Online Quantile Regression

Yinan Shen   YINANSHE@USC.EDU
Department of Mathematics
University of Southern California
Los Angeles, CA 90089, USA

Dong Xia   MADXIA@UST.HK
Department of Mathematics
Hong Kong University of Science and Technology
Hong Kong SAR, China

Wen-Xin Zhou   WENXINZ@UIC.EDU
Department of Information and Decision Sciences
University of Illinois Chicago
Chicago, IL 60607, USA

Editor: Genevera Allen

Abstract

This paper addresses the challenge of integrating sequentially arriving data into the quantile regression framework, where the number of features may increase with the number of observations, the time horizon is unknown, and memory resources are limited. Unlike least squares and robust regression methods, quantile regression models different segments of the conditional distribution, thereby capturing heterogeneous relationships between predictors and responses and providing a more comprehensive view of the underlying stochastic structure. We employ stochastic sub-gradient descent to minimize the empirical check loss and analyze its statistical properties and regret behavior. Our analysis reveals a subtle interplay between updating iterates based on individual observations and on batches of observations, highlighting distinct regularity characteristics in each setting. The proposed method guarantees long-term optimal estimation performance regardless of the chosen update strategy. Our contributions extend existing literature by establishing exponential-type concentration inequalities and by achieving optimal regret and error rates that exhibit only *short-term* sensitivity to initialization. A key insight from our study lies in the refined statistical analysis showing that properly chosen stepsize schemes substantially mitigate the influence of initial errors on subsequent estimation and regret. This result underscores the robustness of stochastic sub-gradient descent in managing initial uncertainties and affirms its effectiveness in sequential learning settings with unknown horizons and data-dependent sample sizes. Furthermore, when the initial estimation error is well-controlled, our analysis reveals a trade-off between short-term error reduction and long-term optimality. For completeness, we also discuss the squared loss case and outline appropriate update schemes, whose analysis requires additional care. Extensive simulation studies corroborate our theoretical findings.

Keywords: online learning, quantile regression, regret, stochastic sub-gradient descent.

1. Introduction

Online learning aims to efficiently incorporate sequentially arriving data and make timely predictions. In contrast to offline learning, where all data are collected and stored prior to analysis, online

learning processes data in a sequential manner without requiring access to the entire sample at once, thereby alleviating both storage and computational burdens. As a result, online methods are particularly well-suited for large-scale datasets (LeCun et al., 1989; Rajalakshmi et al., 2019; Finn et al., 2019), streaming asset pricing data (Soleymani and Paquet, 2020), and the increasingly prominent domain of reinforcement learning (Gao et al., 2019; Han et al., 2022; Ren and Zhou, 2023). Broader treatments and further applications of online learning can be found in Bottou (1999), Cesa-Bianchi and Lugosi (2006), Hoffman et al. (2010), Hazan (2016), and Orabona (2019).

In classical offline linear regression, estimation and inference are based on a pre-collected sample of independent observations $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ satisfying $Y_i = \mathbf{X}_i^\top \boldsymbol{\beta}^* + \xi_i$, where $\boldsymbol{\beta}^* \in \mathbb{R}^d$ is the unknown parameter of interest, and ξ_i denotes the random noise variable, satisfying $\mathbb{E}(\xi_i | \mathbf{X}_i) = 0$. It is well-known that the ordinary least squares estimator $\hat{\boldsymbol{\beta}}$ achieves the minimax rate of convergence, i.e., $\mathbb{E}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2 = O(\sigma^2 d/n)$, where σ^2 is such that $\mathbb{E}(\xi_i^2 | \mathbf{X}_i) \leq \sigma^2$ almost surely. In contrast, the entire sample is inaccessible in the online setting, and the total number of observations may even be unknown. Specifically, at time t, only n_t pairs of observations can be used:

$$Y_i^{(t)} = \mathbf{X}_i^{(t)\top} \boldsymbol{\beta}^* + \xi_i^{(t)}, \quad i = 1, \ldots, n_t. \tag{1}$$

While the data are received sequentially, online learning refrains from repeatedly using the data for model refitting. Instead, it sequentially updates the estimator. In contrast, offline methods leverage the entire sample, and based on the available data, the offline method (He and Shao, 2000; Koenker, 2005) leads to a statistically optimal estimator with mean squared error of order $O(\sigma^2 \cdot d / \sum_{t=0}^T n_t)$, where T is called the *horizon*. This prompts the question of whether online methods can achieve error rates comparable to their offline counterparts. Must we compromise accuracy for more efficient computation and reduced storage in online learning? Furthermore, with the accumulation of more information over time, the theoretical offline error rates decline. Is it feasible to sustain this downward trend in optimal error rates over time? Particularly when the horizon (maximum time T) is unknown, isolated error rates at specific times lack persuasiveness.

In online learning, a learner must generate timely predictions as data arrive sequentially, either individually or in batches. The concept of *regret*, first introduced by Savage (1951) and later popularized in the online learning literature (e.g., Goldenshluger and Zeevi, 2013; Han et al., 2020; Ren and Zhou, 2023), quantifies cumulative prediction error over time and serves as a key measure of learning performance. The regret at horizon T is defined as

$$\mathsf{Regret}_T := \mathbb{E}\left\{ \sum_{t=0}^T f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*) \right\}, \tag{2}$$

where $f_t(\cdot)$ and $\boldsymbol{\beta}_t$ denote the loss function and estimator at time t, respectively. The expectation is taken over the data from $t = 0$ to T. Regret serves as a metric to evaluate the efficacy of a sequential estimation scheme. Optimal regret occurs when error rates exhibit a scaling behavior of $1/t$ as time t becomes sufficiently large (Hazan, 2016). In contrast, Cai et al. (2023) assumes a known horizon and focuses more on the ultimate accuracy of estimation. However, this setting bears resemblance to the sample splitting technique (Xia and Yuan, 2021; Zou et al., 2020).

Online estimation under the squared loss in equation (2) has been extensively studied over the past two decades; see, for example, Bottou (1999), Zinkevich (2003), Hazan et al. (2007b), and Langford et al. (2009), among others. Stochastic gradient descent stands out as a natural and elegant methodology for processing sequentially arriving data, integrating streaming information,

and minimizing the regret in (2). Bottou (1999) lays the foundation for the general theoretical framework of online learning, establishing the asymptotic convergence of gradient descent toward stationary points. Under suitable conditions, the seminal work of Zinkevich (2003) proves the iterative convergence of gradient descent with a stepsize of order $O(t^{-1/2})$. Subsequently, Hazan et al. (2007a) demonstrates that using a stepsize of order $O(t^{-1})$ yields an optimal regret bound of order $O(\log T)$. The pioneering work of Duchi and Singer (2009) further examined the performance of iterative sub-gradient descent applied to non-differentiable loss functions, offering insights into convergence dynamics and establishing regret bounds under various stepsize schemes. We refer to Hazan et al. (2007b), Langford et al. (2009), Streeter and McMahan (2010), McMahan and Streeter (2010), Duchi et al. (2011), Bach and Moulines (2013), and the references therein Orabona (2019) for additional gradient descent-based algorithms in online learning. In the offline learning context, stochastic (sub-)gradient descent serves as an effective approach to mitigate computational burdens (Zhang, 2004). This method updates the current iterate using only a single, randomly selected observation pair, rather than the entire sample. Rakhlin et al. (2011) introduces a stepsize scheme of order $O(1/t)$, yielding a sub-linear convergence rate comparable to that of its online counterpart under a strongly convex objective. In offline stochastic optimization, the total sample size is typically known, allowing such information to be incorporated into the stepsize design and ensuring that the statistical optimality remains stable over time. Examples of constant stepsize schemes that explicitly account for the learning horizon can be found in Duchi and Singer (2009), Cai et al. (2023), and Puchkin et al. (2024). Other studies, such as Delyon and Juditsky (1993), Roux et al. (2012), and Johnson and Zhang (2013), focus on accelerating the sub-linear convergence rate of stochastic descent methods. While both offline stochastic gradient descent and online learning alleviate computational pressures, they differ fundamentally in nature. In online learning, the time horizon and the total number of observations are unknown, and may even be infinite, necessitating procedures that achieve dynamic and statistically optimal error rates over time. Moreover, an online learner often must generate predictions for newly arriving covariate-only inputs, with predictive accuracy evaluated through the notion of regret.

The statistical analysis of the squared loss remains a delicate yet crucial area of study. Although online learning and (stochastic) gradient descent methods have been extensively investigated from an optimization perspective for both squared and non-differentiable loss functions, most of the existing literature assumes that the empirical loss possesses certain regularity properties, such as strong convexity and/or smoothness. Moreover, the feasible domain is often restricted to a bounded convex set (Hazan et al., 2007a; Duchi and Singer, 2009; Langford et al., 2009; Rakhlin et al., 2011). Comprehensive overviews of online learning algorithms from an optimization standpoint can be found in Hazan (2016) and Orabona (2019). However, empirical strong convexity may fail to hold even under the squared loss when the available storage size is limited. Addressing this issue calls for refined statistical analysis that explicitly accounts for the interplay between data dependence, sample size, and memory constraints; see, for example, Chen et al. (2019), Han et al. (2024), and the references therein. Fan et al. (2018) further demonstrates that an expected error rate of order $O(\log(T)/T)$ is attainable for sparse online regression under a stepsize scheme of order $O(1/t)$, yielding an optimal $O(\log T)$ regret.

The preceding statistical online frameworks exhibit notable limitations when applied to quantile regression, often yielding suboptimal error rates and weakened guarantees on success probability, particularly in the presence of heavy-tailed noise. Existing studies on renewable or online quantile regression primarily analyze asymptotic performance within batch learning frameworks, where the

batch size diverges over time. A common feature of these methods is their reliance on storing historical summary statistics, which incurs a memory cost of order $O(d^2)$ (Jiang and Yu, 2022; Sun et al., 2023; Chen and Yuan, 2024). In contrast, the present work complements prior studies by developing online quantile regression methods that eliminate the need to store either historical data or summary statistics. For comparison, Section 3.1 provides a parallel discussion of online least squares methods and their corresponding stepsize schemes. To estimate conditional quantile models sequentially, we adopt a simple yet effective sub-gradient descent approach. Through refined statistical analysis, we design a stepsize sequence that either exhibits geometric decay, remains constant, or decays at the rate of $1/t$, depending on the proximity of the current iterate to the oracle solution. This design departs from the scheme proposed by Duchi and Singer (2009); a detailed comparison between the two approaches is provided in Section 3.1, highlighting the advantages of the proposed statistically informed stepsize rule. Under heavy-tailed noise, we further establish exponential-type tail bounds for the proposed online estimators, improving upon the polynomial-type guarantees in Han et al. (2022). This enhancement in tail behavior underscores the robustness and reliability of our estimators when applied to heavy-tailed data environments.

In this work, we refer to online learning as a setting in which data arrive sequentially over time. The complete dataset and the total number of observations are never simultaneously available, regardless of how many observations can be processed at each step. Under the linear model (1), we consider three representative scenarios: (i) the newly arriving dataset $\mathcal{D}_t := \{(\mathbf{X}_i^{(t)}, Y_i^{(t)})\}_{i=1}^{n_t}$ contains only a single observation pair, requiring $O(d)$ storage; (ii) $\mathcal{D}_t$ contains at least $O(d)$ samples; and (iii) the server can store an unbounded number of samples. The first setting has been extensively examined in the literature (Duchi and Singer, 2009; Hazan et al., 2007a; Hazan, 2016; Cai et al., 2023; Langford et al., 2009; Bach and Moulines, 2013; Mhammedi et al., 2019; Han et al., 2022), yet it continues to demand careful statistical treatment, as discussed earlier. Do et al. (2009) further studies the case where data arrive in batches at each iteration. From an optimization standpoint, however, there is little distinction between single-sample and batch arrivals, since regularity conditions such as smoothness or strong convexity are typically assumed. From a statistical perspective, by contrast, we will demonstrate that these two settings exhibit markedly different behaviors. Specifically, this paper seeks to address the following fundamental questions: Is it possible to achieve a continuously decreasing error rate as new data arrive under limited storage? Can statistical optimality be maintained along the learning trajectory? How much information is lost relative to the offline benchmark? In online learning, how do failure probabilities accumulate across iterations, and how do they determine the maximal horizon for guaranteed convergence? What are the implications for per-iteration success probability, especially under heavy-tailed noise and single-sample updates? Finally, what role does initialization play? Does the initial error necessarily have a long-term influence on subsequent errors and regret, or can arbitrary initialization be tolerated?

We summarize our main contributions addressing these questions as follows.

1. This paper develops a statistical analysis framework for online quantile regression, addressing the challenges arising from the empirical loss function's lack of strong convexity and smoothness. To achieve long-term statistical optimality under heavy-tailed noise, we design a stepsize scheme informed by the statistical regularity properties of the quantile loss. Unlike prior studies, the derived error rates explicitly scale with the noise level, capturing the intrinsic stochastic complexity of the problem. Furthermore, the proposed online approach deliberately trades a marginal loss in statistical efficiency for substantial gains in compu-

tational scalability and memory efficiency relative to offline methods. As demonstrated in Section 3.1, this slight reduction in accuracy becomes asymptotically negligible as the time horizon grows.

2. Our algorithm attains statistically optimal rates even under heavy-tailed noise and admits a high-probability guarantee, with the failure probability decaying exponentially fast in the dimension. This property ensures the validity of our theoretical results even when the unknown horizon grows exponentially with dimension. For instance, when only a single datum arrives at each iteration, the failure probability associated with our established error rate is $O(\exp(-c_0 d))$, where $c_0 > 0$ is a universal constant. Consequently, the horizon can be as large as $T = O(\exp(c_0 d))$ without compromising the theoretical guarantees. In sharp contrast, the results in Han et al. (2022) and Cai et al. (2023) hold with probability $1 - d^{-O(1)}$ under sub-Gaussian noise, implying that their admissible horizon is limited to $T = O(d^{O(1)})$.

3. Our analysis demonstrates that the influence of the initial error on both estimation accuracy and regret is merely transient. Consequently, the derived statistical error bounds remain valid for arbitrary initializations, without requiring the initial estimator to lie within a compact region. This robustness is achieved through refined statistical analyses and the design of a novel stepsize scheme. Stochastic (sub-)gradient descent methods for both quantile and squared losses benefit substantially from stepsize schedules informed by underlying statistical properties. For instance, in the setting where $\xi_i^{(t)} \sim N(0, \sigma^2)$ and $n_t = 1$ in (1), we establish the regret bound $\sum_{t=0}^{T} \mathbb{E}\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|^2 \leq O(\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2^2 + \sigma^2 \log T)$.

The remainder of this paper is organized as follows. Section 2 introduces the quantile (check) loss function and the online sub-gradient descent algorithm. It further examines the convergence dynamics and regret bounds, with detailed discussions of the three distinct settings considered. Section 3.1 presents extensive numerical experiments, including evaluations of the proposed stepsize scheme, comparisons of accuracy with offline regression methods, and analyses of convergence behavior. A brief discussion of online least squares from a statistical perspective is also provided. The proof of Theorem 1 appears in Section 4, while the proofs of the remaining theorems are deferred to the supplementary material.

2. Online quantile regression via adaptive sub-gradient descent

Throughout this section, we consider the conditional quantile model (1), that is, the conditional τ-th quantile of $\xi_i^{(t)}$ given $\mathbf{X}_i^{(t)}$ is zero for some predetermined $\tau \in (0, 1)$. Quantile regression (QR) plays a crucial role in unraveling pathways of dependence between the outcome and a collection of features, which remain elusive through conditional mean regression analysis, such as the least squares estimation. Since its introduction by Koenker and Bassett Jr (1978), QR as undergone extensive study, encompassing theoretical understanding, methodological development for various models and data types, and practical applications in a wide range of fields. The main focus of existing literature centers around the formulation of methodologies and theories for QR utilizing static data, where a complete dataset is available. This is referred to as the offline setting in online learning literature. The most fundamental and well-studied method for estimating $\boldsymbol{\beta}^*$ involves empirical risk minimization, or statistical M-estimation. The associated loss function $\rho_{Q,\tau}(x) := \tau x \mathbb{I}(x \geq 0) + (\tau - 1)x \mathbb{I}(x < 0)$ is known as the check loss or quantile loss. We refer to

Koenker (2005) and Koenker et al. for a comprehensive exploration of offline quantile regression, covering statistical theory, computational methods, as well as diverse extensions and applications.

In contrast to the least squares method, the non-smooth nature of the quantile loss introduces additional challenges to QR, particularly in the era of big data. Established algorithms employed for this purpose include the simplex algorithm (Koenker and d'Orey, 1987), interior point methods with preprocessing (Portnoy and Koenker, 1997), alternating direction method of multipliers, and proximal methods (Parikh and Boyd, 2014). More recently, Fernandes et al. (2021) and He et al. (2023) have shown that convolution-smoothed quantile regression attains desired asymptotic and non-asymptotic properties, provided that the smoothing parameter (bandwidth) is appropriately selected as a function of the sample size and dimensionality. When addressing low-rank matrix regression or completion problems under either absolute or quantile loss, recent studies, including Cambier and Absil (2016), Li et al. (2020), Charisopoulos et al. (2021), Tong et al. (2021), Ma and Fattahi (2023), and Shen et al. (2023), have conducted extensive investigations of sub-gradient descent methods from both computational and statistical perspectives. The offline methods mentioned above crucially hinge on specific regularity properties inherent in the empirical loss function. However, in the online setting, characterized by sequential data arrival and an unknown total number of observations, the assurance of offline regularity properties becomes untenable with only a limited number of available samples. Consequently, neither the offline results nor the corresponding proof techniques are applicable in this online context. As an illustration, in an offline setting based on model (1), Shen et al. (2023) demonstrated the existence of certain parameters $\mu_1, \mu_2 > 0$ such that with high probability,

$$\frac{1}{n} \sum_{i=1}^{n} (|Y_i - \langle \mathbf{X}_i, \boldsymbol{\beta} \rangle| - |Y_i - \langle \mathbf{X}_i, \boldsymbol{\beta}^* \rangle|) \geq \max\{\mu_1 \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 - \gamma, \mu_2 \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2\}$$

holds for all $\boldsymbol{\beta}$ under the sample size requirement $n \geq Cd$, $\gamma = \mathbb{E}|\xi|$ is the first absolute moment of the noise variable. However, in the context of online learning, where the stored data is limited, such as when $n = 1$, the above inequality does not hold in general.

The foundational contributions to online learning by Duchi and Singer (2009), Duchi et al. (2011) and Johnson and Zhang (2013) employ sub-gradient descent to address scenarios characterized by the lack of differentiability in the loss function at specific points. These studies establish the properties of iterates under various stepsize schemes, with a primary emphasis on the excess risk, leaving the analysis of estimation error $\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2$ unaddressed. In the absence of strong convexity and/or certain smoothness, the upper bounds on the excess risk cannot be straightforwardly extended to those on the estimation error. From the statistical viewpoint, Jiang and Yu (2022) and Wang et al. (2022) consider online quantile regression within the batch learning framework in which the data arrive in batches, and establish their asymptotic normality as the batch size goes to infinity. It is worth noticing that the requirement for diverging batch sizes renders these methods impractical in scenarios where only one observation or a few observations arrive at a time.

This paper is dedicated to the exploration of online quantile regression with unknown horizon and sample sizes (batch or total). We aim to provide a non-asymptotic analysis of online sub-gradient descent. By using a customized stepsize scheme with explicit dependencies on dimensionality, sample size, and noise scale, we establish optimal rates of convergence for online QR estimators. Additionally, we seek to demonstrate that near-optimal regret performance can be achieved. Let $\mathcal{S}_t$ be the set of observations acquired at time t, which will be used to update the

current iterate $\boldsymbol{\beta}_t$. Consequently, the loss function at t is given by

$$f_t(\boldsymbol{\beta}) := \sum_{(\mathbf{X}_i^{(t)}, Y_i^{(t)}) \in \mathcal{S}_t} \rho_{Q,\tau}(Y_i^{(t)} - \langle \mathbf{X}_i^{(t)}, \boldsymbol{\beta} \rangle).$$

We will explore three settings: (i) $\mathcal{S}_t = \mathcal{D}_t$, containing only one data vector, referred to as the *online learning* setting, (ii) $\mathcal{S}_t = \mathcal{D}_t$, comprising a set of observations with a size of at least $O(d)$, termed as the *online batch learning* setting, and (iii) $\mathcal{S}_t$ containing all the data vectors accumulated up to time t, expressed as $\mathcal{S}_t = \cup_{l=0}^t \mathcal{D}_l$, recognized as the *sequential learning with infinite storage* setting.

Online sub-gradient descent. Initiated at some $\boldsymbol{\beta}_0$, the online QR estimates are iteratively defined via sub-gradient descent as

$$\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t - \eta_t \mathbf{g}_t, \quad t = 0, 1, \ldots,$$

where $\mathbf{g}_t \in \partial f_t(\boldsymbol{\beta}_t)$ is the sub-gradient of f_t at $\boldsymbol{\beta}_t$, and $\{\eta_t\}_{t=0,1,\ldots}$ constitutes a sequence of stepsize parameters (learning rates). The loss function varies across different settings, exhibiting distinct regularity properties. Consequently, tailored stepsize schemes are imperative in the three settings to attain the desired convergence properties. We will show that these customized stepsize schemes yield statistically optimal estimators and achieve near-optimal regret performances. Importantly, these schemes effectively adapt to varying dimensions and unknown horizons.

Prior to presenting the theoretical guarantees and convergence dynamics, we outline the essential assumptions concerning the model.

Assumption 1 (*Random covariate*) *The covariate vector $\mathbf{X} \in \mathbb{R}^d$ follows the Gaussian distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is symmetric and positive definite. There exist absolute constants $C_l, C_u > 0$ such that $C_l \leq \lambda_{\min}(\boldsymbol{\Sigma}) \leq \lambda_{\max}(\boldsymbol{\Sigma}) \leq C_u$.*

Assumption 2 (*Noise distribution*) *Given $\mathbf{X} \in \mathbb{R}^d$, the noise variable ξ has the conditional density function $h_{\xi|\mathbf{X}}(\cdot)$ and distribution function $H_{\xi|\mathbf{X}}(\cdot)$, respectively. The conditional τ-th quantile of ξ given $\mathbf{X}$ is zero, i.e., $H_{\xi|\mathbf{X}}(0) = \tau$ and $\mathbb{E}|\xi| < +\infty$. Let $\gamma = \mathbb{E}|\xi|$. There exist constants $b_0, b_1 > 0$ such that* [1]

(a) $\inf_{|x| \leq 8(C_u/C_l)^{1/2}\gamma} h_{\xi|\mathbf{X}}(x) \geq b_0^{-1}$,

(b) $\sup_{x \in \mathbb{R}} h_{\xi|\mathbf{X}}(x) \leq b_1^{-1}$.

