}) = 2\lambda \cdot \mathbf{I} + \sum_{\tau=1}^{t} \phi'(\mathbf{x}_\tau^\top \boldsymbol{\mu}) \cdot \mathbf{x}_\tau \mathbf{x}_\tau^\top. \tag{6}$$

We also construct the following sequence $\{\underline{\mathbf{H}}_t\}_{t\in[H]}$:

$$\underline{\mathbf{H}}_t := 2\lambda \cdot \mathbf{I} + \kappa \sum_{\tau=1}^{t} \mathbf{x}_\tau \mathbf{x}_\tau^\top. \tag{7}$$

By the convexity of ϕ, it is easy to show that $\{\underline{\mathbf{H}}_t\}_{t\in[H]}$ is a lower bound of $\{\mathbf{H}_t(\cdot)\}_{t\in[H]}$ since for all $t \in [T]$, $\underline{\mathbf{H}}_t \preceq \mathbf{H}_t(\cdot)$.

4.4.1. Regret Decomposition

We first point out the key place which prevents Ouhamma et al. (2021) obtaining a tight regret bound. In the previous analysis of online learning algorithms in stochastic setting (Ouhamma et al., 2021), the regret is bounded through the summation of instantaneous regret

$$\sum_{t=1}^{T} \left(\ell_t(\widehat{\boldsymbol{\mu}}_t) - \ell_t(\boldsymbol{\mu}^*)\right) \leq \sum_{t=1}^{T} O\left(\|\widehat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}^*\|_{\mathbf{G}_t}^2\right) \cdot \|\mathbf{x}_t\|_{\mathbf{G}_t^{-1}}^2$$

where $\mathbf{G}_t = \lambda + \sum_{\tau=1}^{t-1} \mathbf{x}_\tau \mathbf{x}_\tau^\top$ is the sample covariance matrix. Applying a similar confidence ellipsoid as proposed in Abbasi-Yadkori et al. (2011), $\|\widehat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}^*\|_{\mathbf{G}_t}$ is uniformly bounded by $\widetilde{O}(\sigma^2 d)$. By Matrix Potential Lemma (Abbasi-Yadkori et al., 2011), $\sum_{t=1}^{T} \|\mathbf{x}_t\|_{\mathbf{G}_t^{-1}}^2$ is bounded by $O(d \log T)$. Thus, a quadratic dependence of d is inevitable in their regret upper bound.

To circumvent this issue, we prove the following lemma, which decomposes the cumulative regret into three terms.

Lemma 4.7 (Regret decomposition)**.** *For each $t \in [T]$, let $\mathcal{L}_t$ be the cumulative loss function defined in (5) and $\mathbf{H}_t$ be the corresponding Hessian matrix as shown in (6). There exists a sequence $\{\boldsymbol{\mu}'_t\}_{t\in[T]}$ in $\mathbb{R}^d$ such that the stochastic regret of Algorithm 1 can be decomposed as follows:*

$$\mathcal{R}^{stoc}(T) \leq \lambda B^2 + \frac{1}{2}\sum_{t=1}^{T} \epsilon_t^2 \|\mathbf{x}_t\|_{\mathbf{H}_t^{-1}(\boldsymbol{\mu}'_t)}^2 + \frac{1}{2}\sum_{t=1}^{T} \left(\phi(\mathbf{x}_t^\top \widehat{\boldsymbol{\mu}}_t) - \phi(\mathbf{x}_t^\top \boldsymbol{\mu}^*)\right)^2 \|\mathbf{x}_t\|_{\mathbf{H}_t^{-1}(\boldsymbol{\mu}'_t)}^2.$$

Intuitively speaking, $\sum_{t=1}^{T} \epsilon_t^2 \|\mathbf{x}_t\|_{\mathbf{H}_t^{-1}(\boldsymbol{\mu}'_t)}^2$ represents the part of regret caused by the random noise, while $\sum_{t=1}^{T} \left(\phi(\mathbf{x}_t^\top \widehat{\boldsymbol{\mu}}_t) - \phi(\mathbf{x}_t^\top \boldsymbol{\mu}^*)\right)^2 \|\mathbf{x}_t\|_{\mathbf{H}_t^{-1}(\boldsymbol{\mu}'_t)}^2$ represents the gap between the estimator $\widehat{\boldsymbol{\mu}}_t$ and the hidden vector $\boldsymbol{\mu}^*$. We bound these two terms separately as follows.

4.4.2. Bounding the Estimation Error

To derive a high-probability upper bound for the term $\sum_{t=1}^{T} \left(\phi(\mathbf{x}_t^\top \widehat{\boldsymbol{\mu}}_t) - \phi(\mathbf{x}_t^\top \boldsymbol{\mu}^*)\right)^2 \|\mathbf{x}_t\|_{\mathbf{H}_t^{-1}(\boldsymbol{\mu}'_t)}^2$ in Lemma 4.7, we start by considering the connection between $\left(\phi(\mathbf{x}_t^\top \widehat{\boldsymbol{\mu}}_t) - \phi(\mathbf{x}_t^\top \boldsymbol{\mu}^*)\right)$ and the cumulative regression regret.

Lemma 4.8 (Connection between squared estimation error and regret)**.** *Consider an arbitrary online learner interactively trained with stochastic data for T rounds as described in Section 3. If Condition 3.2 is true for the noise at all the rounds $t \in [T]$, then the following inequality holds with*

probability at least $1-\delta$*:*

$$\sum_{t=1}^{T}\left[\mathbf{x}_t^\top(\widehat{\boldsymbol{\mu}}_t-\boldsymbol{\mu}^*)\right]^2 \le \frac{4}{\kappa}\mathcal{R}^{stoc}(T)+\frac{16}{\kappa^2}\cdot\sigma^2\log(1/\delta).$$

4.4.3. Bounding the weighted sum of squared noise

According to Lemma 4.7, it remains to bound the term $\sum_{t=1}^{T}\epsilon_t^2\|\mathbf{x}_t\|^2_{\mathbf{H}_t^{-1}(\boldsymbol{\mu}_t')}$, which can be regarded as the weighted sum of squared sub-Gaussian random variables.

In our analysis, we first show that $\epsilon_t^2\|\mathbf{x}_t\|^2_{\mathbf{H}_t^{-1}(\boldsymbol{\mu}_t')}$ is a sub-exponential random variable and apply a tail bound on the total summation. The result is presented in the following lemma.

Lemma 4.9. *Suppose that the sequence of noise* $\{\epsilon_t\}_{t\in[T]}$ *satisfies Condition 3.2. For each* $t\in[T]$*, let* $\underline{\mathbf{H}}_t$ *be the matrix defined in (7). With probability at least* $1-\delta$*,*

$$\sum_{t=1}^{T}\epsilon_t^2\|\mathbf{x}_t\|^2_{\underline{\mathbf{H}}_t^{-1}} \le 34\kappa^{-1}\cdot\sigma^2 d\log\frac{d\lambda+T\kappa A^2}{d\lambda} + 24\sigma^2\cdot\frac{A^2}{\lambda}\log(1/\delta).$$

Putting all things together From Lemma 4.7, 4.8, 4.9, we conclude by a union bound that, with probability at least $1-2\delta$,

$$\mathcal{R}^{\text{stoc}}(T)\le O(\kappa^{-1}\cdot\sigma^2 d\log T)+o(\log T).$$

5. Lower Bound

In this section, we present a lower bound of the stochastic regret for stochastic online linear regression, which indicates the FTRL algorithm is already tight and optimal.

Theorem 5.1 (Lower bound for stochastic online linear regression)**.** *Consider the case where* $\kappa=K=1$ *in Assumption 3.1. Then the problem degrades to stochastic online linear regression where the stochastic regret can be defined as follows:*

$$\mathcal{R}^{stoc}(T):=\sum_{t=1}^{T}\ell_t(\widehat{\boldsymbol{\mu}}_t)-\sum_{t=1}^{T}\ell_t(\boldsymbol{\mu}^*)$$

Suppose that the noise sequence $\{\epsilon_t\}_{t\in[T]}$ *is a sequence of i.i.d. Gaussian random variables, i.e.,* $\epsilon_t\sim\mathcal{N}(0,\sigma^2)$ *for all* $t\in[T]$*. When* T *is sufficiently large, for any online regression algorithm, there exists* $\boldsymbol{\mu}^*\in\mathbb{R}^d$ *and a sequence of feature vectors* $\{\mathbf{x}_t\}_{t\in[T]}$ *such that* $\mathbb{E}[\mathcal{R}^{stoc}(T)]\ge\Omega(\sigma^2 d\log(T/d)+B\max_{t\in[T]}\|\mathbf{x}_t\|_2)$.

We notice that Mourtada (2022) provided a lower bound for expected excess risk in a random-design linear prediction problem, which can be seen as an offline version of the considered problem in our work. However, in online linear regression, we have to prove the existence of a $\boldsymbol{\theta}^*$ which makes their result hold for every round $t\in[T]$.

The proof of Theorem 5.1 involves the application of Pinsker's inequality, which is stated in Lemma D.8.

Proof of Theorem 5.1. For $\boldsymbol{\mu}\in\mathbb{R}^d$, $\|\boldsymbol{\mu}\|_2\le B$, we denote by $P_{\boldsymbol{\mu}}$ the measure on $\widehat{y}_1,\cdots,\widehat{y}_T,y_1,\cdots,y_T$ generated by the interaction between the algorithm and the environment.

We suppose that the sequence of feature vectors are fixed and consists of unit vectors. Let $\mathcal{T}_i:=\{t\in[T]|\mathbf{x}_t=\mathbf{e}_i\}$ for each $i\in[d]$, and $t_{i,j}$ be the j-th element in $\mathcal{T}_i$.

Assume that $\boldsymbol{\mu}^*$ is uniformly sampled from $[-B/\sqrt{d},B/\sqrt{d}]^d$ at the beginning of the first round.

We show that for all $i\in[d]$, $\mathbb{E}[(\widehat{y}_{t_{i,j}}-\langle\mathbf{x}_{t_{i,j}},\boldsymbol{\mu}^*\rangle)^2]\ge\Omega\left(\frac{\sigma^2}{j}\right)$ for all sufficiently large j. Consider any pair of $\boldsymbol{\mu}^1,\boldsymbol{\mu}^2\in[-B/\sqrt{d},B/\sqrt{d}]^d$ such that $\boldsymbol{\mu}_k^1=\boldsymbol{\mu}_k^2$ for all $k\in[d]\backslash\{i\}$ and $\left|\boldsymbol{\mu}_i^1-\boldsymbol{\mu}_i^2\right|\in\left[\frac{\sigma}{8\cdot\sqrt{j-1}},\frac{\sigma}{4\cdot\sqrt{j-1}}\right]$. By Lemma D.8, for any event A in the filtration generated by the labels before round $t_{i,j}$,

$$\left|P_{\boldsymbol{\mu}^1}(A)-P_{\boldsymbol{\mu}^2}(A)\right|\le\sqrt{\frac{1}{2}\mathrm{KL}\left(P_{\boldsymbol{\mu}^1}\|P_{\boldsymbol{\mu}^2}\right)} = \sqrt{\frac{1}{2}\sum_{t=1}^{t_{i,j}-1}\frac{\left(\mathbf{x}_t^\top\boldsymbol{\mu}^1-\mathbf{x}_t^\top\boldsymbol{\mu}^2\right)^2}{2\sigma^2}}\le\frac{1}{8}.$$

Thus, for any event A, $P_{\boldsymbol{\mu}^1}(A)+P_{\boldsymbol{\mu}^2}(\overline{A})\ge 7/8$. Let $A=\{\widehat{y}_{t_{i,j}}\ge(\boldsymbol{\mu}_i^1+\boldsymbol{\mu}_i^2)/2\}$. We have

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\mu}^1}\left[(\widehat{y}_{t_{i,j}}-\langle\mathbf{x}_{t_{i,j}},\boldsymbol{\mu}^1\rangle)^2\right]+\mathbb{E}_{\boldsymbol{\mu}^2}\left[(\widehat{y}_{t_{i,j}}-\langle\mathbf{x}_{t_{i,j}},\boldsymbol{\mu}^2\rangle)^2\right]\\&=\mathbb{E}_{\boldsymbol{\mu}^1}\left[(\boldsymbol{\mu}_1^i-\widehat{y}_{t_{i,j}})^2\right]+\mathbb{E}_{\boldsymbol{\mu}^2}\left[(\boldsymbol{\mu}_2^i-\widehat{y}_{t_{i,j}})^2\right]\\&\ge(P_{\boldsymbol{\mu}^1}(A)+P_{\boldsymbol{\mu}^2}(\overline{A}))\frac{\sigma^2}{256(j-1)}\ge\Omega(\sigma^2/j)\end{aligned}\tag{8}$$

For any 'segment' $\mathcal{S}=\{\boldsymbol{\mu}\in[-B/\sqrt{d},B/\sqrt{d}]^d|\boldsymbol{\mu}_i\in\left(a,a+\frac{\sigma}{4\cdot\sqrt{j-1}}\right),\boldsymbol{\mu}_k=c_k\ for\ k\ne i\}$, we denote $\boldsymbol{\mu}_{\mathcal{S}}(u):=(c_1,\cdots,a+u,c_{i+1},\cdots,c_d)^\top$.

We have

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\mu}\in\mathcal{S}}\left[(\widehat{y}_{t_{i,j}}-\langle\mathbf{x}_{t_{i,j}},\boldsymbol{\mu}\rangle)^2\right]\\&=\frac{4\sqrt{j-1}}{\sigma}\int_a^{a+\frac{\sigma}{4\cdot\sqrt{j-1}}}\mathbb{E}_{\boldsymbol{\mu}_{\mathcal{S}}(u)}\left[(\widehat{y}_{t_{i,j}}-(\boldsymbol{\mu}_{\mathcal{S}}(u))_i)^2\right]\mathrm{d}u\\&\ge\Omega(\sigma^2/j),\end{aligned}$$

where the last inequality holds due to (8).

From the arbitrariness of $\{c_k\}$ in $\mathcal{S}$ and the uniform distribution of $\boldsymbol{\mu}^*$, we have the following conclusion

for a 'slice' in the cube $[-B/\sqrt{d}, B/\sqrt{d}]^d$: For $\mathcal{B} = \left\{\boldsymbol{\mu} \in [-B/\sqrt{d}, B/\sqrt{d}]^d | \boldsymbol{\mu}_i \in \left(a, a + \frac{\sigma}{4\cdot\sqrt{j-1}}\right)\right\}$,

$$\mathbb{E}_{\boldsymbol{\mu}\in\mathcal{B}}\left[(\widehat{y}_{t_{i,j}} - \langle \mathbf{x}_{t_{i,j}}, \boldsymbol{\mu}\rangle)^2\right] \geq \Omega(\sigma^2/j).$$

It is then straightforward to conclude that $\mathbb{E}\left[(\widehat{y}_{t_{i,j}} - \langle \mathbf{x}_{t_{i,j}}, \boldsymbol{\mu}\rangle)^2\right] \geq \Omega(\sigma^2/j)$ when j is sufficiently large so that $\frac{\sigma}{4\cdot\sqrt{j-1}}$ is sufficiently small.

If we generate $\{\mathbf{x}_t\}$ such that $\mathcal{T}_i = \lfloor T/d \rfloor$ for all $i \in [d]$, then we have

$$\sum_{t=1}^{T} \mathbb{E}\left[(\widehat{y}_t - \langle \mathbf{x}_t, \boldsymbol{\mu}^*\rangle)^2\right] \geq \Omega(\sigma^2 \log(T/d)).$$

Trivially,

$$\sum_{t=1}^{T} \mathbb{E}\left[(\widehat{y}_t - \langle \mathbf{x}_t, \boldsymbol{\mu}^*\rangle)^2\right] \geq \sum_{i\in[d]} \mathbb{E}\left[\left(\widehat{y}_{t_{i,1}} - \boldsymbol{\mu}_i^*\right)^2\right] \geq \sum_{i\in[d]} \mathbb{E}\left[(\boldsymbol{\mu}_i^*)^2\right] \geq \Omega(B).$$

Applying Lemma 4.8, we can further conclude that $\mathbb{E}[\mathcal{R}^{\text{stoc}}(T)] \geq \Omega(\sigma^2 d \log(T/d) + B)$ when T is sufficiently large.

□

6. Application to Heteroscedastic Bandits

In the last section, it is shown that FTRL achieves nearly optimal regret faced with sequential data with zero-mean noise. Particularly, in subsection 4.3, we show that FTRL is capable of dealing with noise such that $\sigma^2 = \max_{t\in[T]} \text{Var}[\epsilon_t]$. However, it only utilizes the max variance information, which is not satisfactory if we want to see the trend of the change of the variance w.r.t. rounds. Thus, it is natural to ask whether we can design an algorithm for the generalized linear bandits, whose statistical complexity depends on the variance *adaptively*, says, depends on the *total variance* $\Sigma_t \text{Var}[\epsilon_t]$. For linear bandits setting, such a goal has been achieved in Zhou et al. (2021); Zhang et al. (2021).

6.1. Problem Setup

We consider a heteroscedastic variant of the classic stochastic bandit problem with generalized linear reward functions. At each round $t \in [T]$ ($T \in \mathbb{N}$), the agent observes a decision set $\mathcal{D}_t \subseteq \mathbb{R}^d$ which is chosen by the environment. The agent then selects an action $\mathbf{a}_t \in \mathcal{D}_t$ and observes reward r_t together with a corresponding variance upper bound σ_t^2. We assume that $r_t = f^*(\mathbf{a}_t) + \epsilon_t$ where $f^* = f_{\boldsymbol{\theta}^*} \in \mathcal{G}$ defined in (1) is the underlying real-valued reward function and ϵ_t is a random noise. We make the following assumption on ϵ_t.

Assumption 6.1 (Heteroscedastic noise)**.** The noise sequence $\{\epsilon_t\}_{t\in[T]}$ is a sequence of independent zero-mean random variables such that

$$\forall t \geq 1, \quad \mathbb{P}\left(|\epsilon_t| \leq R\right) = 1, \mathbb{E}\left[\epsilon_t^2\right] \leq \sigma_t^2.$$

The goal of the agent is to minimize the following cumulative regret:

$$\text{Regret}(T) := \textstyle\sum_{t=1}^{T} [f^*(\mathbf{a}_t^*) - f^*(\mathbf{a}_t)], \tag{9}$$

where the optimal action $\mathbf{a}_t^*$ at round $t \in [T]$ is defined as $\mathbf{a}_t^* := \text{argmax}_{\mathbf{a}\in\mathcal{D}_t} f^*(\mathbf{a})$.

6.2. The Proposed Algorithm

Existing approach. To tackle the heteroscedastic bandit problem, for the case where the $\mathcal{F}$ is the linear function class (i.e., $f(a) = \langle \boldsymbol{\theta}^*, a\rangle$ for some $\boldsymbol{\theta}^* \in \mathbb{R}^d$), a *weighted linear regression* framework (Kirschner & Krause, 2018; Zhou et al., 2021) has been proposed. Generally speaking, at each round $t \in [T]$, weighted linear regression constructs a confidence set $\mathcal{C}_t$ based on the empirical risk minimization (ERM) for all previous observed actions a_s and rewards r_s as follows:

$$\boldsymbol{\theta}_t \leftarrow \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\text{argmin}}\, \lambda\|\boldsymbol{\theta}\|_2^2 + \textstyle\sum_{s\in[t]} w_s(\langle \boldsymbol{\theta}, a_s\rangle - r_s)^2, \tag{10}$$
$$\mathcal{C}_t \leftarrow \left\{\boldsymbol{\theta} \in \mathbb{R}^d \middle| \textstyle\sum_{s=1}^{t} w_s(\langle \boldsymbol{\theta}, a_s\rangle - \langle \boldsymbol{\theta}_t, a_s\rangle)^2 \leq \beta_t\right\},$$

where w_s is the weight, and β_t, λ are some parameters to be specified. w_s is selected in the order of the inverse of the variance σ_s^2 at round s to let the variance of the rescaled reward $\sqrt{w_s} r_s$ upper bounded by 1. Therefore, after the weighting step, one can regard the heteroscedastic bandit problem as a homoscedastic bandits problem and apply existing theoretical results to it. To deal with the general function case, a direct attempt is to replace the $\langle \boldsymbol{\theta}, a\rangle$ appearing in above construction rules with $f(a)$. However, such an approach requires that $\mathcal{F}$ is *close* under the linear mapping, which does not hold for general function class $\mathcal{F}$.

We propose our algorithm MOR-UCB as displayed in Algorithm 2. At the core of our design is the idea of partitioning the observed data into several layers and 'packing' data with similar variance upper bounds into the same layer as shown in line 7-8 of Algorithm 2. Specifically, for any two data belonging to the same layer, their variance will be at most one time larger than the other. Next in line 9, our algorithm implements FTRL to estimate f^* according to the data points in $\Psi_{t+1,l}$. Then in line 5, the agent makes use of L confidence sets simultaneously to select an action based on the *optimism-in-the-face-of-uncertainty* (OFU) principle over all L number of levels.