Both local lower and global upper bounds on the (conditional) density $h_{\xi|\mathbf{X}}$ are commonly imposed in the QR literature, particularly in non-asymptotic settings (e.g., Belloni and Chernozhukov (2011) and He et al. (2023)).

---

1. As marked in Shen et al. (2023), the noise assumptions can be slightly relaxed to $|H_{\xi|\mathbf{X}}(x) - H_{\xi|\mathbf{X}}(0)| \geq |x|/b_0$ for $|x| \leq 8(C_u/C_l)^{1/2}\gamma$ and $|H_{\xi|\mathbf{X}}(x) - H_{\xi|\mathbf{X}}(0)| \leq |x|/b_1$.

2.1 Online Learning

The server receives a new data point $(\mathbf{X}_t, Y_t)$ that follows the linear model (1). The online sub-gradient descent algorithm calculates the corresponding sub-gradient $\mathbf{g}_t \in \partial f_t(\boldsymbol{\beta}_t)$ and subsequently updates the current estimate $\boldsymbol{\beta}_t$ in the direction of the negative sub-gradient, all *without* storing the new observation. Here, the loss function at time t is given by $f_t(\boldsymbol{\beta}) = \rho_{Q,\tau}(Y_t - \mathbf{X}_t^\top \boldsymbol{\beta})$. More specifically, the algorithm proceeds as follows. We begin with an arbitrary initialization $\boldsymbol{\beta}_0$. At each time step t, given the current estimate $\boldsymbol{\beta}_t$ and a newly arrived observation $(\mathbf{X}_t, Y_t)$, the update is performed as

$$\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t - \eta_t \cdot (\tau - \mathbb{I}\{Y_t - \langle \mathbf{X}_t, \boldsymbol{\beta}_t \rangle\}) \cdot \mathbf{X}_t,$$

where $\eta_t > 0$ denotes the stepsize. The choice of η_t will be discussed in Theorem 1.

The expected excess risk $\mathbb{E}[f_t(\boldsymbol{\beta}) - f_t(\boldsymbol{\beta}^*)]$, as a function of $\boldsymbol{\beta}$, demonstrates a two-phase regularity property, visually depicted in Figure 1. As an illustrative example, under the assumptions $\mathbf{X} \sim N(\mathbf{0}, \mathbf{I}_d)$ and $\xi \sim N(0, \sigma^2)$, Shen et al. (2023) derived the population excess risk as

$$\mathbb{E}\{|\mathbf{X}^\top(\boldsymbol{\beta} - \boldsymbol{\beta}^*) + \xi| - |\xi|\} = \sqrt{\frac{\pi}{2}} \frac{\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|^2}{\sqrt{\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|^2 + \sigma^2} + \sigma}.$$

When $\boldsymbol{\beta}$ is distant from the population risk minimizer $\boldsymbol{\beta}^*$, the expected excess risk exhibits a first-order lower bound, decaying linearly in $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2$. However, as $\boldsymbol{\beta}$ approaches proximity to $\boldsymbol{\beta}^*$, a quadratic lower bound emerges concerning $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2$. For general noise with bounded density, a similar equation is achievable and a two-phase property exists accordingly. Importantly, it is noteworthy that, in contrast to offline works such as (Tong et al., 2021; Shen et al., 2023), this regularity property pertains to the expectation, given that the empirical loss is based on only one single observation. The empirical loss lacks a guaranteed high probability concentration property; for instance, the variance of $f_t(\boldsymbol{\beta})$ may significantly overshadow its expectation. As a result, the commonly observed monotone-decreasing trend in the estimation error, such as $\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2 \leq \|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2$ in offline learning algorithms, does not occur. In fact, a more nuanced analysis is essential to understand the convergence dynamics of online sub-gradient algorithms for quantile regression.

Interestingly, the following theorem illustrates that, despite the online sub-gradient descent algorithm's inability to ensure error contraction as $\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2 < \|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2$, the error rates $\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2$ can be bounded by a monotone-decreasing sequence with high probability. The convergence of the online sub-gradient descent algorithm unfolds in two distinct phases, as suggested by the two-phase regularity properties depicted in Figure 1.

Theorem 1 *Under Assumptions 1 and 2 (a), there exist universal positive constants c_0, c_1, c_2, c_3, C_0 such that, for an arbitrary initialization $\boldsymbol{\beta}_0 \in \mathbb{R}^d$, the sequence $\{\boldsymbol{\beta}_t\}_{t \geq 1}$ generated by the online sub-gradient descent algorithm follows the dynamics outlined below:*

1. in phase one[2], i.e., when $\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2 \geq 8C_l^{-1/2}\gamma$, by selecting a stepsize $\eta_0 = \frac{C\sqrt{C_l}}{C_u \bar{\tau}^2} \frac{D_0}{d}$ and $\eta_t = (1 - c_5 \frac{C_l}{C_u} \frac{1}{d\bar{\tau}^2})\eta_{t-1}$ with any $D_0 \geq \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|$, $C > 1$, $c_5 < 0.1$, it holds with

---

2. With this geometrically decaying stepsize, once $\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2 \asymp O(\gamma)$ is achieved, the estimation error will remain to be this scale both theoretically and practically.

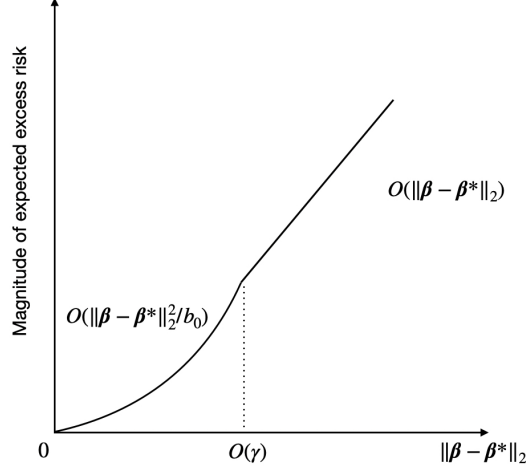Figure 1: *Expected excess risk of the objective function. Y-axis: lower bound of $\mathbb{E}[f_t(\boldsymbol{\beta}) - f_t(\boldsymbol{\beta}^*)]$; X-axis: the value of $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2$. It shows that the lower bound varies from a linear to quadratic dependence on $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2$ as $\boldsymbol{\beta}$ gets closer to $\boldsymbol{\beta}^*$.*

probability at least $1 - \exp(-c_0 d)$ that

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2 \leq C_0 \left(1 - c_5 \frac{C_l}{C_u} \frac{1}{\bar{\tau}^2 d}\right)^{t+1} \cdot \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2^2; \, [3]$$

the conclusion of phase one occurs after $t_1 = O(d \log(\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2/\gamma))$ iterations, where $\bar{\tau} = \max\{\tau, 1 - \tau\}$.

2. *in phase two, i.e., when $\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2 < 8 C_l^{-1/2} \gamma$, by choosing the stepsize $\eta_t = (b_0/C_l) \cdot (C_a/(t - t_1 + C_b d))$ with certain constants $C_a > 12$ and $C_b > C_a/(12d)$, it holds with probability exceeding $1 - \exp(-c_0 d)$ that*

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2 \leq C^* \frac{C_u}{C_l^2} \frac{d}{t + 1 - t_1 + C_b d} b_0^2,$$

where the constant C^* depends on C_a, C_b.

In accordance with Theorem 1, the algorithm exhibits linear convergence during the first phase, concluding after t_1 iterations. This phase requires a total of t_1 observations and employs an approximately geometrically decaying stepsize scheme. In the second phase, the algorithm converges sub-linearly, employing a $O(1/(t + d))$ decaying stepsize scheme. The failure probability in both phases decays exponentially in d. To our knowledge, our result stands as the first instance in online learning literature to provide an exponential-type high probability bound. Previous works employing the squared loss generally guarantee a polynomial-type tail probability, even under sub-Gaussian

---

3. We remark that C_0 depends on the choices of C, c_5 and we leave the details in the proof.

noise (Cai et al., 2023; Han et al., 2022; Bastani and Bayati, 2020). This enhancement holds significant implications for practical applications. It suggests that our algorithm remains applicable even when the ultimate horizon T is as large as $e^{O(d)}$. Theorem 1 asserts that, provided $d \ll T \ll e^{c_0 d}$, the final estimator β_T, with high probability, attains an error rate that is minimax optimal under offline settings. Online methods can adapt to the unknown horizon, and offer substantial reductions in computation costs, as well as savings in memory and storage.

The final error rate achieved after the second phase iterations is proportional to the noise level, independent of the initialization error. In existing works, the initialization error often exerts a lasting impact on the final error rate, even when using the squared loss. For instance, the online algorithm Fan et al. (2018) obtains an expected error rate $\mathbb{E}\|\beta_T - \beta^*\|_2^2 \le d\|\beta_0 - \beta^*\|_2^2 \log(T)/T$. Moreover, the error rate established by Cai et al. (2023) is $\|\beta_T - \beta^*\|_2^2 \le d\|\beta_0 - \beta^*\|_2^2/T$, which holds with high probability as long as $T \ll \mathrm{poly}(d)$. Theorem 1 reveals that the initialization error only has a short-time effect in the online sub-gradient descent algorithm, dissipating after the second phase iterations begin. This two-phase convergence phenomenon is elucidated through a meticulous analysis of the algorithm's dynamics. It is worth noting that the two-phase convergence and the transient impact of initialization error are not exclusive to quantile regression; they also manifest in least squares, as per our analysis framework. A comprehensive discussion of these properties for online least squares will be presented in Section 3.1. Numerical simulations in that section demonstrate that a two-phase stepsize scheme yields significantly improved accuracy.

Corollary 2 *Assume that the same conditions as in Theorem 1 hold, and let the horizon satisfy $T \ge C^* d \max\{1, \log(\sqrt{C_l}\|\beta_0 - \beta^*\|_2/\gamma)\}$ with a constant C^* depending on (C_l, C_u, C_a, C_b). Then, with probability at least $1 - T\exp(-c_0 d)$, the online sub-gradient descent algorithm produces a final estimate with an error rate*

$$\|\beta_T - \beta^*\|_2^2 \le C_3 \frac{d}{T} b_0^2,$$

where $C_3 > 0$ is a constant.

Corollary 2 demonstrates that the online sub-gradient descent algorithm attains the minimax optimal error rate, provided the horizon is not too small. This makes it a favorable choice over the offline approach, considering the benefits of reduced computation and storage costs. However, it is advisable to opt for offline approaches in cases where the horizon is small. A comprehensive examination of their numerical performances is conducted in Section 3.1.

Remark 3 (*Trade-off between short-term accuracy and long-term optimality*) *Suppose the initialization is already situated in the second phase region, i.e., $\|\beta_0 - \beta^*\|_2 < 8C_l^{-1/2}\gamma$. It is advisable to initiate the phase two iterations directly ($t_1 = 0$), where the choice of the parameter C_a plays a pivotal role in determining short-term accuracy and long-term optimality. Selecting $C_a > 12$, according to Theorem 1, the updated estimates achieve an error rate of $\|\beta_t - \beta^*\|_2^2 \le C^* \frac{C_u}{C_l^2} \frac{d}{t + C_b d} b_0^2$ with high probability. This upper bound may exceed the initialization error when t is small, indicating potential accuracy fluctuations in the early stage. Nevertheless, the algorithm eventually converges and yields a statistically optimal estimator. On the contrary, opting for a sufficiently small parameter $C_a < \min\{12, (C_l/\sqrt{C_u})\|\beta_0 - \beta^*\|_2/b_0\}$ results in a more stable convergence*
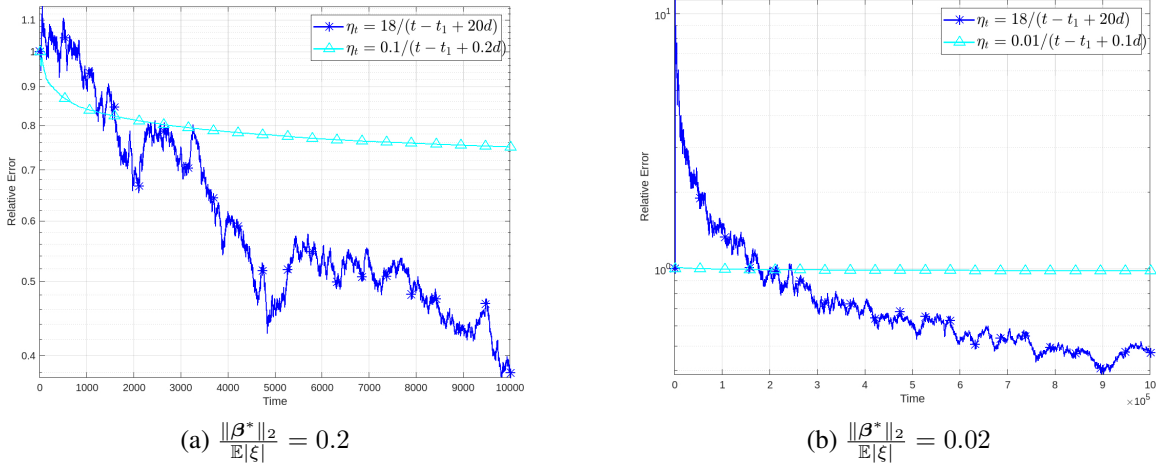
(a) $\frac{\|\boldsymbol{\beta}^*\|_2}{\mathbb{E}|\xi|} = 0.2$ (b) $\frac{\|\boldsymbol{\beta}^*\|_2}{\mathbb{E}|\xi|} = 0.02$

Figure 2: *Trade-off between short-term accuracy and long-term optimality.* The initialization $\boldsymbol{\beta}_0 = \mathbf{0}$ is already in proximity to $\boldsymbol{\beta}^*$, enabling the online sub-gradient descent algorithm to bypass the first phase and start the second phase iterations immediately. A small stepsize can ensure short-term accuracy but may compromise losing long-term optimality, whereas a large stepsize can guarantee long-term optimality but at the expense of short-term accuracy. Here, the dimension is set to $d = 100$, and the noise follows a Student t_ν distribution with $\nu = 1.1$.

in the early stage. However, it sacrifices long-term accuracy, as the updated estimates possess an error bound

$$\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2 \leq \bar{C}^* \left(\frac{d}{t + C_b d} \right)^{\frac{C_a}{12}} \cdot \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2^2,$$

achieving a sub-optimal error rate in the end. This phenomenon is illustrated in Figure 2. Specifically, in Figure 2a, when the initialization satisfies $\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2 \leq 0.2\mathbb{E}|\xi|$, it demonstrates that a small C_a ensures short-term accuracy but compromises long-term optimality. On the other hand, a large C_a guarantees long-term optimality at the expense of short-term accuracy. Figure 2b presents even more pronounced differences, where the initialization satisfies $\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2 \leq 0.02\mathbb{E}|\xi|$.

Theorem 4 *Under the same conditions and stepsize scheme as in Theorem 1, the online sub-gradient descent algorithm achieves an upper bound on regret given by*

$$\textit{Regret}_T \leq C_1(C_u/C_l)\sqrt{C_u}\,d\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2 + C_2 C_a \frac{C_u^2}{C_l^2} \frac{b_0^2}{b_1} d \log\left(1 + \frac{T}{C_b d}\right),$$

for any T with probability over $1 - t_1 \exp(-cd)$.

The regret upper bound exhibits logarithmic growth with respect to the horizon T, aligning with the well-established optimal regret bounds in online optimization (Hazan et al., 2007a; Orabona, 2019). The other term in the regret upper bound is solely dependent on the initialization error and remains unaffected by the horizon T. This indicates that the initialization error does not exert a persistent influence over the regret. In contrast, regret upper bounds established by Hazan et al. (2007a), Orabona (2019), Cai et al. (2023), and Han et al. (2022) all involve a multiplicative term of the form $\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2^2 \log T$, indicating that poor initialization may lead to substantially larger regret.

2.2 Batch Learning

Suppose the server receives a batch of data points $\mathcal{D}_t = \{(\mathbf{X}_i^{(t)}, Y_i^{(t)})\}_{i=1}^{n_t}$ at each time t, consisting of n_t i.i.d. observations that follow the linear model (1). The loss function at time t is given by

$$f_t(\boldsymbol{\beta}) = \frac{1}{n_t} \sum_{i=1}^{n_t} \rho_{Q,\tau}(Y_i^{(t)} - \mathbf{X}_i^{(t)\top}\boldsymbol{\beta}), \tag{3}$$

which represents the empirical risk based on n_t observations. The special case considered in Section 2.1, where only a single data point is observed, corresponds to $n_t = 1$. As n_t increases, the objective function (3) becomes more amenable to analysis due to the concentration phenomenon. As before, we allow for an arbitrary initialization $\boldsymbol{\beta}_0$. At each time t, the current estimate is $\boldsymbol{\beta}_t$, and upon receiving the new batch $\{(\mathbf{X}_i^{(t)}, Y_i^{(t)})\}_{i=1}^{n_t}$, the update is given by

$$\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t - \eta_t \cdot \frac{1}{n_t} \sum_{i=1}^{n_t} (\tau - \mathbb{I}\{Y_i^{(t)} - \langle \mathbf{X}_i^{(t)}, \boldsymbol{\beta}_t \rangle\}) \cdot \mathbf{X}_i^{(t)},$$

where η_t denotes the stepsize. The choice of η_t will be discussed in Theorem 5. In contrast to our focus, Jiang and Yu (2022) and Wang et al. (2022) examine asymptotic batch learning for quantile regression, specifically in the regime where the batch size n_t tends to infinity. Separately, Do et al. (2009) investigates batch learning from an optimization-theoretic perspective. More recently, the advantages of batch learning have been extensively explored in the literature on bandit algorithms (Ren and Zhou, 2023; Gao et al., 2019; Han et al., 2020). Despite these advances, the computational and statistical aspects of batched quantile regression remain largely underexplored. Relative to the setting considered in Section 2.2, access to larger batches provides technical advantages in analyzing the non-smooth objective function, especially when the current iterate is far from the ground truth, and gives rise to distinct theoretical properties.

The excess risk and sub-gradient of the objective function play central roles in characterizing the convergence dynamics of (sub-)gradient descent algorithms. While the expected excess risk properties illustrated in Figure 1 also hold for the objective function (3), the sub-gradient of the objective function $f_t(\boldsymbol{\beta})$, denoted by $\mathbf{g}_t \in \partial f_t(\boldsymbol{\beta})$, behaves distinctly from that in Section 2.1. The three phases in the behavior of the expected sub-gradient norm are illustrated in Figure 3. Essentially, when $\boldsymbol{\beta}$ is sufficiently far from the ground truth, the expected squared norm of the sub-gradient, $\mathbb{E}\|\mathbf{g}\|_2^2$, is bounded above by a constant that is independent of the distance $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2$. When $\boldsymbol{\beta}$ is closer, but not too close, to $\boldsymbol{\beta}^*$, the expected norm is upper bounded by $O(\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2)$, and this bound remains independent of the batch size. Finally, when $\boldsymbol{\beta}$ is very close to the ground truth, the expected sub-gradient norm depends only on the batch size, dimensionality, and noise characteristics.

Hence, the online sub-gradient descent algorithm applied to (3) exhibits a three-phase convergence behavior, provided that the stepsize is carefully chosen to align with the aforementioned statistical properties. For notational convenience, define $\bar{\tau} = \max\{\tau, 1 - \tau\}$.

Theorem 5 *Suppose Assumptions 1 and 2 hold. Let $\boldsymbol{\beta}_0$ be an arbitrary initialization satisfying $\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2 \le D_0$ for some $D_0 > 0$. There exist universal constants $\{c_i\}_{i=0}^5, C_0, C_1, C_2 > 0$ such that if the batch size $n_t \ge n \ge C_0(C_u/C_l)\bar{\tau}^2 d$ for all t, then the sequence $\{\boldsymbol{\beta}_t\}_{t\ge 1}$ generated by the online sub-gradient descent algorithm exhibits the following dynamics:*
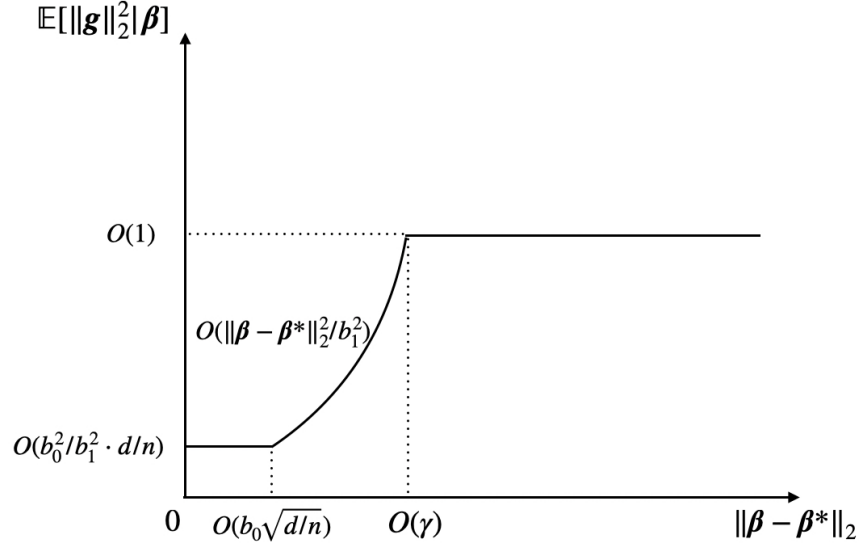
Figure 3: Expected length of the sub-gradient for the empirical quantile loss. Y-axis: upper bound of $\mathbb{E}[\|\mathbf{g}\|_2^2|\boldsymbol{\beta}]$, where $\mathbf{g} \in \partial \frac{1}{n}\sum_{i=1}^{n} \rho_{Q,\tau}(Y_i - \mathbf{X}_i^\top \boldsymbol{\beta})$; X-axis: $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2$. It reveals three phases of properties associated with the sub-gradient, depending on the closeness between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$.

(1). during phase one, where $\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2 \geq 8C_l^{-1/2}\gamma$, by choosing a stepsize $\eta_0 = C\sqrt{C_l}/C_u \cdot D_0$ and $\eta_t = (1 - cC_l/C_u)\eta_{t-1}$ with any $D_0 \geq \|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2$, $C > 1$, $c < 0.1$, it holds with probability exceeding $1 - \exp(-c_0 d) - \exp\big(-\sqrt{n_t/\log(n_t)}\big)$ that

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2 \leq \left(1 - c\frac{C_l}{C_u}\right)D_0;$$

the conclusion of phase one is reached after $t_1 = O(\log(\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2/\gamma))$ iterations;

(2). in phase two, characterized by $C_1 C_u^{1/2} C_l^{-1}\bar{\tau}\sqrt{d/n} \cdot b_0 \leq \|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2 < 8C_l^{-1/2}\gamma$, by selecting a constant stepsize $\eta_t = \eta \in (C_l/C_u^2)(b_1^2/b_0) \cdot [c_1, c_2]$, it holds with probability at least $1 - 3\exp(-c_0 d) - \exp(-\sqrt{n_t/\log n_t})$ that

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2 \leq \left(1 - 0.0005\frac{b_1^2}{b_0^2}\frac{C_l^2}{C_u^2}\right) \cdot \|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2;$$

phase two requires $O\big(\log((n/d)(\gamma/b_0))\big)$ iterations to conclude, providing an estimate $\boldsymbol{\beta}_{t_2}$ that satisfies $\|\boldsymbol{\beta}_{t_2} - \boldsymbol{\beta}^*\|_2 \asymp C_u^{1/2}C_l^{-1}\bar{\tau}\sqrt{d/n} \cdot b_0$ under the specified conditions;

(3). in phase three, characterized by $\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2 \leq C_1 C_u^{1/2}C_l^{-1}\bar{\tau}\sqrt{d/n} \cdot b_0$, by choosing a stepsize $\eta_t := \frac{C_a}{C_l}\frac{b_0}{t+C_b-t_2}$ with arbitrary constants $C_a \geq 12$ and $C_b > C_a/12$, it holds with probability exceeding $1 - c_2\exp(-c_1 d) - \exp(-\sqrt{n_t/\log n_t})$ that

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2 \leq C^* \frac{C_u^2 b_0^2}{C_l^2 b_1^2}\frac{\bar{\tau}^2}{t+C_b+1-t_2}\frac{d}{n} \cdot \frac{C_u}{C_l^2}b_0^2,$$

where the constant C^* depends solely on C_a, C_b, and t_2 denotes the conclusion time of the phase two convergence.

According to Theorem 5, the online sub-gradient descent algorithm exhibits fast linear convergence during phases one and two. The stepsize schemes used in these phases mirror those observed in the two-phase convergence of offline quantile regression (Shen et al., 2023). Specifically, the first phase requires t_1 iterations and uses a total of $O(d \log(\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2/\gamma))$ data vectors. This aligns with the data scale required in phase one of the online learning setting, as outlined in Theorem 1. In the third phase, a stepsize of $O(1/t)$ is essential for attaining long-term statistical optimality. The final error rate exhibits a continual decrease as additional data becomes available, demonstrating the efficacy of the proposed stepsize scheme in seamlessly integrating sequentially arriving data with an unknown horizon. Furthermore, at each iteration, the failure probability diminishes exponentially with respect to the dimension.

Corollary 6 *Under the conditions of Theorem 5, and if the horizon $T \geq C_0(\log(\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2/b_0) + \log(n/d))$, then, with a probability exceeding $1 - c_1 T \exp(-c_0 d) - T \exp(-\sqrt{n/\log n})$, the online sub-gradient descent algorithm produces a final estimate with an error rate given by*

$$\|\boldsymbol{\beta}_T - \boldsymbol{\beta}^*\|_2^2 \leq C^* \frac{C_u^2 b_0^2}{C_l^2 b_1^2} \frac{C_u \bar{\tau}}{C_l^2} \cdot \frac{d}{nT} \cdot b_0^2,$$

where the constant C^* depends only on C_a and C_b, and $\bar{\tau} = \max\{\tau, 1 - \tau\}$.

If all batch sizes satisfy $n_t \asymp n$, then the total number of data points used is $O(Tn)$. In this setting, the error rate established in Corollary 6 matches the minimax-optimal rate in the offline setting (He et al., 2023). Notably, this rate is robust to initialization errors, similar to the rate obtained for online learning in Corollary 2. When the initialization is sufficiently close to $\boldsymbol{\beta}^*$, a trade-off emerges between short-term accuracy and long-term optimality, as discussed in Remark 3.

Theorem 7 *Under the same conditions and stepsize scheme as in Theorem 5, the online sub-gradient descent algorithm attains a regret upper bound that for any $T \geq 1$ with probability over $1 - c_1 t_1 \exp(-c_0 d) - t_1 \exp(-\sqrt{n/\log n})$,*

$$\text{Regret}_T \leq C_1 \frac{C_u^3}{C_l^3} \frac{b_0^2}{b_1^2} \max\{\sqrt{C_u}\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2, \gamma^2/b_0\} + C_2 \frac{C_u^3}{C_l^3} \frac{b_0^3}{b_1^3} \frac{d}{n} b_0 \log\left(\frac{T - t_1 + C_b}{C_b}\right).$$

The regret upper bound consists of two terms: one that depends on the initialization error but is independent of T, and another that grows logarithmically with T yet remains independent of the initialization. When the total number of data points consumed is equal–denoted by N, and assuming $N \geq C_1 \max\{n \log(\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2/b_0), n \log(n/d)\}$–both online and batch learning achieve an optimal estimation error rate of $O(b_0^2 d/N)$. However, batch learning incurs significantly lower regret by leveraging a larger number of data points in each iteration, thereby accelerating the learning process. Specifically, as shown in Theorem 4, online learning achieves a regret upper bound of $O(d\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2 + d \log(N/d))$, whereas batch learning attains a considerably smaller bound of $O(\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2 + (d/n) \log(N/n))$. It is worth noting, however, that batch learning requires substantially greater storage capacity.

2.3 Sequential Learning with Infinite Storage

Consider a scenario in which the server hosting $\boldsymbol{\beta}_{t+1}$ has unlimited storage capacity, retaining all historical data and updating the estimate only upon the arrival of new samples. Let $\boldsymbol{\beta}_0$ be an arbitrary initialization. At each time t, the server incorporates all observations received up to and including time t when computing the update. Let $\{(Y_i^{(t)}, \mathbf{X}_i^{(t)})\}_{i=1}^{n_t}$ denote the n_t observations that arrive at time t. The loss function at time t is then given by

$$f_t(\boldsymbol{\beta}) := \frac{1}{\sum_{l=0}^{t} n_l} \sum_{l=0}^{t} \sum_{i=1}^{n_l} \rho_{Q,\tau}(Y_i^{(l)} - \mathbf{X}_i^{(l)\top}\boldsymbol{\beta}), \tag{4}$$

which corresponds to the empirical risk over all accumulated data. The online sub-gradient descent algorithm updates the iterate via $\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t - \eta_t \cdot \mathbf{g}_t$, where η_t is the stepsize and $\mathbf{g}_t \in \partial f_t(\boldsymbol{\beta}_t)$ is a sub-gradient. Specifically, we use the update rule

$$\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t - \eta_t \cdot \frac{1}{\sum_{l=0}^{t} n_l} \sum_{l=0}^{n_l} \sum_{i=1}^{n_t} (\tau - \mathbb{I}\{Y_i^{(l)} - \langle \mathbf{X}_i^{(l)}, \boldsymbol{\beta}_t \rangle\}) \cdot \mathbf{X}_i^{(l)}.$$

The stepsize scheme η_t will be discussed in Theorem 8. Although this algorithm requires iterating over the entire dataset accumulated so far at each update, it remains computationally more efficient than a full offline optimization approach (Shen et al., 2023).

The objective function in (4) retains its inherent two-phase structure, as illustrated in Figure 1. Notably, although this formulation incurs substantial storage and computational costs, these demands facilitate a faster convergence rate of the algorithm, as established in the following theorem.

Theorem 8 *Suppose Assumptions 1 and 2 hold, and let $\boldsymbol{\beta}_0$ be an arbitrary initialization satisfying $\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2 \leq D_0$ for some $D_0 > 0$. There exist absolute positive constants c_0, c_1, c_2, C_1, C_2 such that, if the initial storage $n_0 \geq C_1 d$, then the sequence $\{\boldsymbol{\beta}_t\}_{t \geq 1}$ generated by the online sub-gradient descent algorithm has the following dynamics:*

(1). During phase one, when $\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2 \geq 8\gamma$, by selecting a stepsize $\eta_t := \frac{\sqrt{C_l}}{8C_u}\left(1 - \frac{1}{100}\frac{C_l}{C_u}\right)^t \cdot D_0$, it holds with probability at least $1 - \exp(-c_0 d) - \exp(-\sqrt{\sum_{l=0}^{t} n_l / \log \sum_{l=0}^{t} n_l})$ that

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2 \leq \left(1 - \frac{1}{100}\frac{C_l}{C_u}\right)^{t+1} \cdot D_0.$$

Phase one concludes after $t_1 = O\big(\log(\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2/\gamma)\big)$ iterations.

(2). In phase two, when $\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2 < 8\gamma$, by selecting a stepsize $\eta_t := \eta \asymp \frac{C_l}{C_u^2} \cdot \frac{b_1^2}{b_0}$, it holds with probability at least $1 - \exp(-c_0 d) - \exp(-\sqrt{\sum_{l=0}^{t} n_l / \log \sum_{l=0}^{t} n_l})$ that

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2 \leq \left(1 - c_1 \frac{b_1^2}{b_0^2} \cdot \frac{C_l^2}{C_u^2}\right) \cdot \|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2 + c_2 \frac{d}{C_u \sum_{l=0}^{t} n_l} \cdot b_1^2.$$

As shown in Theorem 8, the convergence behavior of the algorithm proceeds in two distinct phases. The initial phase requires a geometrically decaying sequence of stepsizes, while the second phase adopts a constant stepsize. Both phases are proven to exhibit linear convergence. For simplicity, assume that $n_l = n$ holds for all $l \geq 1$. During the second phase (i.e., for $t > t_1$), the error rate satisfies

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2 \leq \left(1 - c_1 \frac{b_1^2}{b_0^2} \cdot \frac{C_l^2}{C_u^2}\right)^{t+1-t_1} \cdot \|\boldsymbol{\beta}_{t_1} - \boldsymbol{\beta}^*\|_2^2 + C_2 \frac{d}{n_0 + tn} \cdot \frac{C_u}{C_l^2} \cdot b_0^2 \cdot \left(\frac{b_0}{b_1} \cdot \frac{C_u}{C_l}\right)^2,$$

which leads to the following corollary.

Corollary 9 *Assume the same conditions as in Theorem 8, and suppose that $n_l = n$ for all $l \geq 1$. If the total time horizon satisfies $T \geq C^* \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2/\gamma$, for some constant C^* that may depend on (C_l, C_u, n_0, n), with probability over $1 - \sum_{t=0}^{T} \exp(-\sqrt{(n_0 + tn)/\log(n_0 + tn)})$*
$- T \exp(-c_0 d)$, it holds that

$$\|\boldsymbol{\beta}_T - \boldsymbol{\beta}^*\|_2 \leq C \left(\frac{b_0}{b_1} \cdot \frac{C_u}{C_l}\right)^2 \cdot \frac{C_u}{C_l^2} \cdot \frac{d}{n_0 + Tn} b_0^2.$$

The corollary above demonstrates that, despite the server updating only once upon the arrival of each new batch of samples, the procedure consistently achieves statistical optimality, even when the total time horizon is unknown.

3. Numerical Experiments

In this section, we present experiments to empirically validate our theoretical findings. We first generate synthetic data for simulation studies, followed by an application of our proposed methods to a real-data example.

3.1 Simulations

In this section, we numerically examine the performance of the proposed online QR algorithms. We first demonstrate the effectiveness of our stepsize scheme, and then compare the final accuracy of different online estimators with that of the offline estimate. We further illustrate the advantage of online quantile regression over online least squares regression. It is worth noting that existing online quantile algorithms (Jiang and Yu, 2022; Sun et al., 2023; Chen and Yuan, 2024) depend on storage of $O(d^2)$ summary statistics. Additionally, they require data to arrive in batches with a diverging batch size. Their requirements limit their applicability for empirical performance comparisons. We use the relative error $\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2/\|\boldsymbol{\beta}^*\|_2$ as the main metric. For simplicity, the initialization $\boldsymbol{\beta}_0$ is set to be $\mathbf{0}$ throughout our numerical studies.

STEPSIZE SCHEME. We begin by demonstrating the effectiveness of our stepsize scheme, as outlined in Theorem 1, which is theoretically guided by statistical regularities. The selection of parameters in the proposed stepsize is quite flexible, both theoretically (as outlined in Theorem 1) and in practice. For instance, concerning the parameter C_a required by the stepsize schedule in the second phase of iterations, we observe that our algorithm always performs well as long as it is not excessively small.
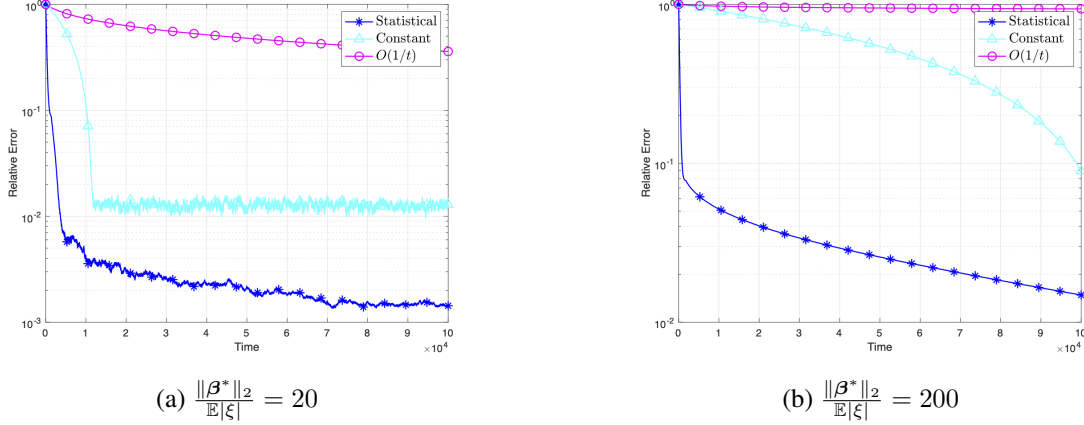
(a) $\frac{\|\boldsymbol{\beta}^*\|_2}{\mathbb{E}|\xi|} = 20$ (b) $\frac{\|\boldsymbol{\beta}^*\|_2}{\mathbb{E}|\xi|} = 200$

Figure 4: Relative error *versus* time/iterations in online (one-sample) learning ($n_t \equiv 1$). The dimension $d = 100$, unknown horizon $T = 10^5$, and quantile loss parameter $\tau = 1/2$. The convergence performances of online sub-gradient descent are examined under three stepsize schemes: *Statistical* stands for our stepsize scheme guided by Theorem 1, *Constant* stands for the stepsize scheme $\eta_t \equiv const$ (Zhang, 2004), $O(1/t)$ means the decaying stepsize scheme $\eta_t = O(1/t)$ (Duchi and Singer, 2009). Left(a): moderate SNR; right(b): strong SNR.

In all the experiments, the stepsize for the first phase is scheduled as $\eta_t = (1 - 0.5/d)^t \eta_0$, where η_0 is the initial stepsize. The choice of stepsize in second phase involves specific parameters where we set $C_a = 20$ and $C_b = 30$. We fix the dimension at $d = 100$, the unknown horizon $T = 10^5$, and sample t_ν-distributed noise with a degree of freedom $\nu = 1.1$. The performance of proposed stepsize scheme is compared with those of two existing alternative stepsize scheme: the $\eta_t = O(1/t)$ decaying scheme and a constant stepsize scheme $\eta_t \equiv const$. For detailed discussions on these stepsize schemes, we refer to Duchi and Singer (2009) and Zhang (2004).

The convergence performances of online sub-gradient descent under the aforementioned three stepsize schemes are presented in Figure 4. For both the moderate and strong SNR cases, our proposed stepsize scheme can ensure a linear convergence of the online sub-gradient descent algorithm in the first phase. The linear convergence behavior stops once the algorithm reaches at a sufficiently accurate estimate, i.e., when the error rate is dominated by the noise scale $\mathbb{E}|\xi|$. Our proposed scheme then resets the stepsize and the algorithm enters second phase of iterations. The error rate continues to decrease in the second phase, which is sub-linear and exhibits an $O(1/t)$ convergence rate. The two-phase convergence phenomenon is consistent with the theoretical discoveries. In contrast, the constant stepsize scheme can also achieve an error rate the same as the one our algorithm achieves at the end of first phase iterations. However, it cannot further improve the estimate or may converge too slowly resulting into a statistically sub-optimal estimate. It is also observed that employing a relatively large constant stepsize can facilitate faster convergence in the initial stage, which was claimed by the theoretical results in Cai et al. (2023). The performance under the $O(1/t)$ stepsize scheme is considerably inferior to those under the other two stepsize schemes.

STATISTICAL ACCURACY COMPARISONS. We now evaluate the statistical accuracy of the final estimator output by the online sub-gradient descent algorithm. The error rate achieved by offline quantile regression is used as the benchmark. The dimension d and noise distribution are set the
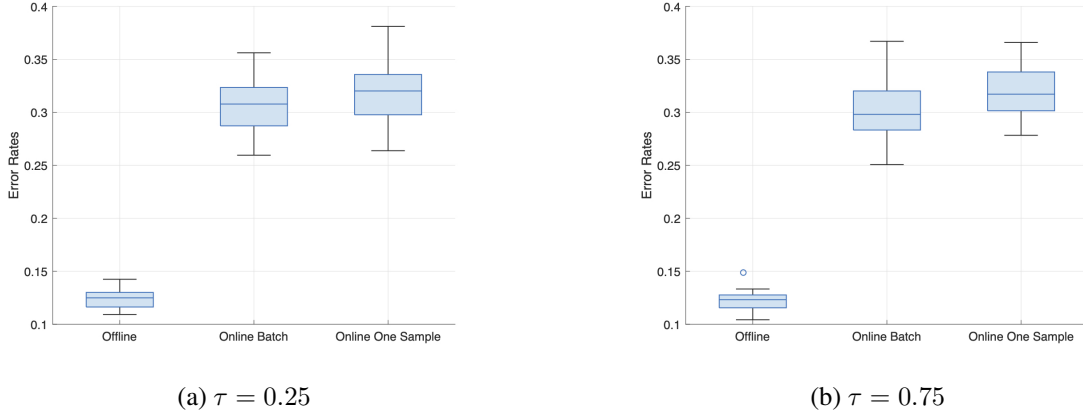
(a) $\tau = 0.25$ (b) $\tau = 0.75$

Figure 5: Error rates of offline and online regression using quantile loss $\rho_{Q,\tau}(\cdot)$. Online One Sample refers to the online learning algorithm studied in Section 2.1. The dimension $d = 100$, total sample size $n = 20,000$, the batch size $n_t \equiv 100$, and noise has a $t_{1.1}$ distribution. Box-plots are drawn based on 50 independent simulations.

same as those in the previous set of simulations. Both the online one-sample learning and batch learning are studied in the experiment. The total sample size is set at $n = 20,000$. Figure 5 displays a box plot of error rates based on 50 replications. For each simulation, the online batch learning algorithm and one-sample learning algorithm take approximately 2 and 10 seconds, respectively, on a MacBook Pro 2020. The offline learning, implemented using the quantreg package (Portnoy and Koenker, 1997), takes more than 2 minutes. Figure 5 shows that offline learning achieves the best statistical accuracy when the total sample size is relatively small, while online one-sample learning and batch learning achieve comparable statistical accuracy. These are consistent with our theoretical findings. Simulation results for the large sample size $n = 50,000$ are displayed in Figure 6, in which case the offline learning achieves only slightly better accuracy than its online counterparts. However, online algorithms enjoy much higher computational efficiencies. On the same Mac Pro, the online one-sample, batch, and offline learning methods take approximately 30, 5, and 520 seconds, respectively.

CONVERGENCE DYNAMICS COMPARISONS. While online QR is motivated primarily for treating heavy-tailed noise, it is still of interest to examine its performance under Gaussian noise. Towards that end, we compare the performance of our proposed online QR algorithms with that of classical online least squares algorithm (Zhang, 2004). We will show that the proposed online QR algorithms are not only robust in the presence of heavy-tailed noise or responses, but also are as efficient as the classical online least squares algorithm if the noise is Gaussian.

While online learning with square loss has been extensively studied (Orabona, 2019; Hazan, 2016), its stepsize scheme guided by statistical properties remains relatively under-explored. Here we briefly explain the appropriate stepsize scheme to achieve long-term statistical optimality for online least squares algorithm with a focus on sub-Gaussian noise and undetermined horizon. Considering the square loss function at time t as $f_t(\boldsymbol{\beta}) := (Y_t - \mathbf{X}_t^\top \boldsymbol{\beta})^2$, the expected length of gradient $\mathbf{g}_t$ satisfies

$$\mathbb{E}\left[\|\mathbf{g}_t\|_2^2 \big| \boldsymbol{\beta}_t\right] \leq 4dC_u(\mathbb{E}\xi^2 + C_u\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2).$$

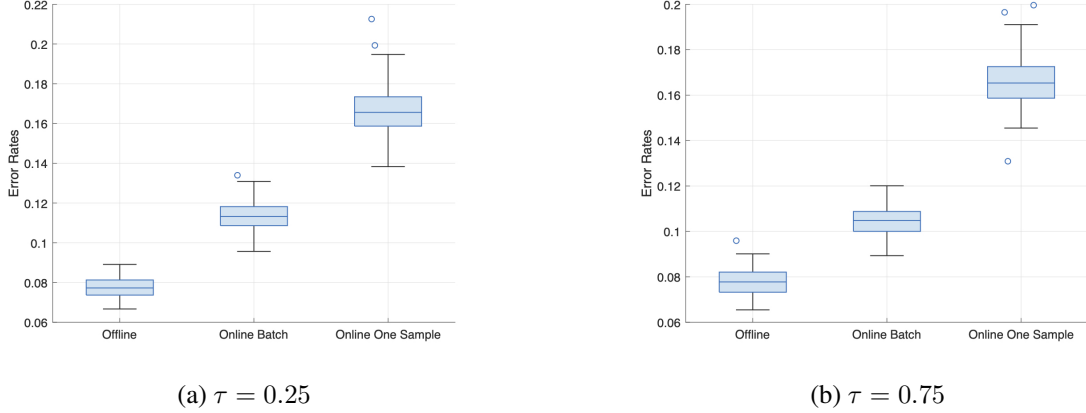(a) $\tau = 0.25$ (b) $\tau = 0.75$

Figure 6: Error rates of offline and online regression using quantile loss $\rho_{Q,\tau}(\cdot)$. Online One Sample refers to the online learning algorithm studied in Section 2.1. The dimension $d = 100$, total sample size $n = 50,000$, the batch size $n_t \equiv 200$, and noise has a $t_{1.1}$ distribution. Box-plots are drawn based on 50 independent simulations.