Remark 6.2. Prior to our work, Takemura et al. (2021) also adopts a multi-layer structure to solve misspecified contextual linear bandits. In the aforementioned paper, the

Algorithm 2 **M**ulti-layer **O**nline **R**egression-UCB

Input: $T, \lambda, R, \overline{\sigma} > 0$.
Initialize: Set $L \leftarrow \lceil \log_2 R/\overline{\sigma} \rceil$ and $\mathcal{C}_{1,l} \leftarrow \mathcal{F}$, $\Psi_{1,l} \leftarrow \varnothing$ for all $l \in [L]$.
for $t = 1 \cdots T$ **do**
  Observes $\mathcal{D}_t$.
  Choose $\mathbf{a}_t \leftarrow \underset{\mathbf{a} \in \mathcal{D}_t}{\operatorname{argmax}} \min_{l \in [L]} \phi\left(\widehat{\boldsymbol{\theta}}_{t,l}^\top \mathbf{a} + \beta_{t,l} \|\mathbf{a}\|_{\boldsymbol{\Sigma}_{t,l}^{-1}}\right)$.
  Observe stochastic reward r_t and σ_t^2.
  Find l_t such that $2^{l_t+1}\overline{\sigma} \geq \max(\overline{\sigma}, \sigma_t) \geq 2^{l_t}\overline{\sigma}$.
  Update $\Psi_{t+1,l_t} \leftarrow \Psi_{t,l_t} \cup \{t\}$ and $\Psi_{t+1,l} \leftarrow \Psi_{t,l}$ for all $l \in [L] \backslash \{l_t\}$.
  Compute $\widehat{\boldsymbol{\theta}}_{t+1,l} \leftarrow \underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\operatorname{argmin}} \lambda \|\boldsymbol{\theta}\|_2^2 + \sum_{\tau \in \Psi_{t+1,l}} \ell_\tau(\boldsymbol{\theta})$ for all $l \in [L]$, where ℓ_τ is defined following (2):

$$\ell_\tau(\boldsymbol{\theta}) := -\mathbf{a}_\tau^\top \boldsymbol{\theta} r_\tau + \int_0^{\mathbf{a}_\tau^\top \boldsymbol{\theta}} \phi(z) \mathrm{d}z.$$

  Compute $\boldsymbol{\Sigma}_{t+1,l} \leftarrow 2\lambda \mathbf{I} + \sum_{\tau \in \Psi_{t+1,l}} \mathbf{a}_\tau \mathbf{a}_\tau^\top$ for all $l \in [L]$.
end for

designed algorithm is a modified version of SupLinUCB algorithm (Chu et al., 2011). From the algorithm design perspective, SupLinUCB algorithm groups the selected contexts into different levels based on their uncertainty, while our algorithm groups the contexts based on the variances of their corresponding rewards. The difference in the algorithm design is due to the different goals: SupLinUCB algorithm aims to reduce the dependence on dimension from d to $\sqrt{d}$ for the finite arm case, while our algorithm aims to reduce the dependence on the variance from $R\sqrt{T}$ to $\sqrt{\sum_{t \in T} \sigma_t^2}$. Although the high-level structures of these two algorithms are similar, they are fundamentally different algorithms.

6.3. Theoretical Results

We provide the theoretical guarantee of MOR here.

Theorem 6.3 (Cumulative regret for generalized linear bandits)**.** *Suppose that* $\|\boldsymbol{\theta}^*\|_2 \leq 1$ *and for all* $\mathbf{a} \in \bigcup_{t \in [T]} \mathcal{D}_t$, $\|\mathbf{a}\|_2 \leq 1$. *Set* $\lambda = 4K^2/\kappa$ *and*

$$\beta_{t,l} = 16 \cdot 2^l \overline{\sigma} \kappa^{-1/2} \sqrt{d \log\left(\frac{2d\lambda + t\kappa A^2}{2d\lambda}\right) \log\left(4t^2 L/\delta\right)} + 4R \cdot \kappa^{-1/2} \log(4t^2 L/\delta) + 2\sqrt{2/\kappa} K$$

in Algorithm 2. With probability at least $1 - \delta$, *the regret of Algorithm 2 in the first* T *rounds satisfies that:*

$$\textit{Regret}(T) \leq \widetilde{O}\left(\frac{K}{\kappa} d \sqrt{\sum_{t=1}^T \sigma_t^2} + (R + K) K \cdot \kappa^{-1} \sqrt{dT}\right)$$

Remark 6.4. In the case of heteroscedastic linear bandits (Zhou et al., 2021) where $\kappa = K = 1$, the regret is bounded by $\widetilde{O}\left(d\sqrt{\sum_{t=1}^T \sigma_t^2} + (R+1)\sqrt{dT}\right)$, which matches with the result in Zhou & Gu (2022) by an $\widetilde{O}((R+1)\sqrt{dT})$ lower-order term. Applying a more fine-grained concentration bound (e.g., Zhou & Gu, 2022) may further remove this term, which we leave for future work.

7. Conclusion and Future Work

In this paper, we study the problem of stochastic online generalized linear regression and provide a novel analysis for FTRL, attaining an $O\left(\sigma^2 d \log T\right) + o(\log T)$ upper bound. In addition, we prove the first lower bound for online linear regression in the stochastic setting, indicating that our regret bound is minimax-optimal.

As an application, we further considered heteroscedastic generalized linear bandit problem. Applying parallel FTRL learners, we design a UCB-based algorithm MOR-UCB, which achieves a tighter instance-dependent regret bound in bandit setting. We believe that our analysis can also be applied to obtain variance-dependent regret bounds in MDP setting (Zhao et al., 2023; Zhou et al., 2023; 2022).

Although a near-optimal regret for stochastic online linear regression is achieved in this paper, the regret of stochastic online regression of general loss functions is still understudied, which we leave for future work.

Acknowledgements

We thank the anonymous reviewers for their helpful comments. HZ, DZ, JH and QG are supported in part by the National Science Foundation CAREER Award 1906169 and the Sloan Research Fellowship. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24:2312–2320, 2011.

Aitken, A. C. Iv.—on least squares and linear combination of observations. *Proceedings of the Royal Society of Edinburgh*, 55:42–48, 1936.

Antos, A., Grover, V., and Szepesvari, C. Active learning in heteroscedastic noise. *Theor. Comput. Sci.*, 411:2712–2728, 2010.

Assael, J.-A. M., Wang, Z., Shahriari, B., and de Freitas,

N. Heteroscedastic treed bayesian optimisation. *arXiv preprint arXiv:1410.7172*, 2014.

Azoury, K. S. and Warmuth, M. K. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43(3):211–246, 2001.

Baby, D. and Wang, Y.-X. Online forecasting of total-variation-bounded sequences. *Advances in Neural Information Processing Systems*, 32, 2019.

Bartlett, P. L., Koolen, W. M., Malek, A., Takimoto, E., and Warmuth, M. K. Minimax fixed-design linear regression. In *Conference on Learning Theory*, pp. 226–239. PMLR, 2015.

Besbes, O., Gur, Y., and Zeevi, A. Non-stationary stochastic optimization. *Operations research*, 63(5):1227–1244, 2015.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, learning, and games*. Cambridge university press, 2006.

Cesa-Bianchi, N., Long, P. M., and Warmuth, M. K. Worst-case quadratic loss bounds for prediction using linear functions and gradient descent. *IEEE Transactions on Neural Networks*, 7(3):604–619, 1996.

Chaudhuri, K., Jain, P., and Natarajan, N. Active heteroscedastic regression. In *International Conference on Machine Learning*, 2017.

Chu, W., Li, L., Reyzin, L., and Schapire, R. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 208–214. JMLR Workshop and Conference Proceedings, 2011.

Cowan, W., Honda, J., and Katehakis, M. N. Normal bandits of unknown means and variances: Asymptotic optimality, finite horizon regret bounds, and a solution to an open problem. *arXiv preprint arXiv:1504.05823*, 2015.

Dai, Y., Wang, R., and Du, S. S. Variance-aware sparse linear bandits. *arXiv preprint arXiv:2205.13450*, 2022.

Filippi, S., Cappe, O., Garivier, A., and Szepesvári, C. Parametric bandits: The generalized linear case. In *NIPS*, volume 23, pp. 586–594, 2010.

Foster, D. P. Prediction in the Worst Case. *The Annals of Statistics*, 19(2):1084 – 1090, 1991. doi: 10.1214/aos/1176348140.

Freedman, D. A. On tail probabilities for martingales. *the Annals of Probability*, pp. 100–118, 1975.

Gaillard, P., Gerchinovitz, S., Huard, M., and Stoltz, G. Uniform regret bounds over r^d for the sequential linear regression problem with the square loss. In *Algorithmic Learning Theory*, pp. 404–432. PMLR, 2019.

Goldberg, P. W., Williams, C. K., and Bishop, C. M. Regression with input-dependent noise: A gaussian process treatment. *Advances in neural information processing systems*, 10:493–499, 1997.

Herbster, M. and Warmuth, M. K. Tracking the best linear predictor. *Journal of Machine Learning Research*, 1(281-309):10–1162, 2001.

Hong, D., Balzano, L., and Fessler, J. A. Towards a theoretical analysis of pca for heteroscedastic data. In *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 496–503. IEEE, 2016.

Hong, D., Fessler, J. A., and Balzano, L. Optimally weighted pca for high-dimensional heteroscedastic data. *arXiv preprint arXiv:1810.12862*, 2018.

Jun, K.-S., Bhargava, A., Nowak, R., and Willett, R. Scalable generalized linear bandits: online computation and hashing. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 98–108, 2017.

Kersting, K., Plagemann, C., Pfaff, P., and Burgard, W. Most likely heteroscedastic gaussian process regression. In *Proceedings of the 24th international conference on Machine learning*, pp. 393–400, 2007.

Kirschner, J. and Krause, A. Information directed sampling and bandits with heteroscedastic noise. In *Conference On Learning Theory*, pp. 358–384. PMLR, 2018.

Kivinen, J. and Warmuth, M. K. Exponentiated gradient versus gradient descent for linear predictors. *information and computation*, 132(1):1–63, 1997.

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

Littlestone, N., Long, P. M., and Warmuth, M. K. On-line learning of linear functions. In *Proceedings of the twenty-third annual ACM symposium on Theory of computing*, pp. 465–475, 1991.

Malek, A. and Bartlett, P. L. Horizon-independent minimax linear regression. *Advances in Neural Information Processing Systems*, 31:5259–5268, 2018.

Mourtada, J. Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices. *The Annals of Statistics*, 50(4):2157–2178, 2022.

Nelder, J. A. and Wedderburn, R. W. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972.

Ouhamma, R., Maillard, O.-A., and Perchet, V. Stochastic online linear regression: the forward algorithm to replace ridge. *Advances in Neural Information Processing Systems*, 34:24430–24441, 2021.

Pinsker, M. S. and Feinstein, A. Information and information stability of random variables and processes. 1964.

Raj, A., Gaillard, P., and Saad, C. Non-stationary online regression. *arXiv preprint arXiv:2011.06957*, 2020.

Takemura, K., Ito, S., Hatano, D., Sumita, H., Fukunaga, T., Kakimura, N., and Kawarabayashi, K.-i. A parameter-free algorithm for misspecified linear contextual bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 3367–3375. PMLR, 2021.

Tirinzoni, A., Pirotta, M., Restelli, M., and Lazaric, A. An asymptotically optimal primal-dual incremental algorithm for contextual linear bandits. *Advances in Neural Information Processing Systems*, 33:1417–1427, 2020.

Vovk, V. Competitive on-line linear regression. In *NIPS*, 1997.

Zhang, L., Yang, T., Zhou, Z.-H., et al. Dynamic regret of strongly adaptive methods. In *International conference on machine learning*, pp. 5882–5891. PMLR, 2018.

Zhang, Z., Yang, J., Ji, X., and Du, S. S. Variance-aware confidence set: Variance-dependent bound for linear bandits and horizon-free bound for linear mixture mdp. *arXiv preprint arXiv:2101.12745*, 2021.

Zhao, H., He, J., Zhou, D., Zhang, T., and Gu, Q. Variance-dependent regret bounds for linear bandits and reinforcement learning: Adaptivity and computational efficiency. *arXiv preprint arXiv:2302.10371*, 2023.

Zhou, D. and Gu, Q. Computationally efficient horizon-free reinforcement learning for linear mixture mdps. *arXiv preprint arXiv:2205.11507*, 2022.

Zhou, D., Gu, Q., and Szepesvari, C. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pp. 4532–4576. PMLR, 2021.

Zhou, R., Wang, R., and Du, S. S. Horizon-free and variance-dependent reinforcement learning for latent markov decision processes. *arXiv preprint arXiv:2210.11604*, 2022.

Zhou, R., Zhang, Z., and Du, S. S. Sharp variance-dependent bounds in reinforcement learning: Best of both worlds in stochastic and deterministic environments. *arXiv preprint arXiv:2301.13446*, 2023.

Zinkevich, M. Online convex programming and generalized infinitesimal gradient ascent. icml'03, 2003.

A. Proofs from Section 4

A.1. Proof of Theorem 4.1

Lemma A.1 (Regret decomposition)**.** *For each $t \in [T]$, let $\mathcal{L}_t$ be the cumulative loss function defined in (5) and $\mathbf{H}_t$ be the corresponding Hessian matrix as shown in (6). There exists a sequence $\{\boldsymbol{\mu}'_t\}_{t\in[T]}$ in $\mathbb{R}^d$ such that the stochastic regret of Algorithm 1 can be decomposed as follows:*

$$\mathcal{R}^{stoc}(T) \leq \lambda B^2 + \frac{1}{2}\sum_{t=1}^{T}\left(\phi(\mathbf{x}_t^\top \widehat{\boldsymbol{\mu}}_t) - \phi(\mathbf{x}_t^\top \boldsymbol{\mu}^*)\right)^2 \|\mathbf{x}_t\|^2_{\mathbf{H}_t^{-1}(\boldsymbol{\mu}'_t)} + \frac{1}{2}\sum_{t=1}^{T}\epsilon_t^2\|\mathbf{x}_t\|^2_{\mathbf{H}_t^{-1}(\boldsymbol{\mu}'_t)}.$$

Proof. From the updating rule of Algorithm 1,

$$\begin{aligned}
\mathcal{R}^{\text{stoc}}(T) &= \sum_{t=1}^{T}\ell_t(\widehat{\boldsymbol{\mu}}_t) - \sum_{t=1}^{T}\ell_t(\boldsymbol{\mu}^*) \\
&= \sum_{t=1}^{T}\left[\ell_t(\widehat{\boldsymbol{\mu}}_t) + \sum_{\tau=1}^{t-1}\ell_\tau(\widehat{\boldsymbol{\mu}}_t) - \sum_{\tau=1}^{t}\ell_\tau(\widehat{\boldsymbol{\mu}}_{t+1})\right] + \sum_{t=1}^{T}\ell_t(\widehat{\boldsymbol{\mu}}_{t+1}) - \sum_{t=1}^{T}\ell_t(\boldsymbol{\mu}^*) \\
&= \sum_{t=1}^{T}\left[\mathcal{L}_t(\widehat{\boldsymbol{\mu}}_t) - \mathcal{L}_t(\widehat{\boldsymbol{\mu}}_{t+1}) - \phi(\widehat{\boldsymbol{\mu}}_t) + \phi(\widehat{\boldsymbol{\mu}}_{t+1})\right] + \sum_{t=1}^{T}\ell_t(\widehat{\boldsymbol{\mu}}_{t+1}) - \sum_{t=1}^{T}\ell_t(\boldsymbol{\mu}^*) \\
&= \sum_{t=1}^{T}\left[\mathcal{L}_t(\widehat{\boldsymbol{\mu}}_t) - \mathcal{L}_t(\widehat{\boldsymbol{\mu}}_{t+1})\right] + \mathcal{L}_T(\widehat{\boldsymbol{\mu}}_{t+1}) - \mathcal{L}_T(\boldsymbol{\mu}^*) + \lambda\|\boldsymbol{\mu}^*\|_2^2 - \lambda\|\widehat{\boldsymbol{\mu}}_1\|_2^2 \\
&\leq \lambda\|\boldsymbol{\mu}^*\|_2^2 + \sum_{t=1}^{T}\left[\mathcal{L}_t(\widehat{\boldsymbol{\mu}}_t) - \mathcal{L}_t(\widehat{\boldsymbol{\mu}}_{t+1})\right], \qquad (11)
\end{aligned}$$

where the third equality holds due to the definition of $\mathcal{L}$ in (5),

Applying Taylor expansion, we have

$$\begin{aligned}
\mathcal{L}_t(\widehat{\boldsymbol{\mu}}_t) - \mathcal{L}_t(\widehat{\boldsymbol{\mu}}_{t+1}) &= \left\langle \frac{\partial\mathcal{L}_t}{\partial\boldsymbol{\mu}}(\widehat{\boldsymbol{\mu}}_t), \widehat{\boldsymbol{\mu}}_t - \widehat{\boldsymbol{\mu}}_{t+1}\right\rangle - (\widehat{\boldsymbol{\mu}}_{t+1} - \widehat{\boldsymbol{\mu}}_t)^\top \mathbf{H}_t(\boldsymbol{\mu}'_t)(\widehat{\boldsymbol{\mu}}_{t+1} - \widehat{\boldsymbol{\mu}}_t) \\
&= \left\langle \left(\phi(\mathbf{x}_t^\top\widehat{\boldsymbol{\mu}}_t) - y_t\right)\mathbf{x}_t, \widehat{\boldsymbol{\mu}}_t - \widehat{\boldsymbol{\mu}}_{t+1}\right\rangle - (\widehat{\boldsymbol{\mu}}_{t+1} - \widehat{\boldsymbol{\mu}}_t)^\top \mathbf{H}_t(\boldsymbol{\mu}'_t)(\widehat{\boldsymbol{\mu}}_{t+1} - \widehat{\boldsymbol{\mu}}_t) \qquad (12)
\end{aligned}$$

for some $\boldsymbol{\mu}'_t \in \mathbb{R}^d, \|\boldsymbol{\mu}'_t\|_2 \leq B$.

Substituting (12) into (11),

$$\begin{aligned}
\mathcal{R}^{\text{stoc}}(T) &\leq \lambda B^2 + \sum_{t=1}^{T}\left[\left\langle \left(\phi(\mathbf{x}_t^\top\widehat{\boldsymbol{\mu}}_t) - y_t\right)\mathbf{x}_t, \widehat{\boldsymbol{\mu}}_t - \widehat{\boldsymbol{\mu}}_{t+1}\right\rangle - (\widehat{\boldsymbol{\mu}}_{t+1} - \widehat{\boldsymbol{\mu}}_t)^\top \mathbf{H}_t(\boldsymbol{\mu}'_t)(\widehat{\boldsymbol{\mu}}_{t+1} - \widehat{\boldsymbol{\mu}}_t)\right] \\
&\leq \lambda B^2 + \frac{1}{4}\sum_{t=1}^{T}\left(\phi(\mathbf{x}_t^\top\widehat{\boldsymbol{\mu}}_t) - y_t\right)^2\|\mathbf{x}_t\|^2_{\mathbf{H}_t^{-1}(\boldsymbol{\mu}'_t)} \\
&\leq \lambda B^2 + \frac{1}{2}\sum_{t=1}^{T}\left(\phi(\mathbf{x}_t^\top\widehat{\boldsymbol{\mu}}_t) - \phi(\mathbf{x}_t^\top\boldsymbol{\mu}^*)\right)^2\|\mathbf{x}_t\|^2_{\mathbf{H}_t^{-1}(\boldsymbol{\mu}'_t)} + \frac{1}{2}\sum_{t=1}^{T}\epsilon_t^2\|\mathbf{x}_t\|^2_{\mathbf{H}_t^{-1}(\boldsymbol{\mu}'_t)}.
\end{aligned}$$

□

Lemma A.2 (Connection between squared estimation error and regret)**.** *Consider an arbitrary online learner interactively trained with stochastic data for T rounds as described in Section 3. If Condition 3.2 is true for the noise at all the rounds $t \in [T]$, then the following inequality holds with probability at least $1-\delta$:*

$$\sum_{t=1}^{T}\left[\mathbf{x}_t^\top(\widehat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}^*)\right]^2 \leq \frac{4}{\kappa}\mathcal{R}^{stoc}(T) + \frac{16}{\kappa^2}\cdot\sigma^2\log(1/\delta).$$

Proof. We start by considering the definition of stochastic regret and making use of the property of our aforementioned loss function:

$$\begin{aligned}
\mathcal{R}^{\text{stoc}}(T) &= \sum_{t=1}^{T} \ell_t(\widehat{\boldsymbol{\mu}}_t) - \sum_{t=1}^{T} \ell_t(\boldsymbol{\mu}^*) \\
&= -\sum_{t=1}^{T} \mathbf{x}_t^\top (\widehat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}^*) y_t + \sum_{t=1}^{T} \int_{\mathbf{x}_t^\top \boldsymbol{\mu}^*}^{\mathbf{x}_t^\top \widehat{\boldsymbol{\mu}}_t} \phi(z) \mathrm{d}z \\
&\geq -\sum_{t=1}^{T} \mathbf{x}_t^\top (\widehat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}^*) y_t + \sum_{t=1}^{T} \int_{\mathbf{x}_t^\top \boldsymbol{\mu}^*}^{\mathbf{x}_t^\top \widehat{\boldsymbol{\mu}}_t} \left(\phi(\mathbf{x}_t^\top \boldsymbol{\mu}^*) + \kappa(z - \mathbf{x}_t^\top \boldsymbol{\mu}^*) \right) \mathrm{d}z \\
&= -\sum_{t=1}^{T} \mathbf{x}_t^\top (\widehat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}^*) \epsilon_t + \sum_{t=1}^{T} \frac{1}{2} \cdot \kappa \left[\mathbf{x}_t^\top (\widehat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}^*) \right]^2, \qquad (13)
\end{aligned}$$

where the first equality follows from the definition of regret (3), the second equality follows from the definition of loss function (2), the inequality holds due to Assumption 3.1.