The gradient length is decided primarily by $\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2$ when it is large. Conversely, when $\boldsymbol{\beta}_t$ is sufficiently close to the oracle, the gradient length is determined by $\mathbb{E}\xi^2$. This implies an appropriate stepsize scheme for online least squares algorithm should also consists of two phases: a constant stepsize $\eta_t \equiv \eta = O(d^{-1})$ in the first phase; a decaying stepsize schedule $\eta_t = O((t+d)^{-1})$ in the second phase. The detailed proof of the convergence dynamics under this stepsize scheme is almost identical to that of Theorem 1 and thus omitted.

We compare the convergence dynamics of online one-sample QR algorithm and online least squares algorithm (equipped with the aforementioned two-phase stepsize scheme). The dimension $d = 100$, quantile loss parameter $\tau = 0.5$, and horizon is unknown. The results under Gaussian noise and $t_{1.1}$ noise are presented in Figure 7. The convergence dynamics of the two algorithms are comparable under Gaussian noise, both showing a linear convergence in the first phase and an $O(1/t)$ decaying rate afterwards. They achieve almost the same statistical accuracy in the end. Under $t_{1.1}$-distributed noise, online least squares algorithm does not converge. In contrast, the proposed online QR algorithm ensures stable convergence and achieves error rates comparable to those under Gaussian noise.

PARAMETER SENSITIVITY. Here we empirically examine the flexibility of tuning parameters discussed in Theorem 1. The simulation results demonstrating parameter sensitivity are presented in Figure 8. Following the same settings as before, we set $d = 100$, $C_l = C_u = 1$ and $\|\boldsymbol{\beta}^*\|_2/\gamma = 20$. Figure 8a illustrates the impact of varying the geometric decay rate $c \in \{0.5, \ldots, 0.05\}$ for the first-phase stepsize, defined as $\eta_t = (1 - c/d)^t \eta_0$, while the second-phase stepsize is fixed as $\eta_t = 15/(t - t_1 + 20d)$. Interestingly, when t is sufficiently large, different choices of the decay rate c result in similar convergence behavior. This suggests that the decay rate c in the first-phase stepsize is indeed flexible in practice. Figures 8b–8d further explore convergence dynamics under various values of C_a and C_b. Specifically, Figure 8c shows the sensitivity to C_b with $C_a = 15$ fixed, while Figure 8d examines the influence of C_a with $C_b = 100$ held constant. These results indicate that as long as C_a and C_b are not too small, the estimation error remains on the same scale, and

19
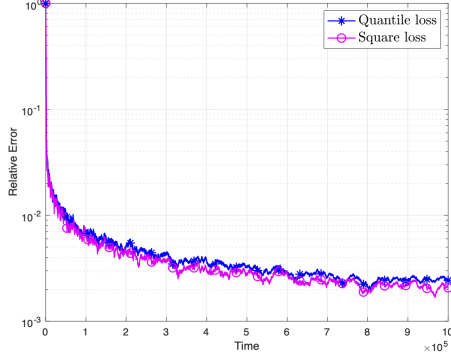
(a) $\frac{\|\boldsymbol{\beta}^*\|_2}{\mathbb{E}|\xi|} = 20$, Gaussian noise
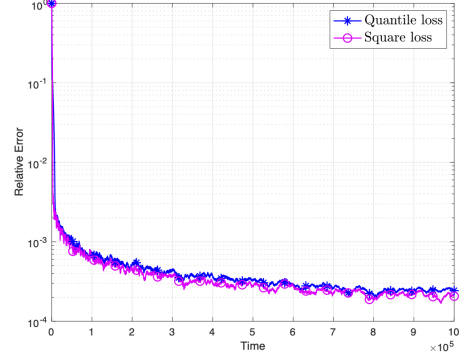
(b) $\frac{\|\boldsymbol{\beta}^*\|_2}{\mathbb{E}|\xi|} = 200$, Gaussian noise
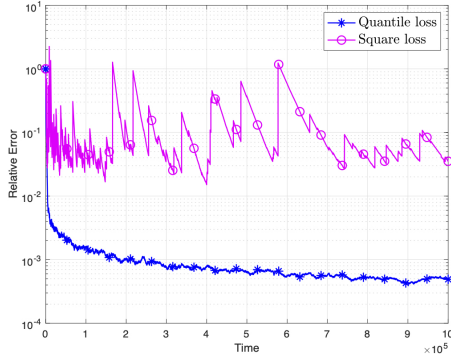
(c) $\frac{\|\boldsymbol{\beta}^*\|_2}{\mathbb{E}|\xi|} = 20$, $t_{1.1}$ noise
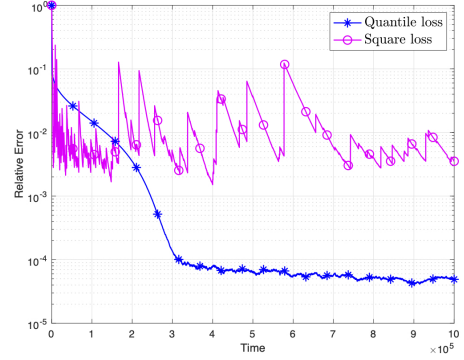
(d) $\frac{\|\boldsymbol{\beta}^*\|_2}{\mathbb{E}|\xi|} = 200$, $t_{1.1}$ noise

Figure 7: Relative error *versus* time/iterations for online QR and least squares algorithms. The dimension is $d = 100$, the (unknown) time horizon is $T = 10^5$, and the quantile level is $\tau = 1/2$. The proposed online QR algorithm is robust to heavy-tailed noise and achieves comparable performance to online least squares under Gaussian noise.
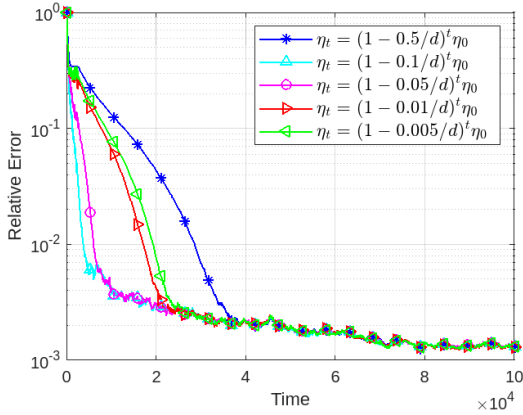
(a) Convergence under varying geometrically decaying step sizes for $t \leq t_1$, followed by a common fixed stepsize for $t \geq t_1$.
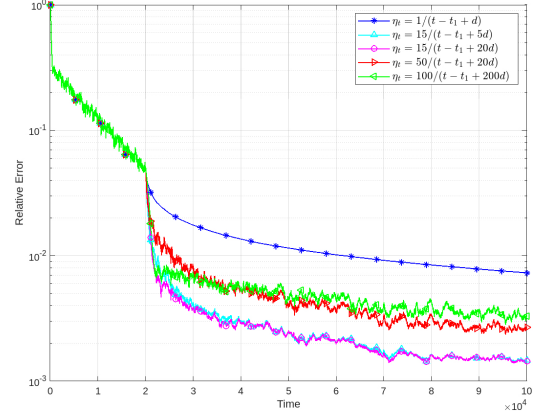
(b) Convergence under varying values of C_a and C_b for $t \geq t_1$, with a common stepsize applied for $t \leq t_1$.
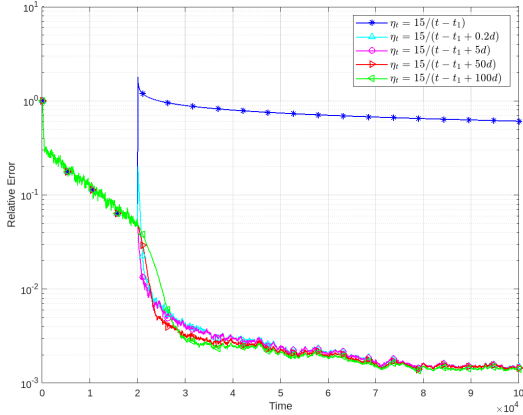
(c) Convergence under a fixed C_a for $t \geq t_1$, and a shared stepsize scheme for $t \leq t_1$.
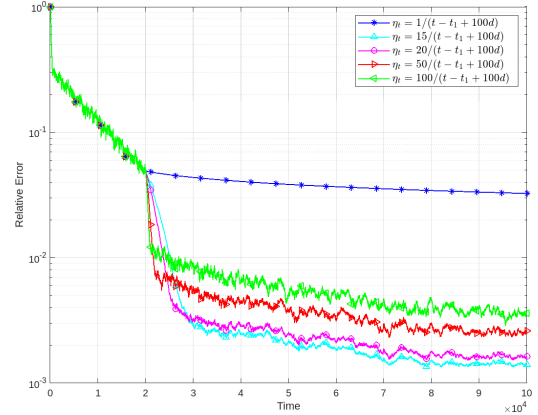
(d) Convergence under a fixed value of C_b for $t \geq t_1$, and a shared stepsize scheme for $t \leq t_1$.

Figure 8: Relative error *versus* time/iteration for online QR. The dimension is $d = 100$, with a signal-to-noise ratio of $\frac{\|\beta^*\|}{\mathbb{E}|\xi|} = 20$. The noise is generated from a t-distribution with 1.1 degrees of freedom. The time horizon is set to $T = 10^5$.

variations in C_b have little effect on performance. However, if both parameters are too small (e.g., $C_a = 1$ and $C_b = 0$), the theoretical $O(1/t)$ convergence rate is no longer guaranteed. Furthermore, Figures 8b–8d depict the convergence dynamics under varying values of C_a and C_b. Specifically, Figure 8c illustrates the effect of different choices of C_b while fixing $C_a = 15$, and Figure 8d examines the sensitivity to C_a with $C_b = 100$ held constant. These results suggest that as long as C_a and C_b are not too small, the estimation error remains on the same scale, and the choice of C_b has minimal impact on performance. In contrast, when both parameters are set to overly small values, for example, $C_a = 1$ and $C_b = 0$, the convergence guarantee at the $O(1/t)$ rate no longer holds. These empirical findings support the theoretical flexibility of C_a and C_b as established in Theorem 1.

REGRET DYNAMICS. In this paragraph, we examine the regret dynamics illustrated in Figure 9, under the setting where $d = 200$, $T = 10^5$, and the noise follows a t-distribution with 1.1 degrees of

(a) Regret Dynamics with Oracle $\log(t)$ Function

(b) Regret Dynamics under Different Stepsizes
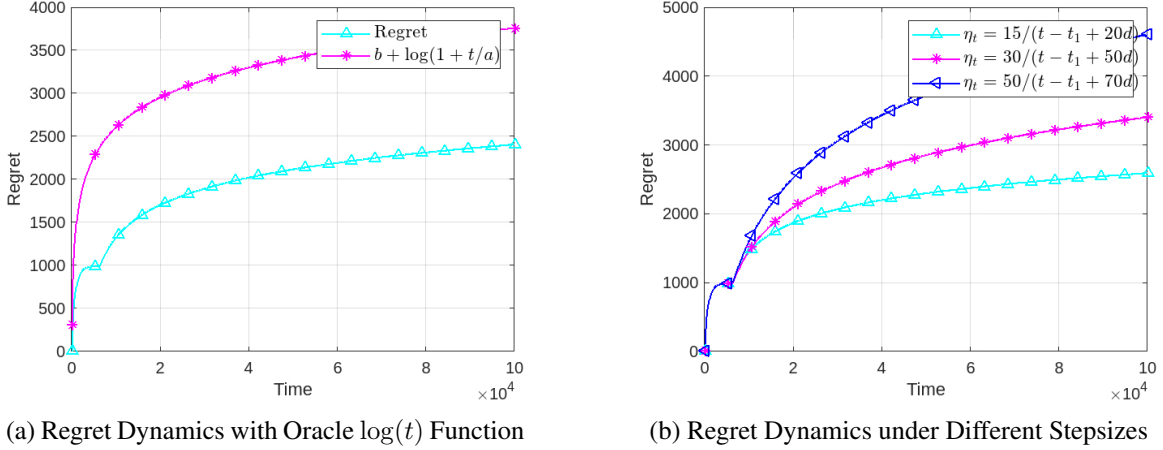
Figure 9: Regret dynamics *versus* time/iteration for online QR. The dimension $d = 100$, signal-to-noise ratio $\frac{\|\boldsymbol{\beta}^*\|}{\mathbb{E}|\xi|} = 20$, unknown horizon $T = 10^5$, $t_{1.1}$ noise.

freedom ($t_{1.1}$). The left panel (Figure 9a) corresponds to the stepsize schemes $\eta_t = (1 - 0.05/d)^t \eta_0$ and $C_a/(t - t_1 + C_b d)$. The pink star-dotted line represents a benchmark function of the form $b + \log(1 + t/a)$. As shown in Figure 9a, the curves exhibit highly parallel growth patterns. Figure 9b displays the regret values over time t under different stepsize schemes, all showing a logarithmic trend of the form $b + \log(1 + t/a)$, albeit with different parameterizations.

3.2 Real Data Example

In this section, we analyze a real-world dataset on news popularity provided by Moniz and Torgo (2018). Our focus is on news articles categorized under Economy that were published on Facebook. The objective is to predict the final popularity of each news item using three sets of predictors: the sentiment score of the headline text, the sentiment score of the body text, and the popularity levels measured at successive time intervals post-publication ($0 - 6h, \ldots, 18 - 24h$). The dataset contains a total of $N = 29928$ samples and $d = 7$ predictors. We randomly split the dataset into a training set with $n_1 = 20000$ observations and a test set with $n_2 = 9928$ observations. We apply the proposed online QR and batch learning algorithms to the training set. As benchmarks, we include the offline QR, implemented using `quantreg` (Portnoy and Koenker, 1997), and the ordinary least squares (OLS) estimators. It is worth noting that this dataset was also examined by Chen and Zhou (2020), although their analysis was based on a different subset of the data, as some records from the original source are no longer available. We report the test errors $n_2^{-1} \sum_{(Y_i, \mathbf{X}_i) \in \text{TestData}} |Y_i - \mathbf{X}_i^\top \widehat{\boldsymbol{\beta}}|$ in Table 1. The outcome variable Y ranges from approximately -1 to 10^4, with an average absolute value on the test set exceeding 50. Despite this wide range, the prediction errors reported in Table 1 are substantially smaller, indicating strong predictive performance. Among the three QR methods–online, batch, and offline–both online and batch QR yield slightly lower prediction errors compared to offline QR. All three quantile-based approaches, however, significantly outperform the OLS in terms of test error. Moreover, consistent with the findings of Chen and Zhou (2020), we observe that the estimated coefficients for the sentiment scores of both the headline and body text are close

Method	Online QR	Online Batch QR	Offline QR	OLS
Test Error	8.5824	8.5084	8.1895	9.8752

Table 1: Prediction performance on the news popularity dataset.

to zero. This suggests that, in this setting, sentiment features provide limited predictive value for news popularity.

4. Proof of Theorem 1

This section presents the proof of Theorem 1. Proofs of the remaining results are deferred to the supplementary material. The update procedure $\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t - \eta_t \mathbf{g}_t$ yields

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2 = \|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2 - 2\eta_t \langle \boldsymbol{\beta}_t - \boldsymbol{\beta}^*, \mathbf{g}_t \rangle + \eta_t^2 \|\mathbf{g}_t\|_2^2, \tag{5}$$

where the sub-gradient $\mathbf{g}_t$ is given by $\mathbf{g}_t = \mathbf{X}_t\{-\tau \cdot 1_{(Y_t > \mathbf{X}_t^\top \boldsymbol{\beta}_t)} + (1-\tau) \cdot 1_{(Y_t < \mathbf{X}_t^\top \boldsymbol{\beta}_t)} + \delta \cdot 1_{(Y_t = \mathbf{X}_t^\top \boldsymbol{\beta}_t)}\}$, and δ can take any value in the range $[-\tau, 1-\tau]$. It follows that

$$\|\mathbf{g}_t\|_2^2 \geq \min\{\tau, 1-\tau\}^2 \|\mathbf{X}_t \cdot 1_{(Y_t \neq \mathbf{X}_t \boldsymbol{\beta}_t)}\|_2^2 + \delta^2 \|\mathbf{X}_t \cdot 1_{(Y_t = \mathbf{X}_t \boldsymbol{\beta}_t)}\|_2^2$$

and

$$\|\mathbf{g}_t\|_2^2 \leq \max\{\tau, 1-\tau\}^2 \|\mathbf{X}_t \cdot 1_{(Y_t \neq \mathbf{X}_t \boldsymbol{\beta}_t)}\|_2^2 + \delta^2 \|\mathbf{X}_t \cdot 1_{(Y_t = \mathbf{X}_t \boldsymbol{\beta}_t)}\|_2^2.$$

By Lemma 20, $\|\mathbf{X}_t\|_2^2 \leq 2C_u d$ holds with probability at least $1 - \exp(-cd)$, which further implies that $\|\mathbf{g}_t\|_2^2 \leq 2\bar{\tau}^2 C_u d$ for all $t \geq 0$, where $\bar{\tau} = \max(\tau, 1-\tau)$.

FIRST PHASE ANALYSIS. We aim to establish the convergence dynamics

$$\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2 \leq C^* \left(1 - c_5 \frac{C_l}{C_u} \frac{1}{\bar{\tau}^2 d}\right)^t \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2,$$

through induction, where the constant $C^* > 1$ remains fixed throughout the proof and $0 < c_5 < 1$ is some sufficiently small constant to be specified later.

Initially, the inequality $\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2 \leq C^* \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2$ holds trivially. For notational convenience, define

$$D_l := (1 - c_5 \frac{C_l}{C_u} \frac{1}{\bar{\tau}^2 d})^l \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2 \text{ for } l = 0, 1, \dots.$$

Assuming the inductive hypothesis $\|\boldsymbol{\beta}_l - \boldsymbol{\beta}^*\|_2 \leq C^* D_l$ holds for all $l = 0, 1, \dots, t$, our goal is to establish that $\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2 \leq C^* D_{t+1}$.

By convexity and the property of sub-gradient, $\langle \boldsymbol{\beta}_t - \boldsymbol{\beta}^*, \mathbf{g}_t \rangle \geq f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*)$. Substituting this into equation (5) yields

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2 \leq \|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2 - 2\eta_t\{(f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*)\} + 2\eta_t^2 C_u \bar{\tau}^2 d$$
$$= \|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2 - 2\eta_t \mathbb{E}\{f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*)|\boldsymbol{\beta}_t\} + 2\eta_t^2 C_u \bar{\tau}^2 d$$
$$- 2\eta_t\big[f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*) - \mathbb{E}\{f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*)|\boldsymbol{\beta}_t\}\big].$$

For the expected difference $\mathbb{E}\{f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*)|\boldsymbol{\beta}_t\}$, Lemma 13 implies

$$\mathbb{E}\{f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*)|\boldsymbol{\beta}_t\} \geq \sqrt{\frac{C_l}{2\pi}}\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2 - \gamma.$$

During phase one, where $\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2 \geq 8C_l^{-1/2}\gamma$, it follows from the above lower bound that

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2 \leq \|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2 - \frac{1}{3}\eta_t\sqrt{C_l}\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2 + 2\eta_t^2 C_u\bar{\tau}^2 d$$
$$- 2\eta_t\big[f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*) - \mathbb{E}\{f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*)|\boldsymbol{\beta}_t\}\big].$$

We further rewrite the above inequality as

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2 \leq \left(1 - \frac{\sqrt{C_l}}{3}\frac{\eta_t}{\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2}\right)\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2 + 2\eta_t^2\bar{\tau}^2 C_u d$$
$$- 2\eta_t\big[f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*) - \mathbb{E}\{f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*)|\boldsymbol{\beta}_t\}\big].$$

By the induction hypothesis, $\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\| \leq C^*D_t$, and hence

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2 \leq \left(1 - \frac{\sqrt{C_l}}{3}\frac{\eta_t}{C^*D_t}\right)\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2 + 2\eta_t^2\bar{\tau}^2 C_u d$$
$$- 2\eta_t\big[f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*) - \mathbb{E}\{f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*)|\boldsymbol{\beta}_t\}\big]$$

Substituting the stepsize $\eta_t = c\frac{\sqrt{C_l}}{C_u}\frac{D_t}{\bar{\tau}^2 d}$ into the inequality above, we obtain

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2 \leq \left(1 - \frac{c}{3C^*}\frac{C_l}{C_u}\frac{1}{\bar{\tau}^2 d}\right)\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2 + 2c^2\frac{C_l}{C_u}\frac{1}{\bar{\tau}^2 d}D_t^2$$
$$- 2c\frac{\sqrt{C_l}}{C_u}\frac{D_t}{\bar{\tau}^2 d}\big[f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*) - \mathbb{E}\{f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*)|\boldsymbol{\beta}_t\}\big].$$

By applying this iterative bound repeatedly starting from $t = 0$, we obtain

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2$$
$$\leq \left(1 - \frac{c}{3C^*}\frac{C_l}{C_u}\frac{1}{\bar{\tau}^2 d}\right)^{t+1}\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2^2 + 2c^2\frac{C_l}{C_u}\frac{1}{d}\frac{1}{\bar{\tau}^2}\sum_{l=0}^{t}\left(1 - \frac{c}{3C^*}\frac{C_l}{C_u}\frac{1}{\bar{\tau}^2 d}\right)^{t-l}D_l^2$$
$$- 2c\frac{\sqrt{C_l}}{C_u}\frac{1}{\bar{\tau}^2 d}\sum_{l=0}^{t}\left(1 - \frac{c}{3C^*}\frac{C_l}{C_u}\frac{1}{\bar{\tau}^2 d}\right)^{t-l}D_l\big[f_l(\boldsymbol{\beta}_l) - f_l(\boldsymbol{\beta}^*) - \mathbb{E}\{f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*)|\boldsymbol{\beta}_t\}\big].$$