Rearranging (13), it follows that

$$\sum_{t=1}^{T} \left[\mathbf{x}_t^\top (\widehat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}^*) \right]^2 \leq \frac{2}{\kappa} \mathcal{R}^{\text{stoc}}(T) + \frac{2}{\kappa} \sum_{t=1}^{T} \epsilon_t \cdot \left[\mathbf{x}_t^\top (\widehat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}^*) \right]. \qquad (14)$$

Since $\widehat{\boldsymbol{\mu}}_t$ is $(\mathbf{x}_{1:t}, y_{1:t-1})$-measurable, we can apply Lemma D.3 to show that

$$\sum_{t=1}^{T} \epsilon_t \cdot \left[\mathbf{x}_t^\top (\widehat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}^*) \right] \leq \sqrt{2\sigma^2 \log(1/\delta) \sum_{t=1}^{T} \left[\mathbf{x}_t^\top (\widehat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}^*) \right]^2} \qquad (15)$$

with probability at least $1 - \delta$.

Substituting (15) into (14), we obtain the following high-probability bound for squared estimation error:

$$\begin{aligned}
\sum_{t=1}^{T} \left[\mathbf{x}_t^\top (\widehat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}^*) \right]^2 &\leq \frac{2}{\kappa} \mathcal{R}^{\text{stoc}}(T) + \frac{2}{\kappa} \sqrt{2\sigma^2 \log(1/\delta) \sum_{t=1}^{T} \left[\mathbf{x}_t^\top (\widehat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}^*) \right]^2} \\
&\leq \frac{4}{\kappa} \mathcal{R}^{\text{stoc}}(T) + \frac{16}{\kappa^2} \cdot \sigma^2 \log(1/\delta),
\end{aligned}$$

where the last inequality follows from Lemma D.2. □

Lemma A.3. *Suppose that the sequence of noise $\{\epsilon_t\}_{t \in [T]}$ satisfies Condition 3.2. For each $t \in [T]$, let $\underline{\mathbf{H}}_t$ be the matrix defined in (7). With probability at least $1 - \delta$,*

$$\sum_{t=1}^{T} \epsilon_t^2 \|\mathbf{x}_t\|^2_{\underline{\mathbf{H}}_t^{-1}} \leq 34\kappa^{-1} \cdot \sigma^2 d \log \frac{d\lambda + T\kappa A^2}{d\lambda} + 24\sigma^2 \cdot \frac{A^2}{\lambda} \log(1/\delta).$$

Proof. We first prove that the random variable ϵ_t^2 is sub-exponential conditioning on $\mathbf{x}_{1:t}, y_{1:t-1}$.

Let $v_t = \mathbb{E}[\epsilon_t^2]$. Considering the moment generating function of ϵ_t^2, we have for all $s \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{E}[\exp(s(\epsilon_t^2 - v_t))] &= 1 + s\mathbb{E}[\epsilon_t^2 - v_t] + \sum_{i=2}^{\infty} \frac{s^i}{i!} \mathbb{E}\left[\left(\epsilon_t^2 - v_t \right)^i \right] \\
&\leq 1 + \sum_{i=2}^{\infty} \frac{s^i}{i!} \mathbb{E}\left[\epsilon_t^{2i} \right].
\end{aligned}$$

For sub-Gaussian noise ϵ_t, we have

$$\begin{aligned}\mathbb{E}\left[|\epsilon_t|^r\right] &= \int_0^\infty \mathbb{P}\left(|\epsilon_t|^r \geq x\right) \mathrm{d}x \\ &= r\int_0^\infty x^{r-1}\mathbb{P}\left(|\epsilon_t| \geq x\right) \mathrm{d}x \\ &\leq 2r\int_0^\infty x^{r-1}\exp\left(-\frac{x^2}{2\sigma^2}\right) \mathrm{d}x \\ &= 2^{r/2}\cdot r\sigma^r \int_0^\infty x^{\frac{r}{2}-1}\exp(-x)\mathrm{d}x \\ &= 2^{r/2}\cdot r\sigma^r \cdot \Gamma(r/2).\end{aligned}$$

Hence, we have

$$\begin{aligned}\mathbb{E}[\exp(s(\epsilon_t^2 - v_t))] &\leq 1 + \sum_{i=2}^\infty 2i \cdot \frac{s^i}{i!} 2^i\sigma^{2i}(i-1)! \\ &\leq 1 + \sum_{i=2}^\infty 2(2s\sigma^2)^i \\ &= 1 + \frac{8s^2\sigma^4}{1-2s\sigma^2},\end{aligned}$$

which implies that $\epsilon_t^2 - v_t$ is $\left(\left(4\sqrt{2}\sigma^2\right)^2, 4\sigma^2\right)$-sub-exponential.

By the composition property of sub-exponential random variables, we have

$$\sum_{t=1}^T \epsilon_t^2 \|\mathbf{x}_t\|_{\underline{\mathbf{H}}_t^{-1}}^2 \sim \mathrm{SE}\left(32\sigma^4\sum_{t=1}^T \|\mathbf{x}_t\|_{\underline{\mathbf{H}}_t^{-1}}^4, 4\sigma^2 \max_{t\in[T]} \|\mathbf{x}_t\|_{\underline{\mathbf{H}}_t^{-1}}^2\right).$$

By Lemma D.6, the following concentration bound holds with probability at least $1-\delta$:

$$\begin{aligned}&\sum_{t=1}^T \epsilon_t^2 \|\mathbf{x}_t\|_{\underline{\mathbf{H}}_t^{-1}}^2 - \sum_{t=1}^T \mathbb{E}\left[\epsilon_t^2\right] \|\mathbf{x}_t\|_{\underline{\mathbf{H}}_t^{-1}}^2 \\ &\qquad \leq \max\left\{8\cdot\sqrt{\log(1/\delta)}\cdot\sigma^2\sqrt{\sum_{t=1}^T \|\mathbf{x}_t\|_{\underline{\mathbf{H}}_t^{-1}}^4},\ 8\sigma^2 \max_{t\in[T]} \|\mathbf{x}_t\|_{\underline{\mathbf{H}}_t^{-1}}^2 \cdot \log(1/\delta)\right\}. \qquad (16)\end{aligned}$$

Applying Lemma D.5 and the definition of $\underline{\mathbf{H}}$, we further have

$$\sum_{t=1}^T \epsilon_t^2 \|\mathbf{x}_t\|_{\underline{\mathbf{H}}_t^{-1}}^2 - \sum_{t=1}^T \mathbb{E}\left[\epsilon_t^2\right] \|\mathbf{x}_t\|_{\underline{\mathbf{H}}_t^{-1}}^2 \leq 8\sigma^2\sqrt{\frac{2A^2d}{\lambda}\kappa^{-1}\log\left(\frac{d\lambda+\kappa TA^2}{d\lambda}\right)\log(1/\delta)} + 8\sigma^2\frac{A^2}{\lambda}\log(1/\delta)$$

with probability at least $1-\delta$.

Since ϵ_t is σ-sub-Gaussian, its variance is no larger than σ^2, which indicates that

$$\begin{aligned}\sum_{t=1}^T \epsilon_t^2 \|\mathbf{x}_t\|_{\underline{\mathbf{H}}_t^{-1}}^2 &\leq \left(\sum_{t=1}^T \epsilon_t^2 \|\mathbf{x}_t\|_{\underline{\mathbf{H}}_t^{-1}}^2 - \sum_{t=1}^T \mathbb{E}\left[\epsilon_t^2\right] \|\mathbf{x}_t\|_{\underline{\mathbf{H}}_t^{-1}}^2\right) + \sigma^2\sum_{t=1}^T \|\mathbf{x}_t\|_{\underline{\mathbf{H}}_t^{-1}}^2 \\ &\leq 8\sigma^2\cdot\kappa^{-1/2}\sqrt{\frac{2A^2d}{\lambda}\log\left(\frac{d\lambda+TA^2}{d\lambda}\right)\log(1/\delta)} + 8\sigma^2\cdot\frac{A^2}{\lambda}\log(1/\delta) \\ &\quad + \sigma^2\cdot 2d\kappa^{-1}\log\frac{d\lambda+T\kappa A^2}{d\lambda} \\ &\leq 34\kappa^{-1}\cdot\sigma^2 d\log\frac{d\lambda+T\kappa A^2}{d\lambda} + 24\sigma^2\cdot\frac{A^2}{\lambda}\log(1/\delta)\end{aligned}$$

with probability at least $1-\delta$.

□

Proof of Theorem 4.1. Based on Lemma 4.8 and Lemma 4.9, the following two inequalities hold simultaneously with probability at least $1-2\delta$ for all $\delta \in (0, \frac{1}{2})$:

$$\sum_{t=1}^{T} \left[\mathbf{x}_t^\top (\widehat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}^*)\right]^2 \leq \frac{4}{\kappa}\mathcal{R}^{\text{stoc}}(T) + \frac{16}{\kappa^2} \cdot \sigma^2 \log(1/\delta), \tag{17}$$

$$\sum_{t=1}^{T} \epsilon_t^2 \|\mathbf{x}_t\|^2_{\underline{\mathbf{H}}_t^{-1}} \leq 34\kappa^{-1} \cdot \sigma^2 d \log \frac{d\lambda + T\kappa A^2}{d\lambda} + 24\sigma^2 \cdot \frac{A^2}{\lambda} \log(1/\delta). \tag{18}$$

In the remaining proof, we assume that (17) and (18) hold.

From Lemma 4.7, we have

$$\begin{aligned}
\mathcal{R}^{\text{stoc}}(T) &\leq \lambda B^2 + \frac{1}{2}\sum_{t=1}^{T} \left(\phi(\mathbf{x}_t^\top \widehat{\boldsymbol{\mu}}_t) - \phi(\mathbf{x}_t^\top \boldsymbol{\mu}^*)\right)^2 \|\mathbf{x}_t\|^2_{\mathbf{H}_t^{-1}(\boldsymbol{\mu}'_t)} + \frac{1}{2}\sum_{t=1}^{T} \epsilon_t^2 \|\mathbf{x}_t\|^2_{\mathbf{H}_t^{-1}(\boldsymbol{\mu}'_t)} \\
&\leq \lambda B^2 + \frac{A^2K^2}{2\lambda}\sum_{t=1}^{T} \left[\mathbf{x}_t^\top (\widehat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}^*)\right]^2 + \frac{1}{2}\sum_{t=1}^{T} \epsilon_t^2 \|\mathbf{x}_t\|^2_{\mathbf{H}_t^{-1}(\boldsymbol{\mu}'_t)} \\
&\leq \lambda B^2 + \frac{2A^2K^2}{\lambda\kappa}\mathcal{R}^{\text{stoc}}(T) + \frac{8A^2K^2}{\lambda\kappa^2}\sigma^2 \log(1/\delta) \\
&\quad + 17\kappa^{-1} \cdot \sigma^2 d \log \frac{d\lambda + T\kappa A^2}{d\lambda} + 12\sigma^2 \cdot \frac{A^2}{\lambda} \log(1/\delta)
\end{aligned} \tag{19}$$

where the first inequality is given by Lemma 4.8 directly, the second inequality follows from the Lipschitz property of the activation function in Assumption 3.1, the third inequality holds due to (17), (18) and the fact that $\underline{\mathbf{H}}_t \preceq \mathbf{H}_t^{-1}(\boldsymbol{\mu}'_t)$.

Substituting $\lambda = 4A^2K^2/\kappa$ into (19), we have

$$\begin{aligned}
\mathcal{R}^{\text{stoc}}(T) &\leq \frac{1}{2}\mathcal{R}^{\text{stoc}}(T) + \left(\frac{2}{\kappa} + \frac{3\kappa}{K^2}\right)\sigma^2 \log(1/\delta) + 17\kappa^{-1} \cdot \sigma^2 d \log \frac{4dK^2 + T\kappa^2}{4dK^2} + 4\frac{K^2A^2B^2}{\kappa} \\
&\leq 34\kappa^{-1} \cdot \sigma^2 d \log \frac{4dK^2 + T\kappa^2}{4dK^2} + 8\frac{K^2A^2B^2}{\kappa} + 2\left(\frac{2}{\kappa} + \frac{3\kappa}{K^2}\right)\sigma^2 \log(1/\delta),
\end{aligned}$$

which completes the proof. □

Theorem A.4 (Theorem 3.1, Ouhamma et al. 2021)**.** *In stochastic online linear regression ($\kappa = K = 1$) in Assumption 3.1 with Condition 3.2, we have with probability at least $1-\delta$,*

$$\mathcal{R}^{adv}(T) - \mathcal{R}^{stoc}(T) \leq O(\sigma^2 d \log T) + o(\log T).$$

B. Proofs from Section 4.3

B.1. Proof of Theorem 4.5

Lemma B.1 (Connection between squared estimation error and regret)**.** *Consider an arbitrary online learner interactively trained with stochastic data for T rounds as described in Section 3. If Condition 3.3 is true for the noise at all the rounds $t \in [T]$, then the following inequality holds with probability at least $1-\delta$:*

$$\sum_{t=1}^{T} \left[\mathbf{x}_t^\top (\widehat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}^*)\right]^2 \leq \frac{4}{\kappa}\mathcal{R}^{stoc}(T) + \frac{16}{\kappa^2} \cdot R^2 \log(1/\delta).$$

Lemma B.2. *Suppose that the sequence of noise $\{\epsilon_t\}_{t\in[T]}$ satisfies Condition 3.3. For each $t \in [T]$, let $\underline{\mathbf{H}}_t$ be the matrix defined in (7). With probability at least $1-\delta$,*

$$\sum_{t=1}^{T} \epsilon_t^2 \|\mathbf{x}_t\|^2_{\underline{\mathbf{H}}_t^{-1}} \le 6\kappa^{-1} \cdot \sigma^2 d \log \frac{d\lambda + T\kappa A^2}{d\lambda} + \frac{5}{3} \cdot \frac{R^2 A^2}{\lambda} \log(1/\delta).$$

Proof. We prove this lemma by bounding $\sum_{t=1}^{T} \mathbb{E}[\epsilon_t^2] \|\mathbf{x}_t\|^2_{\underline{\mathbf{H}}_t^{-1}}$ and $\sum_{t=1}^{T} \left(\epsilon_t^2 - \mathbb{E}[\epsilon_t^2]\right) \|\mathbf{x}_t\|^2_{\underline{\mathbf{H}}_t^{-1}}$ separately.

For the first term, we have

$$\sum_{t=1}^{T} \mathbb{E}[\epsilon_t^2] \|\mathbf{x}_t\|^2_{\underline{\mathbf{H}}_t^{-1}} \le \sum_{t=1}^{T} \sigma^2 \|\mathbf{x}_t\|^2_{\underline{\mathbf{H}}_t^{-1}} \le 2\kappa^{-1} \cdot \sigma^2 d \log \frac{d\lambda + T\kappa A^2}{d\lambda}. \tag{20}$$

For the second term $\sum_{t=1}^{T} \left(\epsilon_t^2 - \mathbb{E}[\epsilon_t^2]\right) \|\mathbf{x}_t\|^2_{\underline{\mathbf{H}}_t^{-1}}$, it holds that

$$\begin{aligned}
\mathbb{E}\left[\left(\epsilon_t^2 - \mathbb{E}[\epsilon_t^2]\right) \|\mathbf{x}_t\|^2_{\underline{\mathbf{H}}_t^{-1}}\right] &= 0, \\
\sum_{t=1}^{T} \mathrm{Var}\left[\left(\epsilon_t^2 - \mathbb{E}[\epsilon_t^2]\right) \|\mathbf{x}_t\|^2_{\underline{\mathbf{H}}_t^{-1}}\right] &\le \sum_{t=1}^{T} \mathbb{E}[\epsilon_t^4] \|\mathbf{x}_t\|^4_{\underline{\mathbf{H}}_t^{-1}} \\
&\le R^2\sigma^2 \frac{A^2}{\lambda} \sum_{t=1}^{T} \|\mathbf{x}_t\|^2_{\underline{\mathbf{H}}_t^{-1}} \\
&\le 2\kappa^{-1}\lambda^{-1} \cdot R^2\sigma^2 A^2 d \log \frac{d\lambda + T\kappa A^2}{d\lambda}, \\
\left|\left(\epsilon_t^2 - \mathbb{E}[\epsilon_t^2]\right) \|\mathbf{x}_t\|^2_{\underline{\mathbf{H}}_t^{-1}}\right| &\le R^2 \cdot \frac{A^2}{\lambda}.
\end{aligned}$$

Applying Lemma D.1, with probability at least $1-\delta$,

$$\begin{aligned}
\sum_{t=1}^{T} \left(\epsilon_t^2 - \mathbb{E}[\epsilon_t^2]\right) \|\epsilon_t\|^2_{\underline{\mathbf{H}}_t^{-1}} &\le 2R\sigma A \sqrt{\kappa^{-1}\lambda^{-1} d \log \frac{d\lambda + T\kappa A^2}{d\lambda} \log(1/\delta)} + \frac{2}{3} \cdot \frac{R^2 A^2}{\lambda} \log(1/\delta) \\
&\le 4\sigma^2\kappa^{-1} \cdot d \log \frac{d\lambda + T\kappa A^2}{d\lambda} + \frac{5}{3} \cdot \frac{R^2 A^2}{\lambda} \log(1/\delta).
\end{aligned} \tag{21}$$

Combining (20) with (21), we can show that with probability at least $1-\delta$,

$$\sum_{t=1}^{T} \epsilon_t^2 \|\mathbf{x}_t\|^2_{\underline{\mathbf{H}}_t^{-1}} \le 6\kappa^{-1} \cdot \sigma^2 d \log \frac{d\lambda + T\kappa A^2}{d\lambda} + \frac{5}{3} \cdot \frac{R^2 A^2}{\lambda} \log(1/\delta).$$

□

Proof of Theorem 4.5. Based on Lemma B.1 and Lemma B.2, the following two inequalities hold simultaneously with probability at least $1-2\delta$ for all $\delta \in (0, \frac{1}{2})$:

$$\sum_{t=1}^{T} \left[\mathbf{x}_t^\top (\widehat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}^*)\right]^2 \le \frac{4}{\kappa} \mathcal{R}^{\mathrm{stoc}}(T) + \frac{16}{\kappa^2} \cdot R^2 \log(1/\delta), \tag{22}$$

$$\sum_{t=1}^{T} \epsilon_t^2 \|\mathbf{x}_t\|^2_{\underline{\mathbf{H}}_t^{-1}} \le 6\kappa^{-1} \cdot \sigma^2 d \log \frac{d\lambda + T\kappa A^2}{d\lambda} + \frac{5}{3} \cdot \frac{R^2 A^2}{\lambda} \log(1/\delta). \tag{23}$$

In the remaining proof, we assume that (17) and (18) hold.