We begin by bounding the second term on the right-hand side of the above inequality as

$$\sum_{l=0}^{t}\left(1 - \frac{c}{3C^*}\frac{C_l}{C_u}\frac{1}{\bar{\tau}^2 d}\right)^{t-l}D_l^2$$
$$= \sum_{l=0}^{t}\left(1 - \frac{c}{3C^*}\frac{C_l}{C_u}\frac{1}{\bar{\tau}^2 d}\right)^{t-l}\left(1 - c_5\frac{C_l}{C_u}\frac{1}{\bar{\tau}^2 d}\right)^{2l}D_0^2$$

$$\leq \left(1 - \frac{c}{3C^*}\frac{C_l}{C_u}\frac{1}{\bar{\tau}^2 d}\right)^t \frac{1}{\frac{c}{3C^*} - c_5} \cdot \frac{C_u}{C_l}\bar{\tau}^2 d$$

$$\leq \frac{6C^*}{c}\left(1 - \frac{c}{3C^*}\frac{C_l}{C_u}\frac{1}{\bar{\tau}^2 d}\right)^{2t} \frac{C_u}{C_l}\bar{\tau}^2 d,$$

where the second line follows from the inequality $1 - 2a \leq (1-a)^2$ and the condition $c_5 \leq c/(6C^*)$. This further implies

$$2c^2 \frac{C_l}{C_u}\frac{1}{d}\frac{1}{\bar{\tau}^2}\sum_{l=0}^{t}\left(1 - \frac{c}{3C^*}\frac{C_l}{C_u}\frac{1}{\bar{\tau}^2 d}\right)^{t-l} D_l^2 \leq 12cC^*\left(1 - \frac{c}{3C^*}\frac{C_l}{C_u}\frac{1}{\bar{\tau}^2 d}\right)^t D_0^2.$$

To bound the last stochastic term, under the inductive hypothesis $\cap_{l=0}^{t}\{\|\boldsymbol{\beta}_l - \boldsymbol{\beta}^*\|_2^2 \leq C^{*2}D_l^2\}$, we have

$$\|f_l(\boldsymbol{\beta}_l) - f_l(\boldsymbol{\beta}^*) - \mathbb{E}\left[f_l(\boldsymbol{\beta}_l) - f_l(\boldsymbol{\beta}^*)|\boldsymbol{\beta}_l\right]\|_{\Psi_2} \leq C_1 C^*\sqrt{C_u}\bar{\tau}D_l, \quad l = 0, 1, \ldots, t,$$

where $\|\cdot\|_{\Psi_2}$ denotes the Orlicz norm associated with the Orlicz function $\Psi(x) = \exp(x^2) - 1$. Then, by applying a variant of Azuma's inequality (Shamir, 2011), it follows that with probability at least $1 - \exp(-c_1 d)$,

$$\left|\sum_{l=0}^{t}\left(1 - \frac{c}{3}\frac{C_l}{C_u}\frac{1}{\bar{\tau}^2 d}\right)^{t-l} D_l\left[f_l(\boldsymbol{\beta}_l) - f_l(\boldsymbol{\beta}^*) - \mathbb{E}\{f_l(\boldsymbol{\beta}_l) - f_l(\boldsymbol{\beta}^*)|\boldsymbol{\beta}_l\}\right]\right|$$

$$\leq \bar{\tau}C^*\sqrt{C_u d\sum_{l=0}^{t}\left(1 - \frac{c}{3C^*}\frac{C_l}{C_u}\frac{1}{\bar{\tau}^2 d}\right)^{2t-2l} D_l^4}$$

$$= \bar{\tau}C^*\sqrt{C_u d\sum_{l=0}^{t}\left(1 - \frac{c}{3C^*}\frac{C_l}{C_u}\frac{1}{\bar{\tau}^2 d}\right)^{2t-2l}\left(1 - c_5\frac{C_l}{C_u}\frac{1}{\bar{\tau}^2 d}\right)^{4l}} \times D_0^2$$

$$\leq \bar{\tau}C^*\sqrt{C_u d\left(1 - \frac{c}{3C^*}\frac{C_l}{C_u}\frac{1}{\bar{\tau}^2 d}\right)^{2t}\cdot \frac{6C^*}{c}\frac{C_u}{C_l}\bar{\tau}^2 d} \times D_0^2$$

$$\leq \bar{\tau}^2 d \times \left(1 - c_5\frac{C_l}{C_u}\frac{1}{\bar{\tau}^2 d}\right)^{2t} D_0^2 \times \sqrt{\frac{C_u^2}{C_l}}\sqrt{\frac{C^{*3}}{c}},$$

where the last two inequalities follow from the bound $1 - 2a \leq (1-a)^2$ for $0 < a < 1$, together with the condition $c_5 \leq c/(6C^*)$.

Putting together the pieces, we conclude that for any C^* satisfying $c \leq C^* \leq c/(6c_5)$, it holds

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2$$

$$\leq \left(1 - \frac{c}{3C^*}\frac{C_l}{C_u}\frac{1}{\bar{\tau}^2 d}\right)^{t+1}\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2^2 + 12cC^*\left(1 - \frac{c}{3C^*}\frac{C_l}{C_u}\frac{1}{\bar{\tau}^2 d}\right)^t D_0^2 + \sqrt{c}C^{*3/2}D_{t+1}^2$$

$$\leq \frac{1}{3}D_{t+1}^2 + 12cC^*\left(1 - c_5\frac{C_l}{C_u}\frac{1}{\bar{\tau}^2 d}\right)^{2t} D_0^2 \exp\left(t\log\left(\frac{1 - \frac{c}{3C^*}\frac{C_l}{C_u}\frac{1}{\bar{\tau}^2 d}}{(1 - c_5\frac{C_l}{C_u}\frac{1}{\bar{\tau}^2 d})^2}\right)\right) + \sqrt{c}C^{*3/2}D_{t+1}^2$$

$$\leq C^{*2}D_{t+1}^2.$$

This completes the proof of convergence in phase one.

SECOND PHASE ANALYSIS. The verification of the second phase is also carried out by induction. The proof for the step $t_1 + 1$ is omitted due to its simplicity, as it follows analogously from the subsequent argument. We assume $\cap_{l=t_1}^{t}\{\|\boldsymbol{\beta}_l - \boldsymbol{\beta}^*\|_2^2 \leq \frac{C^*d}{l-t_1+C_bd}\frac{C_u}{C_l^2}b_0^2\}$ and proceed to analyze the behavior at iteration $t+1$. Within this second phase, Lemma 17 establishes a quadratic lower bound on the expected excess loss, that is,

$$\mathbb{E}\left[f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*)|\boldsymbol{\beta}_t\right] \geq \frac{C_l}{12b_0}\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2.$$

Subsequently, the right-hand side of equation (5) can be upper bounded by

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2 \leq \|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2 - \frac{1}{6b_0}\eta_t C_l\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2 + 2\eta_t^2 C_u\bar{\tau}^2 d$$
$$- 2\eta_t\left[f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*) - \mathbb{E}\left[f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*)|\boldsymbol{\beta}_t\right]\right].$$

Incorporating the chosen stepsize $\eta_t = \frac{C_a}{t-t_1+C_bd}\frac{b_0}{C_l}$ preceding equation, we obtain

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2 \leq \left(1 - \frac{C_a}{12(t - t_1 + C_bd)}\right)\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2 + 2b_0^2\bar{\tau}^2\frac{C_u}{C_l^2}\frac{C_a^2}{(t - t_1 + C_bd)^2}d$$
$$- 2\frac{C_a}{t - t_1 + C_bd}\frac{b_0}{C_l}\left[f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*) - \mathbb{E}\left[f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*)|\boldsymbol{\beta}_t\right]\right].$$

Summing the above inequality over t from t_1 onward, we arrive at

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2$$
$$\leq \prod_{l=t_1}^{t}\left(1 - \frac{C_a}{12(l - t_1 + C_bd)}\right)\|\boldsymbol{\beta}_{t_1} - \boldsymbol{\beta}^*\|_2^2$$
$$+ 2b_0^2\bar{\tau}^2\frac{C_u}{C_l^2}d\sum_{l=t_1}^{t}\left(1 - \frac{C_a}{12(l + 1 - t_1 + C_bd)}\right)\cdots\left(1 - \frac{C_a}{12(t - t_1 + C_bd)}\right)\frac{C_a^2}{(l - t_1 + C_bd)^2}$$
$$- 2\frac{b_0}{C_l}\sum_{l=t_1}^{t}\left(1 - \frac{C_a}{12(l + 1 - t_1 + C_bd)}\right)\cdots\left(1 - \frac{C_a}{12(t - t_1 + C_bd)}\right)\frac{C_a}{l - t_1 + C_bd}$$
$$\times \left[f_l(\boldsymbol{\beta}_l) - f_l(\boldsymbol{\beta}^*) - \mathbb{E}\left[f_l(\boldsymbol{\beta}_l) - f_l(\boldsymbol{\beta}^*)|\boldsymbol{\beta}_l\right]\right].$$
$$(6)$$

A sharp bound on the above equation constitutes a key step in the proof. We proceed by analyzing each of the three terms on the right-hand side separately. Note that when $C_a > 12$, we have

$$\left(1 - \frac{C_a}{12(l - t_1 + C_bd)}\right) \leq \frac{l - t_1 + C_bd}{l + 1 - t_1 + C_bd}.$$

Then, the first term on the right-hand side of (6) satisfies

$$\prod_{l=t_1}^{t}\left(1 - \frac{C_a}{12(l - t_1 + C_bd)}\right)\|\boldsymbol{\beta}_{t_1} - \boldsymbol{\beta}^*\|_2^2 \leq \frac{C_bd}{t + 1 - t_1 + C_bd}\|\boldsymbol{\beta}_{t_1} - \boldsymbol{\beta}^*\|_2^2.$$

Consider the product sequence, which can be bounded as

$$
\left(1 - \frac{C_a}{12(l+1-t_1+C_bd)}\right) \cdots \left(1 - \frac{C_a}{12(t-t_1+C_bd)}\right)
$$

$$
= \exp\left(\sum_{k=l+1}^{t} \log\left(1 - \frac{C_a}{12(k-t_1+C_bd)}\right)\right)
$$

$$
\leq \exp\left(-\sum_{k=l+1}^{t} \frac{C_a}{12(k-t_1+C_bd)}\right) \leq \exp\left(-\frac{C_a}{12}\int_{l+1-t_1}^{t+1-t_1} \frac{1}{x+C_bd}\, dx\right)
$$

$$
\leq \exp\left(-\frac{C_a}{12}\log\left(\frac{t+1-t_1+C_bd}{l+1-t_1+C_bd}\right)\right) = \left(\frac{l+1-t_1+C_bd}{t+1-t_1+C_bd}\right)^{\frac{C_a}{12}},
$$

(7)

where the first inequality uses the bound $\log(1+x) \leq x$, and the third line leverages the relationship between integrals and discrete sums. Consequently, the second term in equation (6) can be bounded as

$$
\sum_{l=t_1}^{t}\left(1 - \frac{C_a}{12(l+1-t_1+C_bd)}\right)\cdots\left(1 - \frac{C_a}{12(t-t_1+C_bd)}\right)\frac{C_a^2}{(l-t_1+C_bd)^2}
$$

$$
\leq \sum_{l=t_1}^{t}\left(\frac{l+1-t_1+C_bd}{t+1-t_1+C_bd}\right)^{\frac{C_a}{12}}\frac{C_a^2}{(l-t_1+C_bd)^2}
$$

$$
\leq \left(\frac{1+C_bd}{C_bd}\right)^{\frac{C_a}{12}}\frac{C_a^2}{(t+1-t_1+C_bd)^{\frac{C_a}{12}}}\sum_{l=t_1}^{t}(l-t_1+C_bd)^{\frac{C_a}{12}-2}.
$$

With $C_a > 12$, we have

$$
\sum_{l=t_1}^{t}(l-t_1+C_bd)^{\frac{C_a}{12}-2} \leq \int_{0}^{t-t_1+1}(x+C_bd)^{\frac{C_a}{12}-2}\, dx = \frac{1}{\frac{C_a}{12}-1}(t-t_1+1+C_bd)^{\frac{C_a}{12}-1}.
$$

Combining these results, we obtain the following upper bound for the second term in equation (6):

$$
\sum_{l=t_1}^{t}\left(1 - \frac{C_a}{12(l+1-t_1+C_bd)}\right)\cdots\left(1 - \frac{C_a}{12(t-t_1+C_bd)}\right)\frac{C_a^2}{(l-t_1+C_bd)^2}
$$

$$
\leq \left(\frac{1+C_bd}{C_bd}\right)^{\frac{C_a}{12}}\frac{C_a^2}{t+1-t_1+C_bd}.
$$

Assuming $\cap_{l=t_1}^{t}\{\|\boldsymbol{\beta}_l - \boldsymbol{\beta}^*\|_2^2 \leq \frac{C^*d}{l-t_1+C_bd}\frac{C_u}{C_l^2}b_0^2\}$, it follows that

$$
\|f_l(\boldsymbol{\beta}_l) - f_l(\boldsymbol{\beta}^*) - \mathbb{E}\left[f_l(\boldsymbol{\beta}_l) - f_l(\boldsymbol{\beta}^*)|\boldsymbol{\beta}_l\right]\|_{\Psi_2} \leq C\bar{\tau}\sqrt{\frac{C^*d}{l-t_1+C_bd}\frac{C_u^2}{C_l^2}}b_0.
$$

By Theorem 2 in Shamir (2011) and equation (7), with probability exceeding $1 - c\exp(-cd)$, the third term in equation (6) can be bounded by

$$
\frac{b_0}{C_l}\sum_{l=t_1}^{t}\left(1 - \frac{C_a}{12(l+1-t_1+C_bd)}\right)\cdots\left(1 - \frac{C_a}{12(t-t_1+C_bd)}\right)\frac{C_a}{l-t_1+C_bd}
$$

$$\times \left[f_l(\boldsymbol{\beta}_l) - f_l(\boldsymbol{\beta}^*) - \mathbb{E}\left[f_l(\boldsymbol{\beta}_l) - f_l(\boldsymbol{\beta}^*) \middle| \boldsymbol{\beta}_l \right] \right]$$

$$\leq \bar{\tau} C_a \frac{C_u}{C_l^2} \sqrt{\sum_{l=t_1}^{t} \left(\frac{l+1-t_1+C_b d}{t+1-t_1+C_b d} \right)^{\frac{C_a}{6}} \frac{1}{(l-t_1+C_b d)^2} \frac{C^* d^2}{l-t_1+C_b d} b_0^2}$$

$$= \bar{\tau} C_a \frac{C_u}{C_l^2} b_0^2 \frac{\sqrt{C^*} d}{(t+1-t_1+C_b d)^{\frac{C_a}{12}}} \sqrt{\left(\frac{1+C_b d}{C_b d} \right)^{\frac{C_a}{6}} \sum_{l=t_1}^{t} (l-t_1+C_b d)^{\frac{C_a}{6}-3}}$$

$$\leq \bar{\tau} C_a \frac{C_u}{C_l^2} b_0^2 \frac{\sqrt{C^*} d}{t+1-t_1+C_b d} \times \left(\frac{1+C_b d}{C_b d} \right)^{\frac{C_a}{12}}.$$

Finally, combining the upper bounds of the three terms, we obtain

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2 \leq \frac{C_b d}{t+1-t_1+C_b d} \|\boldsymbol{\beta}_{t_1} - \boldsymbol{\beta}^*\|_2^2 + 2 \left(\frac{1+C_b d}{C_b d} \right)^{\frac{C_a}{12}} \frac{C_a^2}{t+1-t_1+C_b d}$$

$$+ C_a \frac{C_u}{C_l^2} b_0^2 \frac{\sqrt{C^*} d}{t+1-t_1+C_b d} \times \left(\frac{1+C_b d}{C_b d} \right)^{\frac{C_a}{12}}$$

$$\leq C^* \frac{C_u}{C_l^2} b_0^2 \frac{d}{t+1-t_1+C_b d},$$

where $\{(1+C_b d)/(C_b d)\}^{\frac{C_a}{12}} < 2^{C_a/12}$ is independent of d, and C^* is a sufficiently large constant that may depend on C_a, C_b. This completes the proof of Theorem 1.

5. Discussions

This paper focuses on online quantile regression in low-dimensional settings. The analytical framework developed herein can also be extended to the study of stochastic sub-gradient methods for low-rank regression under quantile loss. Let $\mathbf{M}^* \in \mathbb{R}^{d_1 \times d_2}$ denote the true low-rank matrix with $\text{rank}(\mathbf{M}^*) = r$. At each time step t, we observe data $(Y_t, \mathbf{X}_t)$, where $Y_t = \langle \mathbf{X}_t, \mathbf{M}^* \rangle + \xi_t$, and $\mathbf{X}_t \in \mathbb{R}^{d_1 \times d_2}$ is the sensing matrix. The corresponding loss function at time t is defined as $f_t(\mathbf{M}) = \rho_{Q,\tau}(Y_t - \langle \mathbf{X}_t, \mathbf{M} \rangle)$. A central challenge in this setting is to maintain the low-rank structure of the iterates over time. To this end, the online Riemannian optimization framework updates the estimate according to

$$\mathbf{M}_{t+1} = \text{SVD}_r(\mathbf{M}_t - \eta_t \mathcal{P}_{\mathbb{T}_t}(\mathbf{G}_t)),$$

where $\mathbf{G}_t \in \partial f_t(\mathbf{M}_t)$ is the sub-gradient, and $\mathcal{P}_{\mathbb{T}_t}(\cdot)$ denotes the projection onto the tangent space at $\mathbf{M}_t$. Detailed expressions for $\mathcal{P}_{\mathbb{T}_t}(\mathbf{G}_t)$ and further discussion of Riemannian optimization techniques can be found in Vandereycken (2013) and Mishra et al. (2014). Each iteration retains only the leading r singular components, raising an important conceptual question: to what extent does this truncation preserve essential statistical information? A comprehensive theoretical investigation of online low-rank regression under quantile loss is beyond the scope of this paper and is left for future research.

Acknowledgments

References

Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. *Advances in Neural Information Processing Systems*, 26, 2013.

Hamsa Bastani and Mohsen Bayati. Online decision making with high-dimensional covariates. *Operations Research*, 68(1):276–294, 2020.

Alexandre Belloni and Victor Chernozhukov. ℓ_1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82–130, 2011.

Léon Bottou. Online learning and stochastic approximations. In David Saad, editor, *On-line Learning in Neural Networks*, pages 9–42. Cambridge University Press, 1999.

Jian-Feng Cai, Jingyang Li, and Dong Xia. Online tensor learning: Computational and statistical trade-offs, adaptivity and optimal regret. *arXiv preprint arXiv:2306.03372*, 2023.

Léopold Cambier and P.-A. Absil. Robust low-rank matrix completion by riemannian optimization. *SIAM Journal on Scientific Computing*, 38(5):S440–S460, 2016.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Vasileios Charisopoulos, Yudong Chen, Damek Davis, Mateo Diaz, Lijun Ding, and Dmitriy Drusvyatskiy. Low-rank matrix recovery with composite optimization: good conditioning and rapid convergence. *Foundations of Computational Mathematics*, 21(6):1505–1593, 2021.

Lanjue Chen and Yong Zhou. Quantile regression in big data: A divide and conquer based strategy. *Computational Statistics & Data Analysis*, 144:106892, 2020.

Xi Chen, Weidong Liu, and Yichen Zhang. Quantile regression under memory constraint. *The Annals of Statistics*, 47(6):3244–3273, 2019.

Xuerong Chen and Senlin Yuan. Renewable quantile regression with heterogeneous streaming datasets. *Journal of Computational and Graphical Statistics*, 33(4):1185–1201, 2024.

Bernard Delyon and Anatoli Juditsky. Accelerated stochastic approximation. *SIAM Journal on Optimization*, 3(4):868–881, 1993.

Chuong B Do, Quoc V Le, and Chuan-Sheng Foo. Proximal regularization for online and batch learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 257–264, 2009.

John Duchi and Yoram Singer. Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research*, 10(99):2899–2934, 2009.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(7):2121–2159, 2011.

Jianqing Fan, Wenyan Gong, Chris Junchi Li, and Qiang Sun. Statistical sparse online regression: A diffusion approximation perspective. In *International Conference on Artificial Intelligence and Statistics*, pages 1017–1026. PMLR, 2018.

Marcelo Fernandes, Emmanuel Guerre, and Eduardo Horta. Smoothing quantile regressions. *Journal of Business and Economic Statistics*, 39(1):338–357, 2021.

Chelsea Finn, Aravind Rajeswaran, Sham Kakade, and Sergey Levine. Online meta-learning. In *International Conference on Machine Learning*, pages 1920–1930. PMLR, 2019.

Zijun Gao, Yanjun Han, Zhimei Ren, and Zhengqing Zhou. Batched multi-armed bandits problem. *Advances in Neural Information Processing Systems*, 32, 2019.

Alexander Goldenshluger and Assaf Zeevi. A linear response bandit problem. *Stochastic Systems*, 3(1):230–261, 2013.

Qiyu Han, Will Wei Sun, and Yichen Zhang. Online statistical inference for matrix contextual bandit. *arXiv preprint arXiv:2212.11385*, 2022.

Ruijian Han, Lan Luo, Yuanyuan Lin, and Jian Huang. Online inference with debiased stochastic gradient descent. *Biometrika*, 111(1):93–108, 2024.

Yanjun Han, Zhengqing Zhou, Zhengyuan Zhou, Jose Blanchet, Peter W Glynn, and Yinyu Ye. Sequential batch learning in finite-action linear contextual bandits. *arXiv preprint arXiv:2004.06321*, 2020.

Elad Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69:169–192, 2007a.

Elad Hazan, Alexander Rakhlin, and Peter Bartlett. Adaptive online gradient descent. *Advances in Neural Information Processing Systems*, 20, 2007b.

Xuming He and Qi-Man Shao. On parameters of increasing dimensions. *Journal of Multivariate Analysis*, 73(1):120–135, 2000.

Xuming He, Xiaoou Pan, Kean Ming Tan, and Wen-Xin Zhou. Smoothed quantile regression with large-scale inference. *Journal of Econometrics*, 232(2):367–388, 2023.

Matthew Hoffman, Francis Bach, and David Blei. Online learning for latent dirichlet allocation. *Advances in Neural Information Processing Systems*, 23, 2010.

Rong Jiang and Keming Yu. Renewable quantile regression for streaming data sets. *Neurocomputing*, 508:208–224, 2022.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in Neural Information Processing Systems*, 26, 2013.

Roger Koenker. *Quantile Regression*. Cambridge University Press, 2005.

Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica*, 46(1):33–50, 1978.

Roger Koenker, Victor Chernozhukov, Xuming He, and Liming Peng. *Handbook of Quantile Regression*. CRC Press.

Roger W Koenker and Vasco d'Orey. Algorithm as 229: Computing regression quantiles. *Applied Statistics*, pages 383–393, 1987.

John Langford, Lihong Li, and Tong Zhang. Sparse online learning via truncated gradient. *Journal of Machine Learning Research*, 10(3):777–801, 2009.

Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989.