From Lemma 4.7, we have

$$\begin{aligned}
\mathcal{R}^{\text{stoc}}(T) &\leq \lambda B^2 + \frac{1}{2}\sum_{t=1}^{T}\left(\phi(\mathbf{x}_t^\top \widehat{\boldsymbol{\mu}}_t) - \phi(\mathbf{x}_t^\top \boldsymbol{\mu}^*)\right)^2 \|\mathbf{x}_t\|^2_{\mathbf{H}_t^{-1}(\boldsymbol{\mu}_t')} + \frac{1}{2}\sum_{t=1}^{T}\epsilon_t^2 \|\mathbf{x}_t\|^2_{\mathbf{H}_t^{-1}(\boldsymbol{\mu}_t')} \\
&\leq \lambda B^2 + \frac{A^2K^2}{2\lambda}\sum_{t=1}^{T}\left[\mathbf{x}_t^\top(\widehat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}^*)\right]^2 + \frac{1}{2}\sum_{t=1}^{T}\epsilon_t^2 \|\mathbf{x}_t\|^2_{\mathbf{H}_t^{-1}(\boldsymbol{\mu}_t')} \\
&\leq \lambda B^2 + \frac{2A^2K^2}{\lambda\kappa}\mathcal{R}^{\text{stoc}}(T) + \frac{8A^2K^2}{\lambda\kappa^2}R^2\log(1/\delta) \\
&\quad + 3\kappa^{-1}\cdot\sigma^2 d\log\frac{d\lambda + T\kappa A^2}{d\lambda} + \frac{5}{6}\cdot\frac{R^2A^2}{\lambda}\log(1/\delta)
\end{aligned} \tag{24}$$

where the first inequality is given by Lemma 4.8 directly, the second inequality follows from the Lipschitz property of the activation function in Assumption 3.1, the third inequality holds due to (22), (18) and the fact that $\underline{\mathbf{H}}_t \preceq \mathbf{H}_t^{-1}(\boldsymbol{\mu}_t')$.

Substituting $\lambda = 4A^2K^2/\kappa$ into (19), we have

$$\begin{aligned}
\mathcal{R}^{\text{stoc}}(T) &\leq \frac{1}{2}\mathcal{R}^{\text{stoc}}(T) + \left(\frac{2}{\kappa} + \frac{5\kappa}{24K^2}\right)R^2\log(1/\delta) + 3\kappa^{-1}\cdot\sigma^2 d\log\frac{4dK^2 + T\kappa^2}{4dK^2} + 4\frac{K^2A^2B^2}{\kappa} \\
&\leq 6\kappa^{-1}\cdot\sigma^2 d\log\frac{4dK^2 + T\kappa^2}{4dK^2} + 8\frac{K^2A^2B^2}{\kappa} + 2\left(\frac{2}{\kappa} + \frac{5\kappa}{24K^2}\right)R^2\log(1/\delta),
\end{aligned}$$

which completes the proof. □

Theorem B.3 (Confidence ellipsoid for ridge regression estimator)**.** *Set $\lambda = 4A^2K^2/\kappa$ and assume that the noise ϵ_t satisfy Condition 3.3 at all rounds $t \in [T]$, then with probability at least $1-\delta$, for all $t \in [T]$, it holds that*

$$\|\widehat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}^*\|_{\underline{\mathbf{H}}_t} \leq 8\sigma\sqrt{d\log\left(\frac{2d\lambda + t\kappa A^2}{2d\lambda}\right)\log\left(4t^2/\delta\right)} + 4R\log(4t^2/\delta) + \sqrt{2\lambda}B.$$

Remark B.4. This theorem elucidates how to construct a confidence ellipsoid with predictions given by FTRL. Similar variance-aware confidence sets have been shown by Zhou et al. (2021); Zhang et al. (2021) in linear regression, while Theorem B.3 is applicable to generalized linear function class. Later in section 6, we will show how to make use of this theorem in bandit setting.

Proof. According to Algorithm 1, $\boldsymbol{\mu}_{t+1}$ is the minimizer of $\lambda\|\boldsymbol{\mu}\|_2^2 + \sum_{\tau=1}^{t}\ell_\tau(\boldsymbol{\mu})$.

Taking the derivative, we have

$$\begin{aligned}
0 &= 2\lambda\widehat{\boldsymbol{\mu}}_{t+1} + \sum_{\tau\in[t]}\left(-y_\tau\cdot\mathbf{x}_\tau + \phi(\mathbf{x}_\tau^\top\widehat{\boldsymbol{\mu}}_{t+1})\cdot\mathbf{x}_\tau\right) \\
&= 2\lambda\widehat{\boldsymbol{\mu}}_{t+1} + \sum_{\tau\in[t]}\left(-y_\tau + \phi(\mathbf{x}_\tau^\top\boldsymbol{\mu}^*)\right)\cdot\mathbf{x}_\tau + \sum_{\tau\in[t]}\left(\phi(\mathbf{x}_\tau^\top\widehat{\boldsymbol{\mu}}_{t+1}) - \phi(\mathbf{x}_\tau^\top\boldsymbol{\mu}^*)\right)\cdot\mathbf{x}_\tau
\end{aligned}$$

Rearranging the equality,

$$\sum_{\tau\in[t]}\left(\phi(\mathbf{x}_\tau^\top\widehat{\boldsymbol{\mu}}_{t+1}) - \phi(\mathbf{x}_\tau^\top\boldsymbol{\mu}^*)\right)\cdot\mathbf{x}_\tau + 2\lambda\left(\widehat{\boldsymbol{\mu}}_{t+1} - \boldsymbol{\mu}^*\right) = \sum_{\tau\in[t]}\epsilon_\tau\mathbf{x}_\tau - 2\lambda\cdot\boldsymbol{\mu}^*$$

For short, we let $\kappa_{\tau,t+1} = \frac{\phi(\mathbf{x}_\tau^\top\widehat{\boldsymbol{\mu}}_{t+1}) - \phi(\mathbf{x}_\tau^\top\boldsymbol{\mu}^*)}{\mathbf{x}_\tau^\top\widehat{\boldsymbol{\mu}}_{t+1} - \mathbf{x}_\tau^\top\boldsymbol{\mu}^*}$. By Assumption 3.1, $\kappa_{\tau,t+1} \in [\kappa, K]$.

Thus, we have

$$\left\|\left(2\lambda\cdot\mathbf{I}+\sum_{\tau\in[t]}\kappa_{\tau,t+1}\mathbf{x}_\tau\mathbf{x}_\tau^\top\right)\cdot(\widehat{\boldsymbol{\mu}}_{t+1}-\boldsymbol{\mu}^*)\right\|_{\underline{\mathbf{H}}_t^{-1}} = \left\|\sum_{\tau\in[t]}\epsilon_\tau\mathbf{x}_\tau-2\lambda\cdot\boldsymbol{\mu}^*\right\|_{\underline{\mathbf{H}}_t^{-1}}$$
$$\leq \left\|\sum_{\tau\in[t]}\epsilon_\tau\mathbf{x}_\tau\right\|_{\underline{\mathbf{H}}_t^{-1}}+\sqrt{2\lambda}\cdot\|\boldsymbol{\mu}^*\|_2\,.$$

Since $2\lambda\cdot\mathbf{I}+\sum_{\tau\in[t]}\kappa_{\tau,t+1}\mathbf{x}_\tau\mathbf{x}_\tau^\top\succeq\underline{\mathbf{H}}_t$, with probability at least $1-\delta$, for all $t\geq 1$, it holds that

$$\|\widehat{\boldsymbol{\mu}}_{t+1}-\boldsymbol{\mu}^*\|_{\underline{\mathbf{H}}_t^{-1}}\leq\left\|\sum_{\tau\in[t]}\epsilon_\tau\mathbf{x}_\tau\right\|_{\underline{\mathbf{H}}_t^{-1}}+\sqrt{2\lambda}\cdot\|\boldsymbol{\mu}^*\|_2$$
$$\leq 8\sigma\cdot\kappa^{-1/2}\sqrt{d\log\left(\frac{2d\lambda+t\kappa A^2}{2d\lambda}\right)\log\left(4t^2/\delta\right)}+4\cdot\kappa^{-1/2}R\log(4t^2/\delta)+\sqrt{2\lambda}B,$$

where the second inequality holds due to Theorem 4.1 in Zhou et al. (2021). □

C. Proofs from Section 6

Lemma C.1 (Variance-aware confidence ellipsoid for generalized linear bandits)**.** *Suppose that $\|\boldsymbol{\theta}^*\|_2\leq 1$ and for all $\mathbf{a}\in\bigcup_{t\in[T]}\mathcal{D}_t$, $\|\mathbf{a}\|_2\leq 1$. Set $\lambda=4K^2/\kappa$ in Algorithm 2. With probability at least $1-\delta$, it holds for all $t\in[T]$ that*

$$\|\widehat{\boldsymbol{\theta}}_{t,\ell}-\boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}_{t,l}}\leq 16\cdot 2^l\overline{\sigma}\cdot\kappa^{-1/2}\sqrt{d\log\left(\frac{2d\lambda+t\kappa A^2}{2d\lambda}\right)\log\left(4t^2L/\delta\right)}+4R\cdot\kappa^{-1/2}\log(4t^2L/\delta)+2\sqrt{2/\kappa}\cdot K$$

Proof. This lemma can be proved by a direct application of Theorem B.3 and a union bound over L layers. □

Theorem C.2 (Cumulative regret for generalized linear bandits)**.** *Suppose that $\|\boldsymbol{\theta}^*\|_2\leq 1$ and for all $\mathbf{a}\in\bigcup_{t\in[T]}\mathcal{D}_t$, $\|\mathbf{a}\|_2\leq 1$. Set $\lambda=4K^2/\kappa$ and*

$$\beta_{t,l}=16\cdot 2^l\overline{\sigma}\cdot\kappa^{-1/2}\sqrt{d\log\left(\frac{2d\lambda+t\kappa A^2}{2d\lambda}\right)\log\left(4t^2L/\delta\right)}+4R\cdot\kappa^{-1/2}\log(4t^2L/\delta)+2\sqrt{2/\kappa}\cdot K \tag{25}$$

in Algorithm 2. With probability at least $1-\delta$, the regret of Algorithm 2 at the first T rounds satisfies that:

$$\textit{Regret}(T)\leq\widetilde{O}\left(\frac{K}{\kappa}\cdot d\sqrt{\sum_{t=1}^T\sigma_t^2}+(R+K)\,K\cdot\kappa^{-1}\sqrt{dT}\right)$$

Proof. Based on the definition of regret (9),

$$\begin{aligned}
\text{Regret}(T) &= \sum_{t=1}^{T} \phi\left(\left(\boldsymbol{\theta}^*\right)^\top \mathbf{a}_t^*\right) - \phi\left(\left(\boldsymbol{\theta}^*\right)^\top \mathbf{a}_t\right) \\
&\leq \sum_{t=1}^{T} \phi\left(\widehat{\boldsymbol{\theta}}_{t,l_t}^\top \mathbf{a}_t + \beta_{t,l_t} \cdot \|\mathbf{a}_t\|_{\boldsymbol{\Sigma}_{t,l_t}^{-1}}\right) - \phi\left(\left(\boldsymbol{\theta}^*\right)^\top \mathbf{a}_t\right) \\
&\leq \sum_{t=1}^{T} 2K \cdot \beta_{t,l_t} \cdot \|\mathbf{a}_t\|_{\boldsymbol{\Sigma}_{t,l_t}^{-1}} \\
&\leq 2K \sum_{l\in[L]} \beta_{T,l} \sum_{t\in\Psi_{T+1,l}} \|\mathbf{a}_t\|_{\boldsymbol{\Sigma}_{t,l}^{-1}} \\
&\leq 2K \sum_{l\in[L]} \beta_{T,l} \sqrt{|\Psi_{T+1,l}|} \sqrt{\sum_{t\in\Psi_{T+1,l}} \min\left\{1/\kappa, \|\mathbf{a}_t\|_{\boldsymbol{\Sigma}_{t,l}^{-1}}^2\right\}} \\
&\leq 4K \sum_{l\in[L]} \beta_{T,l} \sqrt{|\Psi_{T+1,l}|} \cdot \sqrt{d \cdot \kappa^{-1} \log\left(\frac{2d\lambda + T\kappa A^2}{2d\lambda}\right)}, \qquad (26)
\end{aligned}$$

where the first inequality holds due to Lemma C.1, the second inequality follows from Assumption 3.1, the fourth inequality is obtained by applying Cauchy-Schwarz inequality, the last inequality follows from Lemma D.4.