Xiao Li, Zhihui Zhu, Anthony Man-Cho So, and Rene Vidal. Nonconvex robust low-rank matrix recovery. *SIAM Journal on Optimization*, 30(1):660–686, 2020.

Jianhao Ma and Salar Fattahi. Global convergence of sub-gradient method for robust matrix recovery: Small initialization, noisy measurements, and over-parameterization. *Journal of Machine Learning Research*, 24(96):1–84, 2023.

H Brendan McMahan and Matthew Streeter. Adaptive bound optimization for online convex optimization. In *COLT*, 2010.

Zakaria Mhammedi, Wouter M Koolen, and Tim Van Erven. Lipschitz adaptivity with multiple learning rates in online learning. In *Conference on Learning Theory*, pages 2490–2511. PMLR, 2019.

Bamdev Mishra, Gilles Meyer, Silvere Bonnabel, and Rodolphe Sepulchre. Fixed-rank matrix factorizations and riemannian low-rank optimization. *Computational Statistics*, 29:591–621, 2014.

Nuno Moniz and Luis Torgo. Multi-source social feedback of online news feeds. *arXiv preprint arXiv:1801.07055*, 2018.

Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.

Stephen Portnoy and Roger Koenker. The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error estimators. *Statistical Science*, 12(4):279–300, 1997.

Nikita Puchkin, Eduard Gorbunov, Nickolay Kutuzov, and Alexander Gasnikov. Breaking the heavy-tailed noise barrier in stochastic optimization problems. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238, pages 856–864. PMLR, 2024.

M Rajalakshmi, P Saranya, and P Shanmugavadivu. Pattern recognition-recognition of handwritten document using convolutional neural networks. In *2019 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)*, pages 1–7. IEEE, 2019.

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. *arXiv preprint arXiv:1109.5647*, 2011.

Zhimei Ren and Zhengyuan Zhou. Dynamic batch learning in high-dimensional sparse linear contextual bandits. *Management Science*, 70(2), 2023.

Nicolas Roux, Mark Schmidt, and Francis Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. *Advances in Neural Information Processing Systems*, 25, 2012.

Leonard J Savage. The theory of statistical decision. *Journal of the American Statistical Association*, 46(253):55–67, 1951.

Ohad Shamir. A variant of Azuma's inequality for martingales with subgaussian tails. *arXiv preprint arXiv:1110.2392*, 2011.

Yinan Shen, Jingyang Li, Jian-Feng Cai, and Dong Xia. Computationally efficient and statistically optimal robust high-dimensional linear regression. *arXiv preprint arXiv:2305.06199*, 2023.

Farzan Soleymani and Eric Paquet. Financial portfolio optimization with online deep reinforcement learning and restricted stacked autoencoder—deepbreath. *Expert Systems with Applications*, 156: 113456, 2020.

Matthew Streeter and H Brendan McMahan. Less regret via online conditioning. *arXiv preprint arXiv:1002.4862*, 2010.

Xiaofei Sun, Hongwei Wang, Chao Cai, Mei Yao, and Kangning Wang. Online renewable smooth quantile regression. *Computational Statistics & Data Analysis*, 185:107781, 2023.

Tian Tong, Cong Ma, and Yuejie Chi. Low-rank matrix recovery with scaled subgradient methods: Fast and robust convergence without the condition number. *IEEE Transactions on Signal Processing*, 69:2396–2409, 2021.

Bart Vandereycken. Low-rank matrix completion by riemannian optimization. *SIAM Journal on Optimization*, 23(2):1214–1236, 2013.

Roman Vershynin. *High-dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Kangning Wang, Hongwei Wang, and Shaomin Li. Renewable quantile regression for streaming datasets. *Knowledge-Based Systems*, 235:107675, 2022.

Dong Xia and Ming Yuan. Statistical inferences of linear forms for noisy matrix completion. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(1):58–77, 2021.

Tong Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the 21st International Conference on Machine Learning*, page 116, 2004.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning*, pages 928–936, 2003.

Changliang Zou, Guanghui Wang, and Runze Li. Consistent selection of the number of change-points via sample-splitting. *The Annals of Statistics*, 48(1):413, 2020.

Supplementary Material to "Online Quantile Regression"

Appendix A. Proof of Main Results

This section presents proofs of the main results for each setting.

A.1 Proof of Online Learning

In this subsection, we shall first prove Remark 3, and then prove the regret bound, which is a subsequent result of the expected estimation error rates.

A.1.1 PROOF OF REMARK 3: ONLINE LEARNING WITH GOOD INITIALIZATION

We are going to prove the convergence dynamics when $\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2 < 8\sqrt{C_l^{-1}\gamma}$. According to proof of Theorem 1, we have

$$
\begin{aligned}
\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2 &\leq \prod_{l=0}^{t}\left(1 - \frac{C_a}{12(l + C_b d)}\right)\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2^2 \\
&+ 2b_0^2\bar{\tau}^2\frac{C_u}{C_l^2}d\sum_{l=0}^{t}\left(1 - \frac{C_a}{12(l+1 + C_b d)}\right)\cdots\left(1 - \frac{C_a}{12(t + C_b d)}\right)\frac{C_a^2}{(l + C_b d)^2} \\
&- 2\frac{b_0}{C_l}\sum_{l=0}^{t}\left(1 - \frac{C_a}{12(l+1 + C_b d)}\right)\cdots\left(1 - \frac{C_a}{12(t + C_b d)}\right)\frac{C_a}{l + C_b d} \\
&\times \left[f_l(\boldsymbol{\beta}_l) - f_l(\boldsymbol{\beta}^*) - \mathbb{E}\left[f_l(\boldsymbol{\beta}_l) - f_l(\boldsymbol{\beta}^*)\big|\boldsymbol{\beta}_l\right]\right],
\end{aligned} \tag{8}
$$

also, the following upper bound of the product series holds

$$
\left(1 - \frac{C_a}{12(l + C_b d)}\right)\cdots\left(1 - \frac{C_a}{12(t + C_b d)}\right) \leq \left(\frac{l+1 + C_b d}{t+1 + C_b d}\right)^{\frac{C_a}{12}}.
$$

<u>Case One: large C_a</u> If $C_a > 12$, the analyses are exactly the ones in Theorem 1 and the upper bound is

$$
\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2 \leq C^*\frac{C_u}{C_l^2}\frac{d}{t+1 - t_1 + C_b d}b_0^2.
$$

<u>Case One: small C_a</u> If $C_a < 12$, the first term of equation (8) is bounded with

$$
\prod_{l=0}^{t}\left(1 - \frac{C_a}{12(l + C_b d)}\right)\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2^2 \leq \left(\frac{1 + C_b d}{t+1 + C_b}\right)^{\frac{C_a}{12}}.
$$

In addition, the second term of equation (8) has the following upper bound,

$$
b_0^2\bar{\tau}^2\frac{C_u}{C_l^2}d\sum_{l=0}^{t}\left(1 - \frac{C_a}{12(l+1 + C_b d)}\right)\cdots\left(1 - \frac{C_a}{12(t + C_b d)}\right)\frac{C_a^2}{(l + C_b d)^2}
$$

$$\leq b_0^2 \bar{\tau}^2 \frac{C_u}{C_l^2} d \sum_{l=0}^{t} \left(\frac{l+1+C_b d}{t+1+C_b d} \right)^{\frac{C_a}{12}} \frac{C_a^2}{(l+C_b d)^2}$$

$$\leq b_0^2 \bar{\tau}^2 \frac{C_u}{C_l^2} d \frac{C_a^2}{(t+1+C_b d)^{\frac{C_a}{12}}} \sum_{l=0}^{t} \frac{1}{(l+1+C_b d)^{2-\frac{C_a}{12}}}$$

$$\leq 4 b_0^2 \bar{\tau}^2 \frac{C_u}{C_l^2} \frac{d^{\frac{C_a}{12}} C_a^2}{(t+1+C_b d)^{\frac{C_a}{12}}},$$

where the last line uses $C_b \geq 1$ and

$$\sum_{l=0}^{t} \frac{1}{(l+1+C_b d)^{2-\frac{C_a}{12}}} \leq \int_0^{t+1} \frac{1}{(x+C_b d)^{2-\frac{C_a}{12}}} \, dx \leq \frac{1}{1-C_a/12} \frac{1}{(C_b d)^{1-\frac{C_a}{12}}}.$$

It is worth noting that, under the event of $\cap_{l=0}^{t} \{ \| \boldsymbol{\beta}_l - \boldsymbol{\beta}^* \|_2^2 \leq C^* (\frac{d}{1+C_b d})^{\frac{C_a}{12}} \| \boldsymbol{\beta}_0 - \boldsymbol{\beta}^* \|_2^2 \}$, it has

$$\| f_l(\boldsymbol{\beta}_l) - f_l(\boldsymbol{\beta}^*) - \mathbb{E} \left[f_l(\boldsymbol{\beta}_l) - f_l(\boldsymbol{\beta}^*) | \boldsymbol{\beta}_l \right] \|_{\psi_2} \leq \bar{\tau} \sqrt{ C_u C^* \left(\frac{d}{l+C_b d} \right)^{\frac{C_a}{12}} \| \boldsymbol{\beta}_0 - \boldsymbol{\beta}^* \|_2^2 }.$$

Thus according to Theorem 2 in Shamir (2011), the upper bound of the last term for equation (8) is obtained with probability exceeding $1 - \exp(-cd)$,

$$\frac{b_0}{C_l} \sum_{l=0}^{t} \left(1 - \frac{C_a}{12(l+1+C_b d)} \right) \cdots \left(1 - \frac{C_a}{12(t+C_b d)} \right) \frac{C_a}{l+C_b d}$$

$$\times \left[f_l(\boldsymbol{\beta}_l) - f_l(\boldsymbol{\beta}^*) - \mathbb{E} \left[f_l(\boldsymbol{\beta}_l) - f_l(\boldsymbol{\beta}^*) | \boldsymbol{\beta}_l \right] \right]$$

$$\leq \bar{\tau} C_a \frac{b_0}{C_l} \sqrt{ C^* C_u d \sum_{l=0}^{t} \left(\frac{l+1+C_b d}{t+1+C_b d} \right)^{\frac{C_a}{6}} \frac{1}{(l+C_b d)^2} \left(\frac{d}{l+C_b d} \right)^{\frac{C_a}{12}} \| \boldsymbol{\beta}_0 - \boldsymbol{\beta}^* \|_2^2 }$$

$$\leq \bar{\tau} \sqrt{C^*} \sqrt{C_u} \frac{1}{C_l} \frac{d^{\frac{C_a}{12}}}{(t+1+C_b d)^{\frac{C_a}{12}}} \cdot C_a b_0 \| \boldsymbol{\beta}_0 - \boldsymbol{\beta}^* \|_2.$$

Thus altogether, by having $C_a < \min\{12, (C_l/\sqrt{C_u}) \| \boldsymbol{\beta}_0 - \boldsymbol{\beta}^* \|_2 / b_0 \}$, we obtain the following bound for equation (8) with some sufficient large constant C^*, where C^* may depend on the constants C_a, C_b,

$$\| \boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^* \|_2^2 \leq C^* \left(\frac{d}{t+1+C_b d} \right)^{\frac{C_a}{12}} \| \boldsymbol{\beta}_0 - \boldsymbol{\beta}^* \|_2^2.$$

A.1.2 PROOF OF THEOREM 4

Before proving the regret bound, we first prove the following lemma, which establishes the expected error rate dynamics. It is worth noting that it is not a consequence of the empirical dynamics Theorem 1.

Lemma 10 *Assume the same conditions and stepsizes as Theorem 1. Then in the first phase we have* $\sum_{t \leq t_1} \mathbb{E} \| \boldsymbol{\beta}_t - \boldsymbol{\beta}^* \|_2 \leq C \max\{\tau^2, (1-\tau)^2\} d \frac{C_u}{C_l} D_0 + C t_1 \exp(-cd) D_0$ *and in the second phase we have* $\mathbb{E} \| \boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^* \|_2^2 \leq C \frac{C_u}{C_l^2} \frac{d}{t+1-t_1+C_b d} b_0^2.$

Proof Note that the expectation of $\mathbb{E}\left[\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2\right]$ is

$$
\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2\right] &= \mathbb{E}\left[\mathbb{E}\left[\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2\big|\boldsymbol{\beta}_t\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2 - 2\eta_t\langle\boldsymbol{\beta}_t - \boldsymbol{\beta}^*, \mathbf{g}_t\rangle + \eta_t^2\|\mathbf{g}_t\|_2^2\big|\boldsymbol{\beta}_t\right]\right].
\end{aligned}
$$

First, we would consider the inner conditional expectation. According to the sub-gradient definition, and the independence between η_t and $(\mathbf{X}_t, Y_t)$, we have

$$
\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2 - 2\eta_t\langle\boldsymbol{\beta}_t - \boldsymbol{\beta}^*, \mathbf{g}_t\rangle + \eta_t^2\|\mathbf{g}_t\|_2^2\big|\boldsymbol{\beta}_t\right] \\
\le \|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2 - 2\eta_t\mathbb{E}\left[f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*)\big|\boldsymbol{\beta}_t\right] + \eta_t^2\mathbb{E}\left[\|\mathbf{g}_t\|_2^2\big|\boldsymbol{\beta}_t\right].
\end{aligned}
$$

FIRST PHASE ANALYSIS. We inherit the notations in Section 4. Denote the event $\mathcal{E}_t := \{\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2 \le CD_t\}$ and Theorem 1 shows $\mathbb{P}(\cup_{t=0,\dots,t_1}\mathcal{E}_t^c) \le t_1\exp(-c_0 d)$. Then we have

$$
\begin{aligned}
\mathbb{E}\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2 &= \mathbb{E}\left\{\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2 \cdot \mathbb{I}\{\mathcal{E}_{t+1}\}\right\} + \mathbb{E}\left\{\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2 \cdot \mathbb{I}\{\mathcal{E}_{t+1}^c\}\right\} \\
&\le CD_{t+1} + \mathbb{P}\left(\mathcal{E}_{t+1}^c\right) \cdot \left(\mathbb{E}\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2\right)^{\frac{1}{2}}.
\end{aligned}
\tag{9}
$$

Hence we only need to provide a rough upper bound for $\mathbb{E}\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|^2$. Note that $\mathbb{E}[f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*)\big|\boldsymbol{\beta}_t] \ge 0$ and therefore, we have

$$
\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2\right] &\le \mathbb{E}\left[\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|^2\right] + \eta_t^2 \cdot \max\{\tau, 1-\tau\}^2 C_u d \\
&\le \mathbb{E}\left[\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|^2\right] + \frac{C}{\max\{\tau, 1-\tau\}^2}\frac{C_l}{C_u}\frac{1}{d}D_t^2,
\end{aligned}
$$

which implies

$$
\mathbb{E}\left[\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2\right] \le \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2^2 + \frac{C}{\max\{\tau, 1-\tau\}^2}\frac{C_l}{C_u}\frac{1}{d}\sum_{l=0}^{t}D_l^2 \le CD_0^2.
$$

Inserting the inequality above into equation (9), we obtain

$$
\mathbb{E}\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2 \le CD_{t+1} + \mathbb{P}\left(\mathcal{E}_{t+1}^c\right)D_0.
$$

Thus we have

$$
\begin{aligned}
\sum_{t=0}^{t_1}\mathbb{E}\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2 &\le \sum_{t=0}^{t_1}D_t + C\mathbb{P}(\cup_{t=0,\dots,t_1}\mathcal{E}_t^c)D_0 \\
&\le C\max\{\tau^2, (1-\tau)^2\}d\frac{C_u}{C_l}D_0 + Ct_1\exp(-cd)D_0.
\end{aligned}
$$

SECOND PHASE ANALYSIS. Lemma 17 shows that in the second phase, the lower bound of the expected excess risk is a second-order one, $\mathbb{E}[f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*)|\boldsymbol{\beta}_t] \ge \frac{1}{12b_0}C_l\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2$. Thus we have

$$
\mathbb{E}\left[\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2\big|\boldsymbol{\beta}_t\right] \le \|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2 - \frac{1}{6b_0}\eta_t C_l\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2 + \eta_t^2 C_u\max\{\tau^2, (1-\tau)^2\}d.
$$

By specifying the stepsize $\eta_t = \frac{C_a}{t-t_1+C_b d}\frac{b_0}{C_l}$, we obtain

$$\mathbb{E}\left[\|\boldsymbol{\beta}_{t+1}-\boldsymbol{\beta}^*\|_2^2\big|\boldsymbol{\beta}_t\right] \leq \left(1 - \frac{C_a}{6(t-t_1+C_b d)}\right)\|\boldsymbol{\beta}_t-\boldsymbol{\beta}^*\|_2^2 + \frac{C_u}{C_l^2}\frac{C_a^2}{(t-t_1+C_b d)^2}db_0^2.$$

Taking the expectation over $\boldsymbol{\beta}_t$ on both sides of the equation and substituting the upper bound at time t, $\mathbb{E}[\|\boldsymbol{\beta}_t-\boldsymbol{\beta}^*\|_2^2] \leq C\frac{C_u}{C_l^2}\frac{C_a d}{t-t_1+C_b d}b_0^2$, into the above equation, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\|\boldsymbol{\beta}_{t+1}-\boldsymbol{\beta}^*\|_2^2\right] \\
&\leq C\left(1-\frac{C_a}{6(t-t_1+C_b d)}\right)\frac{C_u}{C_l^2}\frac{C_a d}{t-t_1+C_b d}b_0^2 + \frac{C_u}{C_l^2}\frac{C_a^2}{(t-t_1+C_b d)^2}db_0^2 \\
&= C\left(1-\frac{C_a/2}{6(t-t_1+C_b d)}\right)\frac{C_u}{C_l^2}\frac{C_a d}{t-t_1+C_b d}b_0^2.
\end{aligned}$$

Note that $1 - \frac{C_a/2}{6(t-t_1+C_b d)} \leq 1 - \frac{1}{t-t_1+C_b d} \leq \frac{t-t_1+C_b d}{t+1-t_1+C_b d}$, and hence we finally obtain $\mathbb{E}[\|\boldsymbol{\beta}_{t+1}-\boldsymbol{\beta}^*\|_2^2] \leq C\frac{C_u}{C_l^2}\frac{C_a d}{t+1-t_1+C_b d}b_0^2$, which completes the proof for the second phase. ∎

We are now ready to establish the regret bound stated in Theorem 4. By the expected excess risk bound in Lemma 17, we have $\boldsymbol{\beta}^* = \arg\max_{\boldsymbol{\beta}} \mathbb{E}\sum_{t=0}^T f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta})$. In the first phase, it has

$$\begin{aligned}
&\sum_{t=0}^{t_1}\mathbb{E}\left[f_t(\boldsymbol{\beta}_t)-f_t(\boldsymbol{\beta}^*)\right] \\
&= \sum_{t=0}^{t_1}\mathbb{E}\left[\mathbb{E}\left[f_t(\boldsymbol{\beta}_t)-f_t(\boldsymbol{\beta}^*)\big|\boldsymbol{\beta}_t\right]\right] \leq \max\{\tau,1-\tau\}\sqrt{C_u}\sum_{t=0}^{t_1}\mathbb{E}\|\boldsymbol{\beta}_t-\boldsymbol{\beta}^*\|_2 \\
&\leq C\max\{\tau^3,(1-\tau)^3\}\sqrt{C_u}d\frac{C_u}{C_l}D_0 + \max\{\tau,1-\tau\}\sqrt{C_u}t_1\exp(-cd),
\end{aligned}$$

where we use the Lipschitz continuity of $f_t(\cdot)$, and the final inequality follows from Lemma 10. For the second phase ($t > t_1$), we have

$$\begin{aligned}
\mathbb{E}\left[f_t(\boldsymbol{\beta}_t)-f_t(\boldsymbol{\beta}^*)\right] = \mathbb{E}\left[\mathbb{E}\left[f_t(\boldsymbol{\beta}_t)-f_t(\boldsymbol{\beta}^*)\big|\boldsymbol{\beta}_t\right]\right] &\leq \frac{C_u}{b_1}\mathbb{E}\|\boldsymbol{\beta}_t-\boldsymbol{\beta}^*\|_2^2 \\
&\leq 2\frac{C_u^2}{C_l^2}\frac{d}{t-t_1+C_b d}b_0^2,
\end{aligned}$$

where the first inequality follows from Lemma 17, and the second from Lemma 10. The proof is thus complete.

A.2 Proof of Batch Learning

In this subsection, we develop the batch learning framework in a sequential manner. We begin with the proof of Theorem 5, proceed to analyze convergence under well-controlled initialization, and conclude with the proof of Theorem 7.

A.2.1 PROOF OF THEOREM 5

The update rule $\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t - \eta_t \mathbf{g}_t$ ensures that

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2 = \|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2 - 2\eta_t\langle\boldsymbol{\beta}_t - \boldsymbol{\beta}^*, \mathbf{g}_t\rangle + \eta_t^2\|\mathbf{g}_t\|_2^2. \tag{10}$$

The behavior of the inner product term $\langle\boldsymbol{\beta}_t - \boldsymbol{\beta}^*, \mathbf{g}_t\rangle$ and the gradient norm $\|\mathbf{g}_t\|_2^2$ varies significantly depending on the proximity of the current iterate $\boldsymbol{\beta}_t$ to the true parameter $\boldsymbol{\beta}^*$. Therefore, we analyze these quantities by distinguishing between regimes where $\boldsymbol{\beta}_t$ is either close to or far from $\boldsymbol{\beta}^*$.

FIRST PHASE ANALYSIS. We shall prove the convergence dynamics by induction. Initially, it holds obviously for $\boldsymbol{\beta}_0$. We continue to assume for the iteration of t, it has $\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2 \leq (1 - \frac{1}{100}\frac{C_l}{C_u})^t \cdot D_0$ and we are going to prove the convergence dynamics at $t + 1$. For convenience, we denote

$$D_t := (1 - \frac{1}{100}\frac{C_l}{C_u})^t \cdot D_0, \quad D_{t+1} := (1 - \frac{1}{100}\frac{C_l}{C_u})^{t+1} \cdot D_0.$$

We proceed to assume the event

$$\mathcal{E}_t := \left\{ \sup_{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{R}^d} |f_t(\boldsymbol{\beta}_1) - f_t(\boldsymbol{\beta}_2) - \mathbb{E}[f_t(\boldsymbol{\beta}_1) - f_t(\boldsymbol{\beta}_2)]| \right.$$
$$\left. \cdot\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2^{-1} \leq C_1 \max\{\tau, 1 - \tau\}\sqrt{\frac{C_u d}{n_t}} \right\}$$

holds. Specifically, Proposition 14 establishes that $\mathbb{P}(\mathcal{E}_t) \geq 1 - \exp(-C_2 d) - \exp(-\sqrt{n_t/\log n_t})$. Under the event $\mathcal{E}_t$, the intermediate term in (10) can be lower bounded as

$$\langle\boldsymbol{\beta}_t - \boldsymbol{\beta}^*, \mathbf{g}_t\rangle \geq f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*) \geq \mathbb{E}[f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*)|\boldsymbol{\beta}_t] - C_1 \max\{\tau, 1-\tau\}\sqrt{\frac{C_u d}{n_t}}\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2,$$

where the first inequality follows from the sub-gradient definition, and the second holds on the event $\mathcal{E}_t$. Additionally, the convexity of the quantile loss, combined with Lemma 12, yields

$$\mathbb{E}[f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*)|\boldsymbol{\beta}_t] \geq \frac{1}{n_t}\sum_{i=1}^{n_t}\mathbb{E}\left[\rho_{Q,\tau}(\langle\mathbf{X}_i^{(t)}, \boldsymbol{\beta}^* - \boldsymbol{\beta}_t\rangle) - \rho_{Q,\tau}(-\xi_i) - \rho_{Q,\tau}(\xi_i)|\boldsymbol{\beta}_t\right]$$
$$\geq \frac{1}{\sqrt{2\pi}}\sqrt{C_l}\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2 - \gamma.$$

Thus, if the batch size satisfies $n_t \geq C(C_u/C_l)\max\{\tau^2, (1-\tau)^2\}d$, and the iterate lies in the outer phase region $\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2 \geq 8\sqrt{C_l^{-1}}\gamma$, it holds that

$$\langle\boldsymbol{\beta}_t - \boldsymbol{\beta}^*, \mathbf{g}_t\rangle \geq \frac{1}{6}\sqrt{C_l}\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2.$$

Moreover, Lemma 16 shows that under the event $\mathcal{E}_t$, $\|\mathbf{g}_t\|_2 \leq \sqrt{C_u}$. Combining these, the update equation (10) can be upper bounded as

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2 \leq \|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2 - \frac{1}{3}\eta_t\sqrt{C_l}\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2 + \eta_t^2 C_u.$$

We now treat the right-hand side of the above inequality as a quadratic function in $\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2$. Using the inductive assumption $\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2 \leq D_t$, we obtain

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2 \leq D_t^2 - \frac{1}{3}\eta_t \sqrt{C_l} D_t + \eta_t^2 C_u.$$

The stepsize $\eta_t = (1 - \frac{1}{100}\frac{C_l}{C_u})^t \eta_0 \in \frac{\sqrt{C_l}}{C_u} D_t \cdot [1/8, 5/24]$ implies that

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2 \leq \left(1 - \frac{1}{50}\frac{C_l}{C_u}\right) D_t^2,$$

which leads to the inductive conclusion $\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2 \leq (1 - \frac{1}{100}\frac{C_l}{C_u})^{t+1} D_0 = D_{t+1}$.

SECOND PHASE ANALYSIS. In the second phase, we shall still assume event $\mathcal{E}_t$ holds, which is defined and discussed in the first phase analyses. Note that according to Lemma 17, in this case, where $\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2 \leq 8\sqrt{C_l^{-1}}\gamma$, the expectation of excess risk has a quadratic lower bound,

$$\mathbb{E}\left[f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*)\big|\boldsymbol{\beta}_t\right] \geq \frac{1}{12b_0}C_l\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2.$$

The second phase region has the constraints $\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2 \geq C\max\{\tau, 1-\tau\}\sqrt{(C_u/C_l^2)d/nb_0}$. Then under event $\mathcal{E}_t$, we have

$$f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*) \geq \frac{1}{12b_0}C_l\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2 - C_1\max\{\tau, 1-\tau\}\sqrt{C_u d/n_t}\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2$$

$$\geq \frac{1}{24b_0}C_l\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2.$$

In addition, Lemma 16 proves that under the event $\mathcal{E}_t$, the following holds in the second phase region:

$$\|\mathbf{g}_t\|_2^2 \leq 4\frac{1}{b_1^2}C_u^2\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2.$$

Then together with the sub-gradient definition $\langle \mathbf{g}_t, \boldsymbol{\beta}_t - \boldsymbol{\beta}^* \rangle \geq f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*)$, equation (10) can be upper bounded as

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2 \leq \|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2 - \eta_t \frac{1}{12b_0}C_l\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2 + \eta_t^2 \frac{4}{b_1^2}C_u^2\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2.$$

By having stepsize $\eta_t \in \frac{C_l}{C_u^2}\frac{b_1^2}{b_0} \cdot [c_1, c_2]$, we obtain

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2 \leq \left(1 - 0.0005\frac{b_1^2}{b_0^2}\frac{C_l^2}{C_u^2}\right)\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2.$$

THIRD PHASE ANALYSIS. We shall prove the convergence dynamics by induction. The proof at t_2 is trivial, which can be obtained similarly to the following analyses, and hence it is skipped. Then we are going to prove $\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2$ when the desired inequality holds for $t_2, \ldots, t$. It is worth

noting that Lemma 16 and Proposition 14 show: in the region of $\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2 \leq C_1 \max\{\tau, 1 - \tau\}\sqrt{(C_u/C_l^2)d/n} \cdot b_0$, with probability exceeding $1 - \exp(-C_1 d) - \exp(-\sqrt{n_t/\log n_t})$, it has

$$\|\mathbf{g}_t\|_2^2 \leq C_3^2 \bar{\tau}^2 C_u \frac{d}{n} \frac{C_u^2 b_0^2}{C_l^2 b_1^2},$$

where we denote $\bar{\tau} := \max\{\tau, 1 - \tau\}$, for convenience. According to the loss function expectation calculations in Lemma 17, we have

$$\mathbb{E}\left[f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*)\big|\boldsymbol{\beta}_t\right] \geq \frac{1}{24b_0}\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2.$$

Then together with the sub-gradient definition, we have

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2 \leq \|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2 - \eta_t \frac{1}{12b_0}C_l\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2 + C_3^2\eta_t^2\bar{\tau}^2 C_u \frac{d}{n}\frac{C_u^2 b_0^2}{C_l^2 b_1^2}$$
$$+ 2\eta_t\left[f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*) - \mathbb{E}\left[f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*)\big|\boldsymbol{\beta}_t\right]\right].$$

Insert the stepsize $\eta_t = \frac{C_a}{t - t_2 + C_b}\frac{b_0}{C_l}$ into the above equation and then it arrives at

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2 \leq \left(1 - \frac{C_a}{12(t - t_2 + C_b)}\right)\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2 + C_3^2\bar{\tau}^2\frac{C_a^2}{(t - t_2 + C_b)^2}\frac{C_u}{C_l^2}\frac{d}{n}\frac{C_u^2 b_0^2}{C_l^2 b_1^2}b_0^2$$
$$+ 2\frac{b_0}{C_l}\frac{C_a}{t - t_2 + C_b}\left[f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*) - \mathbb{E}\left[f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*)\big|\boldsymbol{\beta}_t\right]\right].$$

Applying the above bound repeatedly from t_2, we have

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2$$
$$\leq \prod_{l=t_2}^{t}\left(1 - \frac{C_a}{12(l - t_2 + C_b)}\right)\|\boldsymbol{\beta}_{t_2} - \boldsymbol{\beta}^*\|_2^2$$
$$+ C_3^2\bar{\tau}^2\frac{C_u}{C_l^2}\frac{C_u^2 b_0^2}{C_l^2 b_1^2}\frac{d}{n}b_0^2\sum_{l=t_2}^{t}\left(1 - \frac{C_a}{12(l + 1 - t_2 + C_b)}\right)\cdots\left(1 - \frac{C_a}{12(t - t_2 + C_b)}\right)\frac{C_a^2}{(l - t_2 + C_b)^2}$$
$$+ \frac{b_0}{C_l}\sum_{l=t_2}^{t}\left(1 - \frac{C_a}{12(l + 1 - t_2 + C_b)}\right)\cdots\left(1 - \frac{C_a}{12(t - t_2 + C_b)}\right)\frac{C_a}{l - t_2 + C_b}$$
$$\times\left[f_l(\boldsymbol{\beta}_l) - f_l(\boldsymbol{\beta}^*) - \mathbb{E}\left[f_l(\boldsymbol{\beta}_l) - f_l(\boldsymbol{\beta}^*)\big|\boldsymbol{\beta}_l\right]\right].$$
$$(11)$$

A sharp analysis of equation (11) is key to the proof. Notice that with $C_a \geq 12$, we have $1 - \frac{C_a}{12(l - t_2 + C_b)} \leq 1 - \frac{1}{l - t_2 + C_b} \leq \frac{l - t_2 + C_b}{l + 1 - t_2 + C_b}$. In this way, the first term on the right hand side could be bounded with

$$\prod_{l=t_2}^{t}\left(1 - \frac{C_a}{12(l - t_2 + C_b)}\right)\|\boldsymbol{\beta}_{t_2} - \boldsymbol{\beta}^*\|_2^2 \leq \frac{C_b}{t + 1 - t_2 + C_b}\|\boldsymbol{\beta}_{t_2} - \boldsymbol{\beta}^*\|_2^2.$$

The product sequence satisfies, for each $l = t_2, \ldots, t$, that

$$
\left(1 - \frac{C_a}{12(l + 1 - t_2 + C_b)} \right) \cdots \left(1 - \frac{C_a}{12(t - t_2 + C_b)} \right)
$$
$$
= \exp \left(\sum_{k=l+1}^{t} \log \left(1 - \frac{C_a}{12(k - t_2 + C_b)} \right) \right) \leq \exp \left(- \sum_{k=l+1}^{t} \frac{C_a}{12(k - t_2 + C_b)} \right)
$$
$$
\leq \exp \left(- \frac{C_a}{12} \int_{l+1-t_2}^{t+1-t_2} \frac{1}{x + C_b} \, dx \right) = \exp \left(- \frac{C_a}{12} \log \left(\frac{t + 1 - t_2 + C_b}{l + 1 - t_2 + C_b} \right) \right)
$$
$$
= \left(\frac{l + 1 - t_2 + C_b}{t + 1 - t_2 + C_b} \right)^{\frac{C_a}{12}}.
$$

The second term of equation (11) thus has the following upper bound:

$$
\sum_{l=t_2}^{t} \left(1 - \frac{C_a}{12(l + 1 - t_2 + C_b)} \right) \cdots \left(1 - \frac{C_a}{12(t - t_2 + C_b)} \right) \frac{C_a^2}{(l - t_2 + C_b)^2}
$$
$$
\leq \sum_{l=t_2}^{t} \left(\frac{l + 1 - t_2 + C_b}{t + 1 - t_2 + C_b} \right)^{\frac{C_a}{12}} \frac{C_a^2}{(l - t_2 + C_b)^2}
$$
$$
\leq \left(\frac{C_b + 1}{C_b} \right)^{\frac{C_a}{12}} \frac{C_a^2}{(t + 1 - t_2 + C_b)^{\frac{C_a}{12}}} \sum_{l=t_2}^{t} (l - t_2 + C_b)^{\frac{C_a}{12} - 2}.
$$

With $C_a > 12$, we have

$$
\sum_{l=t_2}^{t} (l - t_2 + C_b)^{\frac{C_a}{12} - 2} \leq \int_{0}^{t - t_2 + 1} (x + C_b)^{\frac{C_a}{12} - 2} \, dx = \frac{1}{\frac{C_a}{12} - 1} (t - t_2 + 1 + C_b)^{\frac{C_a}{12} - 1}.
$$

Hence, we have the upper bound of the second term

$$
\sum_{l=t_2}^{t} \left(1 - \frac{C_a}{12(l + 1 - t_2 + C_b)} \right) \cdots \left(1 - \frac{C_a}{12(t - t_2 + C_b)} \right) \frac{C_a^2}{(l - t_2 + C_b)^2}
$$
$$
\leq \frac{1}{\frac{C_a}{12} - 1} \left(\frac{C_b + 1}{C_b} \right)^{\frac{C_a}{12}} \frac{1}{t + 1 + C_b - t_2}.
$$

Then, it suffices to bound the last term of equation (11). It is worth noting that under the event of $\cap_{l=t_2}^{t} \left\{ \| \boldsymbol{\beta}_l - \boldsymbol{\beta}^* \|_2^2 \leq C^* \frac{C_u}{C_l^2} \frac{C_u^2 b_0^2}{C_l^2 b_1^2} \frac{1}{l - t_2 + C_b} \frac{d}{n} b_0^2 \right\}$, we have

$$
\| f_l(\boldsymbol{\beta}_l) - f_l(\boldsymbol{\beta}^*) - \mathbb{E} \left[f_l(\boldsymbol{\beta}_l) - f_l(\boldsymbol{\beta}^*) | \boldsymbol{\beta}_l \right] \|_{\Psi_2} \leq \sqrt{C^* \frac{C_u^2}{C_l^2} \frac{C_u^2 b_0^2}{C_l^2 b_1^2} \frac{1}{l - t_2 + C_b} \frac{d}{n^2}} b_0.
$$

Furthermore, invoking Azuma's sub-Gaussian inequality (e.g., Theorem 2 in Shamir (2011)) in conjunction with equation (12), it holds with probability exceeding $1 - \exp(-cd)$ that

$$
\frac{b_0}{C_l} \sum_{l=t_2}^{t} \left(1 - \frac{C_a}{12(l + 1 - t_2 + C_b)} \right) \cdots \left(1 - \frac{C_a}{12(t - t_2 + C_b)} \right) \frac{C_a}{l - t_2 + C_b}
$$

$$\times \left[f_l(\boldsymbol{\beta}_l) - f_l(\boldsymbol{\beta}^*) - \mathbb{E}\left[f_l(\boldsymbol{\beta}_l) - f_l(\boldsymbol{\beta}^*)|\boldsymbol{\beta}_l\right]\right]$$

$$\leq \frac{b_0}{C_l} \left(d \sum_{l=t_2}^{t} \prod_{k=l+1}^{t} \left(1 - \frac{C_a}{12(k-t_2+C_b)}\right)^2 \frac{C_a^2}{(l-t_2+C_b)^2} \right.$$

$$\left. \times \left\| f_l(\boldsymbol{\beta}_l) - f_l(\boldsymbol{\beta}^*) - \mathbb{E}\left[f_l(\boldsymbol{\beta}_l) - f_l(\boldsymbol{\beta}^*)|\boldsymbol{\beta}_l\right]\right\|_{\Psi_2}^2 \right)^{1/2}$$

$$\leq \bar{\tau}^2 \frac{C_u}{C_l^2} \frac{C_u b_0}{C_l b_1} \frac{d}{n} b_0^2 \frac{CC_a\sqrt{C^*}}{(t+1-t_2+C_b)^{\frac{C_a}{12}}} \sqrt{\sum_{l=t_2}^{t}(l+1-t_2+C_b)^{\frac{C_a}{6}-3}}$$

$$\leq \bar{\tau}^2 \frac{C_u}{C_l^2} \frac{C_u b_0}{C_l b_1} \frac{d}{n} b_0^2 \frac{CC_a\sqrt{C^*}}{t+1-t_2+C_b}.$$

Therefore, combining the pieces, equation (11) can be bounded from above as

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2 \leq C^* \bar{\tau}^2 \frac{C_u}{C_l^2} \frac{d}{n} \frac{C_u^2 b_0^2}{C_l^2 b_1^2} \frac{1}{t+1-t_2+C_b} b_0^2,$$

where $C^* > C_a^2$ is some sufficiently large constant.

A.2.2 ANALYSIS OF BATCH LEARNING WITH WELL-CONTROLLED INITIAL ERRORS

We will elucidate the convergence dynamics under the condition of well-controlled initial error rates, specifically when $\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2^2 \leq C_1^2 \max\{\tau^2, (1-\tau)^2\}(C_u/C_l^2)(d/n) \cdot b_0^2$. Employing a similar analysis as applied in the third phase of Theorem 5, we obtain

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2 \leq \|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2 - \eta_t \frac{1}{12 b_0} C_l \|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2 + C_3^2 \eta_t^2 \bar{\tau}^2 C_u \frac{d}{n} \frac{C_u^2 b_0^2}{C_l^2 b_1^2}$$

$$- 2\eta_t \left[f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*) - \mathbb{E}\left[f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*)|\boldsymbol{\beta}_t\right]\right].$$

Inserting the stepsize $\eta_t = \frac{C_a}{t+C_b}\frac{b_0}{C_l}$ into the above equation results in

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2 \leq \left(1 - \frac{C_a}{12(t+C_b)}\right) \|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2 + C_3^2 \bar{\tau}^2 \frac{C_u}{C_l^2} \frac{C_a^2}{(t+C_b)^2} \frac{d}{n} b_0^2$$

$$- 2\frac{C_a}{t+C_b}\frac{b_0}{C_l} \left[f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*) - \mathbb{E}\left[f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*)|\boldsymbol{\beta}_t\right]\right].$$

Moreover, the preceding equation implies

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2$$

$$\leq \prod_{l=0}^{t} \left(1 - \frac{C_a}{12(l+C_b)}\right) \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2^2$$

$$+ C_3^2 \bar{\tau}^2 \frac{C_u^2 b_0^2}{C_l^2 b_1^2} \frac{C_u}{C_l^2} \sum_{l=0}^{t} \left(1 - \frac{C_a}{12(l+1+C_b)}\right) \cdots \left(1 - \frac{C_a}{12(t+C_b)}\right) \frac{C_a^2}{(l+C_b)^2} \frac{d}{n} b_0^2$$

$$- \frac{b_0}{C_l} \sum_{l=0}^{t} \left(1 - \frac{C_a}{12(l+1+C_b)}\right) \cdots \left(1 - \frac{C_a}{12(t+C_b)}\right) \frac{C_a}{l+C_b}$$

$$\times \left[f_l(\boldsymbol{\beta}_l) - f_l(\boldsymbol{\beta}^*) - \mathbb{E} \left[f_l(\boldsymbol{\beta}_l) - f_t(\boldsymbol{\beta}^*) \middle| \boldsymbol{\beta}_l \right] \right].$$

In what follows, we are going to upper bound each term of the right hand side equation.

Case One: large C_a. With $C_a > 12$, the first term on the right-hand side can be bounded as

$$\prod_{l=0}^{t} \left(1 - \frac{C_a}{12(l + C_b)} \right) \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2^2$$

$$\leq \prod_{l=0}^{t} \left(1 - \frac{1}{l + C_b} \right) \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2^2 \leq \frac{C_b}{t + 1 + C_b} \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2^2.$$

The product sequence is bounded by

$$\left(1 - \frac{C_a}{12(l + 1 + C_b)} \right) \cdots \left(1 - \frac{C_a}{12(t + C_b)} \right) \leq \exp \left(-\frac{C_a}{12} \sum_{k=l+1}^{t} \frac{1}{k + C_b} \right)$$

$$\leq \exp \left(-\frac{C_a}{12} \int_{l+1}^{t+1} \frac{1}{x + C_b} \, dx \right) = \left(\frac{l + 1 + C_b}{t + 1 + C_b} \right)^{\frac{C_a}{12}}.$$

Subsequently, the second term can be bounded as

$$\sum_{l=0}^{t} \left(1 - \frac{C_a}{12(l + 1 + C_b)} \right) \cdots \left(1 - \frac{C_a}{12(t + C_b)} \right) \frac{C_a^2}{(l + C_b)^2}$$

$$\leq \sum_{l=0}^{t} \left(\frac{l + C_b + 1}{t + C_b + 1} \right)^{\frac{C_a}{12}} \frac{C_a^2}{(l + C_b)^2} \leq \left(\frac{C_b + 1}{C_b} \right)^{\frac{C_a}{12}} \frac{C_a^2}{t + C_b + 1}.$$

Next, consider the last term. It is noteworthy that under $\cap_{l=0}^{t} \{ \|\boldsymbol{\beta}_l - \boldsymbol{\beta}^*\|_2^2 \leq C^* \bar{\tau}^2 \frac{C_u}{C_l^2} \frac{1}{l+C_b} \frac{d}{n} b_0^2 \}$, we have

$$\| f_l(\boldsymbol{\beta}_l) - f_l(\boldsymbol{\beta}^*) - \mathbb{E} \left[f_l(\boldsymbol{\beta}_l) - f_t(\boldsymbol{\beta}^*) \middle| \boldsymbol{\beta}_l \right] \|_{\Psi_2}^2 \leq C^* \bar{\tau}^4 \frac{d}{n^2} \frac{C_u^2}{C_l^2} \frac{1}{l + C_b} b_0^2.$$

Therefore, in accordance with Theorem 2 in Shamir (2011), with probability at least $1 - \exp(-cd)$, the last term satisfies

$$\frac{b_0}{C_l} \sum_{l=0}^{t} \left(1 - \frac{C_a}{12(l + 1 + C_b)} \right) \cdots \left(1 - \frac{C_a}{12(t + C_b)} \right) \frac{C_a}{l + C_b}$$

$$\times \left[f_l(\boldsymbol{\beta}_l) - f_l(\boldsymbol{\beta}^*) - \mathbb{E} \left[f_l(\boldsymbol{\beta}_l) - f_t(\boldsymbol{\beta}^*) \middle| \boldsymbol{\beta}_l \right] \right]$$

$$\leq C_a \bar{\tau}^2 \frac{C_u}{C_l^2} \frac{C_u b_0}{C_l b_1} \frac{d}{n} b_0^2 \sqrt{ \sum_{l=0}^{t} \left(\frac{l + 1 + C_b}{t + 1 + C_b} \right)^{\frac{C_a}{6}} \frac{C^*}{(l + C_b)^3} }$$

$$\leq C_a \bar{\tau}^2 \frac{C_u}{C_l^2} \frac{C_u b_0}{C_l b_1} \frac{\sqrt{C^*}}{t + 1 + C_b} \frac{d}{n} b_0^2.$$

43

Putting together the pieces, we conclude that

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2 \leq C^* \bar{\tau}^2 \frac{C_u}{C_l^2} \frac{C_u^2 b_0^2}{C_l^2 b_1^2} \frac{1}{t+1+C_b} \frac{d}{n} b_0^2.$$

<u>Case Two: small C_a</u> In this setting, with $C_a < 12$, the second term satisfies

$$\sum_{l=0}^{t} \left(1 - \frac{C_a}{12(l+1+C_b)}\right) \cdots \left(1 - \frac{C_a}{12(t+C_b)}\right) \frac{C_a^2}{(l+C_b)^2}$$

$$\leq \sum_{l=0}^{t} \left(\frac{l+C_b+1}{t+C_b+1}\right)^{\frac{C_a}{12}} \frac{C_a^2}{(l+C_b)^2} \leq \frac{C_a^2}{(t+C_b+1)^{\frac{C_a}{12}}}.$$

Under the event $\cap_{l=0}^{t}\{\|\boldsymbol{\beta}_l - \boldsymbol{\beta}^*\|_2^2 \leq \frac{c}{(l+C_b)^{\frac{C_a}{12}}} \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2^2\}$, we have

$$\|f_l(\boldsymbol{\beta}_l) - f_l(\boldsymbol{\beta}^*) - \mathbb{E}\left[f_l(\boldsymbol{\beta}_l) - f_t(\boldsymbol{\beta}^*)|\boldsymbol{\beta}_l\right]\|_{\Psi_2}^2 \leq C^* \bar{\tau} \frac{C_u}{n} \frac{1}{(l+C_b)^{\frac{C_a}{12}}} \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2^2.$$

Thus, with probability exceeding $1 - \exp(-cd)$, it holds

$$\frac{b_0}{C_l} \sum_{l=0}^{t} \left(1 - \frac{C_a}{12(l+1+C_b)}\right) \cdots \left(1 - \frac{C_a}{12(t+C_b)}\right) \frac{C_a}{l+C_b}$$

$$\times \left[f_l(\boldsymbol{\beta}_l) - f_l(\boldsymbol{\beta}^*) - \mathbb{E}\left[f_l(\boldsymbol{\beta}_l) - f_t(\boldsymbol{\beta}^*)|\boldsymbol{\beta}_l\right]\right]$$

$$\leq \bar{\tau} C_a \frac{\sqrt{d}}{\sqrt{n}} \frac{\sqrt{C_u}}{C_l} b_0 \cdot \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2 \cdot \sqrt{\sum_{l=0}^{t} \left(\frac{l+1+C_b}{t+1+C_b}\right)^{\frac{C_a}{6}} \frac{C^*}{(l+C_b)^{2+\frac{C_a}{12}}}}$$

$$\leq \bar{\tau} C_a \frac{\sqrt{d}}{\sqrt{n}} \frac{\sqrt{C_u}}{C_l} b_0 \cdot \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2 \cdot \frac{\sqrt{C^*}}{(t+1+C_b)^{\frac{C_a}{12}}}.$$

To sum up, we establish the upper bound

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2 \leq \left(\frac{1+C_b}{t+1+C_b}\right)^{\frac{C_a}{12}} \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2^2 + C_3^2 \bar{\tau}^2 \frac{C_u}{C_l^2} \frac{C_u^2 b_0^2}{C_l^2 b_1^2} \frac{d}{n} b_0^2 \cdot \frac{C_a^2}{(t+C_b+1)^{\frac{C_a}{12}}}$$

$$+ \bar{\tau} \frac{\sqrt{d}}{\sqrt{n}} \frac{\sqrt{C_u}}{C_l} b_0 \cdot \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2 \cdot \frac{C_a \sqrt{C^*}}{(t+1+C_b)^{\frac{C_a}{12}}}.$$

Therefore, by choosing sufficiently small $C_a < (C_l/C_u)(b_1/b_0)(C_l/\sqrt{C_u})\sqrt{n/d}\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2/\bar{\tau} b_0$ and $C_a < C_b \leq 1$, we obtain

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2 \leq \frac{C^*}{(t+1+C_b)^{\frac{C_a}{12}}} \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2^2.$$

A.2.3 PROOF OF THEOREM 7

Prior to demonstrating Theorem 7, it is imperative to establish the ensuing lemma.