Substituting (25) into (26), we obtain

$$\begin{aligned}
\text{Regret}(T) &\leq 4K \sqrt{d \cdot \kappa^{-1} \log\left(\frac{2d\lambda + T\kappa A^2}{2d\lambda}\right)} \sum_{l\in[L]} \sqrt{|\Psi_{T+1,l}|} \cdot \widetilde{O}\left(2^l \overline{\sigma} \kappa^{-1/2} \sqrt{d} + R \cdot \kappa^{-1/2} + \sqrt{K^2/\kappa}\right) \\
&\leq 4K\sqrt{L} \sqrt{d \cdot \kappa^{-1} \log\left(\frac{2d\lambda + T\kappa A^2}{2d\lambda}\right)} \sqrt{\sum_{l\in[L]} \sum_{t\in\Psi_{T+1,l}} \widetilde{O}\left(\sigma_t^2 \cdot \kappa^{-1} d + R^2/\kappa + K^2/\kappa\right)} \\
&\leq \widetilde{O}\left(\frac{K}{\kappa} \cdot d \sqrt{\sum_{t=1}^{T} \sigma_t^2} + (R+K)\, K \cdot \kappa^{-1} \sqrt{dT}\right) \qquad (27)
\end{aligned}$$

□

D. Auxiliary Lemmas

Lemma D.1 (Freedman 1975)**.** *Let $M, v > 0$ be fixed constants. Let $\{x_i\}_{i=1}^n$ be a stochastic process, $\{\mathcal{G}_i\}_i$ be a filtration so that for all $i \in [n]$, x_i is $\mathcal{G}_i$-measurable, while most surely $\mathbb{E}[x_i|\mathcal{G}_{i-1}] = 0$, $|x_i| \leq M$ and $\sum_{i=1}^n \mathbb{E}(x_i^2|\mathcal{G}_i) \leq v$. Then, for any $\delta > 0$, with probability $1-\delta$,*

$$\sum_{i=1}^{n} x_i \leq \sqrt{2v \log(1/\delta)} + 2/3 \cdot M \log(1/\delta).$$

Lemma D.2. *Suppose $a, b \geq 0$. If $x^2 \leq a + b \cdot x$, then $x^2 \leq 2b^2 + 2a$.*

Proof. By solving the root of quadratic polynomial $q(x) := x^2 - b \cdot x - a$, we obtain $\max\{x_1, x_2\} = (b + \sqrt{b^2 + 4a})/2$. Hence, we have $x \leq (b + \sqrt{b^2 + 4a})/2$ provided that $q(x) \leq 0$. Then we further have

$$x^2 \leq \frac{1}{4}\left(b + \sqrt{b^2 + 4a}\right)^2 \leq \frac{1}{4} \cdot 2\left(b^2 + b^2 + 4a\right) \leq 2b^2 + 2a. \qquad (28)$$

□

Lemma D.3 (Hoeffding's inequality)**.** *Let $\{x_i\}_{i=1}^n$ be a stochastic process, $\{\mathcal{G}_i\}_i$ be a filtration so that for all $i \in [n]$, x_i is $\mathcal{G}_i$-measurable, while $\mathbb{E}[x_i|\mathcal{G}_{i-1}] = 0$ and $x_i|\mathcal{G}_{i-1}$ is a σ_i-sub-Gaussian random variable. Then, for any $t > 0$, with probability at least $1-\delta$, it holds that*

$$\sum_{i=1}^n x_i \le \sqrt{2\sum_{i=1}^n \sigma_i^2 \log(1/\delta)}.$$

Lemma D.4 (Lemma 11, Abbasi-Yadkori et al. 2011)**.** *For any $\lambda > 0$ and sequence $\{\mathbf{x}_t\}_{t=1}^T \subset \mathbb{R}^d$ for $t \in \{0, 1, \cdots, T\}$, define $\mathbf{Z}_t = \lambda\mathbf{I} + \sum_{i=1}^t \mathbf{x}_i\mathbf{x}_i^\top$. Then, provided that $\|\mathbf{x}_t\|_2 \le M$ for all $t \in [T]$, we have*

$$\sum_{t=1}^T \min\{1, \|\mathbf{x}_t\|^2_{\mathbf{Z}_{t-1}^{-1}}\} \le 2d\log\frac{d\lambda + TM^2}{d\lambda}.$$

Lemma D.5. *For any $\lambda > 0$ and sequence $\{\mathbf{x}_t\}_{t=1}^T \subset \mathbb{R}^d$ for $t \in \{0, 1, \cdots, T\}$, define $\mathbf{Z}_t = \lambda\mathbf{I} + \sum_{i=1}^t \mathbf{x}_i\mathbf{x}_i^\top$. Then, provided that $\|\mathbf{x}_t\|_2 \le M$ for all $t \in [T]$, we have*

$$\sum_{t=1}^T \|\mathbf{x}_t\|^2_{\mathbf{Z}_t^{-1}} \le 2d\log\frac{d\lambda + TM^2}{d\lambda}.$$

Proof. Applying matrix inversion lemma,

$$\begin{aligned}
\sum_{t=1}^T \|\mathbf{x}_t\|^2_{\mathbf{Z}_t^{-1}} &= \sum_{t=1}^T \mathbf{x}_t^\top \left(\mathbf{Z}_{t-1}^{-1} - \frac{\mathbf{Z}_{t-1}^{-1}\mathbf{x}_t\mathbf{x}_t^\top\mathbf{Z}_{t-1}^{-1}}{1+\mathbf{x}_t^\top\mathbf{Z}_{t-1}^{-1}\mathbf{x}_t}\right)\mathbf{x}_t \\
&= \sum_{t=1}^T \frac{\|\mathbf{x}_t\|^2_{\mathbf{Z}_{t-1}^{-1}}}{1+\|\mathbf{x}_t\|^2_{\mathbf{Z}_{t-1}^{-1}}} \\
&\le \sum_{t=1}^T \min\{1, \|\mathbf{x}_t\|^2_{\mathbf{Z}_{t-1}^{-1}}\} \\
&\le 2d\log\frac{d\lambda + TM^2}{d\lambda},
\end{aligned}$$

where the first equality follows from matrix inversion lemma, the second inequality holds by Lemma D.4. □

Lemma D.6 (Concentration bound for sub-exponential random variables)**.** *Let X be a sub-exponential random variable such that $X \sim SE(\sigma^2, \alpha)$. Then we have*

$$\mathbb{P}\left(X - \mathbb{E}[X] \ge \beta\right) \le \begin{cases} \exp\left(-\beta^2/(2\sigma^2)\right), & 0 < \beta \le \sigma^2/\alpha \\ \exp\left(-\beta/2\alpha\right), & t > \sigma^2/\alpha \end{cases}$$

Lemma D.7 (Confidence Ellipsoid, Theorem 2, Abbasi-Yadkori et al. 2011)**.** *Let $\{\mathcal{G}_k\}_{k=1}^\infty$ be a filtration, and $\{\mathbf{x}_k, \eta_k\}_{k\ge1}$ be a stochastic process such that $\mathbf{x}_k \in \mathbb{R}^d$ is $\mathcal{G}_k$-measurable and $\eta_k \in \mathbb{R}$ is $\mathcal{G}_{k+1}$-measurable. Let $L, \sigma, \lambda, \epsilon > 0$, $\boldsymbol{\mu}^* \in \mathbb{R}^d$. For $k \ge 1$, let $y_k = \langle\boldsymbol{\mu}^*, \mathbf{x}_k\rangle + \eta_k$ and suppose that $\eta_k, \mathbf{x}_k$ also satisfy*

$$\mathbb{E}[\eta_k|\mathcal{G}_k] = 0,\ \eta_k|\mathcal{G}_k \sim \mathrm{subG}(R),\ \|\mathbf{x}_k\|_2 \le L. \tag{29}$$

For $k \ge 1$, let $\mathbf{Z}_k = \lambda\mathbf{I} + \sum_{i=1}^k \mathbf{x}_i\mathbf{x}_i^\top$, $\mathbf{b}_k = \sum_{i=1}^k y_i\mathbf{x}_i$, $\boldsymbol{\mu}_k = \mathbf{Z}_k^{-1}\mathbf{b}_k$, and

$$\beta_k = R\sqrt{d\log\left(\frac{1+kL^2/\lambda}{\delta}\right)}.$$

Then, for any $0 < \delta < 1$, we have with probability at least $1-\delta$ that,

$$\forall k \ge 1,\ \left\|\sum_{i=1}^k \mathbf{x}_i\eta_i\right\|_{\mathbf{Z}_k^{-1}} \le \beta_k,\ \|\boldsymbol{\mu}_k - \boldsymbol{\mu}^*\|_{\mathbf{Z}_k} \le \beta_k + \sqrt{\lambda}\|\boldsymbol{\mu}^*\|_2.$$

Lemma D.8 (Pinsker & Feinstein 1964)**.** *If P and Q are two probability distributions on a measurable space $(X, \mathbf{\Sigma})$, then for any measurable event $A \in \mathbf{\Sigma}$, it holds that*

$$|P(A) - Q(A)| \leq \sqrt{\frac{1}{2} KL(P\|Q)} := \sqrt{\frac{1}{2} \mathbb{E}_P \left(\log \frac{dP}{dQ} \right)}.$$