Lemma 11 *Assume the same conditions and stepsize schemes as in Theorem 5. In the second phase,* $\mathbb{E}\|\boldsymbol{\beta}_{t+1}-\boldsymbol{\beta}^*\|_2^2 \leq (1-c_1(C_l/C_u)/(b_1^2/b_0^2))^{t+1}\|\boldsymbol{\beta}_{t_1}-\boldsymbol{\beta}^*\|_2^2$; *and in phrase three,* $\mathbb{E}\|\boldsymbol{\beta}_{t+1}-\boldsymbol{\beta}^*\|_2^2 \leq C\frac{C_u}{C_l^2}\frac{1}{t+1-t_2+C_b}\frac{d}{n}b_0^2.$

Proof Note that the expectation can be expressed as

$$\mathbb{E}\left[\|\boldsymbol{\beta}_{t+1}-\boldsymbol{\beta}^*\|_2^2\right] = \mathbb{E}\left[\mathbb{E}\left[\|\boldsymbol{\beta}_{t+1}-\boldsymbol{\beta}^*\|_2^2\big|\boldsymbol{\beta}_t\right]\right]$$
$$= \mathbb{E}\left[\mathbb{E}\left[\|\boldsymbol{\beta}_t-\boldsymbol{\beta}^*\|_2^2 - 2\eta_t\langle\boldsymbol{\beta}_t-\boldsymbol{\beta}^*,\mathbf{g}_t\rangle + \eta_t^2\|\mathbf{g}_t\|_2^2\big|\boldsymbol{\beta}_t\right]\right]$$

We first consider the inside conditional expectation. According to the sub-gradient definition and considering that η_t is independent of $(\mathbf{X}_t, Y_t)$, we have

$$\mathbb{E}\left[\|\boldsymbol{\beta}_t-\boldsymbol{\beta}^*\|_2^2 - 2\eta_t\langle\boldsymbol{\beta}_t-\boldsymbol{\beta}^*,\mathbf{g}_t\rangle + \eta_t^2\|\mathbf{g}_t\|_2^2\big|\boldsymbol{\beta}_t\right]$$
$$\leq \|\boldsymbol{\beta}_t-\boldsymbol{\beta}^*\|_2^2 - 2\eta_t\mathbb{E}\left[f_t(\boldsymbol{\beta}_t)-f_t(\boldsymbol{\beta}^*)\big|\boldsymbol{\beta}_t\right] + \eta_t^2\mathbb{E}\left[\|\mathbf{g}_t\|_2^2\big|\boldsymbol{\beta}_t\right] \tag{13}$$

The sub-gradient value at 0 won't affect the expectation calculations, so for convenience, we let $\partial\rho_{Q,\tau}(x)|_{x=0} = 0$. Then the sub-gradient at $\boldsymbol{\beta}_t$ could be written as

$$n_t\mathbf{g}_t = \sum_{i=1}^{n_t}\mathbf{X}_i^{(t)}\cdot\left(\tau 1_{Y_i^{(t)}<\langle\mathbf{X}_i^{(t)},\boldsymbol{\beta}_t\rangle} + (1-\tau)1_{Y_i^{(t)}>\langle\mathbf{X}_i^{(t)},\boldsymbol{\beta}_t\rangle}\right).$$

Its length is bounded with

$$n_t^2\|\mathbf{g}_t\|_2^2 = \sum_{i=1}^{n_t}\|\mathbf{X}_i^{(t)}\|_2^2\cdot\left(\tau^2 1_{Y_i^{(t)}<\langle\mathbf{X}_i^{(t)},\boldsymbol{\beta}_t\rangle} + (1-\tau)^2 1_{Y_i^{(t)}>\langle\mathbf{X}_i^{(t)},\boldsymbol{\beta}_t\rangle}\right)$$
$$+ \sum_{i\neq j}\langle\mathbf{X}_i^{(t)},\mathbf{X}_j^{(t)}\rangle\cdot\left(\tau 1_{Y_i^{(t)}<\langle\mathbf{X}_i^{(t)},\boldsymbol{\beta}_t\rangle} + (1-\tau)1_{Y_i^{(t)}>\langle\mathbf{X}_i^{(t)},\boldsymbol{\beta}_t\rangle}\right)$$
$$\times\left(\tau 1_{Y_j^{(t)}<\langle\mathbf{X}_j^{(t)},\boldsymbol{\beta}_t\rangle} + (1-\tau)1_{Y_j^{(t)}>\langle\mathbf{X}_j^{(t)},\boldsymbol{\beta}_t\rangle}\right).$$

Taking the $\boldsymbol{\beta}_t$ conditional expectation on each side of the above equation leads to

$$n_t^2\mathbb{E}\left[\|\mathbf{g}_t\|_2^2\big|\boldsymbol{\beta}_t\right] \leq \max\{\tau^2, (1-\tau)^2\}C_u n_t d + \frac{n_t(n_t-1)}{2}\mathbb{E}\left[\langle\mathbf{X}_i^{(t)},\mathbf{X}_j^{(t)}\rangle\right.$$
$$\times\left.\left(H_\xi(\langle\mathbf{X}_i^{(t)},\boldsymbol{\beta}_t-\boldsymbol{\beta}^*\rangle)-H_\xi(0)\right)\cdot\left(H_\xi(\langle\mathbf{X}_j^{(t)},\boldsymbol{\beta}_t-\boldsymbol{\beta}^*\rangle)-H_\xi(0)\right)\big|\boldsymbol{\beta}_t\right].$$

Notice that $\mathbf{X}_i^{(t)}$ follows Assumption 1 and then the transformed vector $\mathbf{Z}_i^{(t)} := \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{X}_i^{(t)}$ is isotropic. Thus we have

$$E := \mathbb{E}\left[\langle\mathbf{X}_i^{(t)},\mathbf{X}_j^{(t)}\rangle\cdot\left(H_\xi(\langle\mathbf{X}_i^{(t)},\boldsymbol{\beta}_t-\boldsymbol{\beta}^*\rangle)-H_\xi(0)\right)\cdot\left(H_\xi(\langle\mathbf{X}_j^{(t)},\boldsymbol{\beta}_t-\boldsymbol{\beta}^*\rangle)-H_\xi(0)\right)\big|\boldsymbol{\beta}_t\right]$$
$$= \mathbb{E}\left[\langle\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{Z}_i^{(t)},\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{Z}_j^{(t)}\rangle\times\left(H_\xi(\langle\mathbf{Z}_i^{(t)},\boldsymbol{\Sigma}^{\frac{1}{2}}(\boldsymbol{\beta}_t-\boldsymbol{\beta}^*)\rangle)-H_\xi(0)\right)\right.$$

$$\times \left(H_\xi(\langle \mathbf{Z}_j^{(t)}, \mathbf{\Sigma}^{\frac{1}{2}}(\boldsymbol{\beta}_t - \boldsymbol{\beta}^*)\rangle) - H_\xi(0) \right) \big| \boldsymbol{\beta}_t \right].$$

There exists a set of unit length orthogonal vectors $\{\mathbf{e}_1 := \frac{\mathbf{\Sigma}^{\frac{1}{2}}(\boldsymbol{\beta}_t - \boldsymbol{\beta}^*)}{\|\mathbf{\Sigma}^{\frac{1}{2}}(\boldsymbol{\beta}_t - \boldsymbol{\beta}^*)\|_2}, \mathbf{e}_2, \ldots, \mathbf{e}_d\}$ such that any two of them satisfy $\mathbf{e}_i^\top \mathbf{e}_j = 0$ with $i \neq j$. Then the vector $\mathbf{Z}_i$ can be decomposed into d independent random vectors

$$\mathbf{Z}_i^{(t)} = \frac{\langle \mathbf{Z}_i^{(t)}, \mathbf{\Sigma}^{\frac{1}{2}}(\boldsymbol{\beta}_t - \boldsymbol{\beta}^*)\rangle}{\|\mathbf{\Sigma}^{\frac{1}{2}}(\boldsymbol{\beta}_t - \boldsymbol{\beta}^*)\|_2^2} \mathbf{\Sigma}^{\frac{1}{2}}(\boldsymbol{\beta}_t - \boldsymbol{\beta}^*) + \langle \mathbf{Z}_i^{(t)}, \mathbf{e}_2\rangle \mathbf{e}_2 + \cdots + \langle \mathbf{Z}_i^{(t)}, \mathbf{e}_d\rangle \mathbf{e}_d.$$

And $\mathbf{Z}_j^{(t)}$ could be written in a parallel way. It is worth noting that the term $H_\xi(\langle \mathbf{Z}_i^{(t)}, \mathbf{\Sigma}^{\frac{1}{2}}(\boldsymbol{\beta}_t - \boldsymbol{\beta}^*)\rangle) - H_\xi(0)$ is independent of the last $d-1$ components of the decomposition and $\mathbb{E}\langle \mathbf{Z}_i^{(t)}, \mathbf{e}_l\rangle = \mathbf{0}$ holds with $l = 2, \ldots, d$. Thus we have

$$E = \frac{C_u}{\|\mathbf{\Sigma}^{\frac{1}{2}}(\boldsymbol{\beta}_t - \boldsymbol{\beta}^*)\|_2^2} \mathbb{E}\left[\langle \mathbf{Z}_i^{(t)}, \mathbf{\Sigma}^{\frac{1}{2}}(\boldsymbol{\beta}_t - \boldsymbol{\beta}^*)\rangle \langle \mathbf{Z}_j^{(t)}, \mathbf{\Sigma}^{\frac{1}{2}}(\boldsymbol{\beta}_t - \boldsymbol{\beta}^*)\rangle \right.$$
$$\left. \times \left(H_\xi(\langle \mathbf{Z}_i^{(t)}, \mathbf{\Sigma}^{\frac{1}{2}}(\boldsymbol{\beta}_t - \boldsymbol{\beta}^*)\rangle) - H_\xi(0) \right) \cdot \left(H_\xi(\langle \mathbf{Z}_j^{(t)}, \mathbf{\Sigma}^{\frac{1}{2}}(\boldsymbol{\beta}_t - \boldsymbol{\beta}^*)\rangle) - H_\xi(0) \right) \big| \boldsymbol{\beta}_t \right].$$

There are two methods to upper bound the above equation, which are presented as follows.

1. Bound $H_\xi(\langle \mathbf{Z}_i^{(t)}, \mathbf{\Sigma}^{\frac{1}{2}}(\boldsymbol{\beta}_t - \boldsymbol{\beta}^*)\rangle) - H_\xi(0)$ with constant 1. Then we have

$$E \leq \frac{C_u}{\|\mathbf{\Sigma}^{\frac{1}{2}}(\boldsymbol{\beta}_t - \boldsymbol{\beta}^*)\|_2^2} \mathbb{E}\left[|\langle \mathbf{Z}_i^{(t)}, \mathbf{\Sigma}^{\frac{1}{2}}(\boldsymbol{\beta}_t - \boldsymbol{\beta}^*)\rangle| |\langle \mathbf{Z}_j^{(t)}, \mathbf{\Sigma}^{\frac{1}{2}}(\boldsymbol{\beta}_t - \boldsymbol{\beta}^*)\rangle| \big| \boldsymbol{\beta}_t \right]$$
$$\leq C_u.$$

   In this way, we have

$$\mathbb{E}\left[\|\mathbf{g}_t\|_2^2 \big| \boldsymbol{\beta}_t \right] \leq \max\{\tau^2, (1-\tau)^2\} C_u \frac{d}{n_t} + C_u.$$

2. Bound $H_\xi(\langle \mathbf{Z}_i^{(t)}, \mathbf{\Sigma}^{\frac{1}{2}}(\boldsymbol{\beta}_t - \boldsymbol{\beta}^*)\rangle) - H_\xi(0)$ using conditions in Assumption 2, which implies $H_\xi(\langle \mathbf{Z}_i^{(t)}, \mathbf{\Sigma}^{\frac{1}{2}}(\boldsymbol{\beta}_t - \boldsymbol{\beta}^*)\rangle) - H_\xi(0) \leq \frac{1}{b_1} |\langle \mathbf{Z}_i^{(t)}, \mathbf{\Sigma}^{\frac{1}{2}}(\boldsymbol{\beta}_t - \boldsymbol{\beta}^*)\rangle|$. Then we have

$$E \leq \frac{C_u}{b_1^2 \|\mathbf{\Sigma}^{\frac{1}{2}}(\boldsymbol{\beta}_t - \boldsymbol{\beta}^*)\|_2^2} \mathbb{E}\left[|\langle \mathbf{Z}_i^{(t)}, \mathbf{\Sigma}^{\frac{1}{2}}(\boldsymbol{\beta}_t - \boldsymbol{\beta}^*)\rangle|^2 |\langle \mathbf{Z}_j^{(t)}, \mathbf{\Sigma}^{\frac{1}{2}}(\boldsymbol{\beta}_t - \boldsymbol{\beta}^*)\rangle|^2 \big| \boldsymbol{\beta}_t \right]$$
$$\leq \frac{1}{b_1^2} C_u^2 \|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2.$$

   This method leads to

$$\mathbb{E}\left[\|\mathbf{g}_t\|_2^2 \big| \boldsymbol{\beta}_t \right] \leq \max\{\tau^2, (1-\tau)^2\} C_u \frac{d}{n_t} + \frac{C_u^2}{b_1^2} \|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2.$$

Thus, altogether, we have

$$\mathbb{E}\left[\|\mathbf{g}_t\|_2^2 | \boldsymbol{\beta}_t\right] \le \max\{\tau^2, (1-\tau)^2\} C_u \frac{d}{n_t} + \min\left\{C_u, \frac{C_u^2}{b_1^2}\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2\right\}. \tag{14}$$

Then we are ready to analyze the convergence dynamics in each phase.

SECOND PHASE ANALYSIS. In this region, equation (14) yields $\mathbb{E}\left[\|\mathbf{g}_t\|_2^2 | \boldsymbol{\beta}_t\right] \le 2\frac{C_u^2}{b_1^2}\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2$ and Lemma 17 provides $\mathbb{E}[f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*) | \boldsymbol{\beta}_t] \ge \frac{C_l}{b_0}\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2$. Hence, equation (13) is upper bounded with

$$\mathbb{E}\left[\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2 | \boldsymbol{\beta}_t\right] \le \|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2 - \eta_t \frac{C_l}{12 b_0}\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2 + \eta_t^2 \frac{C_u}{b_1^2}\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2$$

$$\le \left(1 - \frac{C_l}{C_u}\frac{b_1^2}{b_0^2}\right)\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2.$$

Take expectation over $\boldsymbol{\beta}_t$ and then we obtain

$$\mathbb{E}\left[\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2\right] \le \left(1 - \frac{C_l}{C_u}\frac{b_1^2}{b_0^2}\right)\mathbb{E}\left[\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2\right].$$

THIRD PHASE ANALYSIS. In this region, (14) guarantees $\mathbb{E}[\|\mathbf{g}_t\|_2^2 | \boldsymbol{\beta}_t] \le 2(C_u^2/C_l^2)(b_0^2/b_1^2)\bar{\tau}^2 C_u \frac{d}{n_t}$ and Lemma 17 proves $\mathbb{E}[f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*) | \boldsymbol{\beta}_t] \ge \frac{C_l}{b_0}\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2$. Hence, equation (13) is upper bounded with

$$\mathbb{E}\left[\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2 | \boldsymbol{\beta}_t\right]$$

$$\le \|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2 - \eta_t \frac{C_l}{12 b_0}\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2 + \eta_t^2 \frac{C_u^2 b_0^2}{C_l^2 b_1^2}\max\{\tau^2, (1-\tau)^2\} C_u \frac{d}{n}$$

$$\le \left(1 - \frac{C_a}{12(t - t_2 + C_b)}\right)\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2 + \frac{C_a}{(t - t_2 + C_b)^2}\frac{C_u^2 b_0^2}{C_l^2 b_1^2}\max\{\tau^2, (1-\tau)^2\}\frac{C_u}{C_l^2}\frac{d}{n}b_0^2.$$

Take expectation over $\boldsymbol{\beta}_t$ and then we have

$$\mathbb{E}\left[\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2\right]$$

$$\le \left(1 - \frac{C_a}{12(t - t_2 + C_b)}\right)\mathbb{E}\left[\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2\right] + \frac{C_a}{(t - t_2 + C_b)^2}\max\{\tau^2, (1-\tau)^2\}\frac{C_u^2 b_0^2}{C_l^2 b_1^2}\frac{C_u}{C_l^2}\frac{d}{n}b_0^2$$

$$\le C^*\left(1 - \frac{C_a}{24(t - t_2 + C_b)}\right)\frac{C_u^2 b_0^2}{C_l^2 b_1^2}\frac{C_u}{C_l^2}\frac{1}{t - t_2 + C_b}\frac{d}{n}b_0^2 \le C^*\frac{C_u^2 b_0^2}{C_l^2 b_1^2}\frac{C_u}{C_l^2}\frac{1}{t + 1 - t_2 + C_b}\frac{d}{n}b_0^2.$$

$\blacksquare$

Then we are ready to prove Theorem 7. According to the expected excess risk bound in Lemma 17, we have $\boldsymbol{\beta}^* = \arg\max_{\boldsymbol{\beta}} \mathbb{E}\sum_{t=0}^T f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta})$. In the first phase, it has

$$\mathbb{E}\left[f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*)\right] \le \max\{\tau, 1-\tau\}\sqrt{C_u}\left(\mathbb{E}\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2\right)^{1/2}$$

$$\le \max\{\tau, 1-\tau\}\sqrt{C_u}\left(1 - c(C_l/C_u)/(\max\{\tau^2, (1-\tau)^2\})\right)^t \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2,$$

where the first line is obtained in the same way as Theorem 4 and the last line follows from Lemma 10. In the second phase and the third phase, according to Lemma 17, we have

$$\mathbb{E}\left[f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*)\right] = \mathbb{E}\left[\mathbb{E}\left[f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*)|\boldsymbol{\beta}_t\right]\right] \leq \frac{C_u}{b_1}\mathbb{E}\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2.$$

As a result, applying Lemma 11, the regret can ultimately be bounded by

$$
\begin{aligned}
\mathsf{Regret}_T &= \mathbb{E}\sum_{t=0}^{T} f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*)\\
&\leq \sqrt{C_u}\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2 \sum_{t=0}^{t_1}(1 - c(C_l/C_u))^t\\
&\quad + \frac{C_u}{b_1}\sum_{t=t_1}^{t_2}\left(1 - c\frac{b_1^2}{b_0^2}\frac{C_l^2}{C_u^2}\right)^{t-t_1}\|\boldsymbol{\beta}_{t_1} - \boldsymbol{\beta}^*\|_2^2 + C\frac{C_u^3}{C_l^3}\frac{b_0^3}{b_1^3}\frac{d}{n}b_0\sum_{t=t_1}^{T}\frac{1}{t-t_1+C_b}\\
&\leq C\frac{C_u}{C_l}\sqrt{C_u}\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2 + C\frac{ku^3}{C_l^3}\frac{b_0^2}{b_1^2}\gamma^2/b_1 + C\frac{C_u^3}{C_l^3}\frac{b_0^3}{b_1^3}\frac{d}{n}b_0\log\left(\frac{T+1-t_1+C_b}{C_b}\right)\\
&\leq C\frac{C_u^3}{C_l^3}\frac{b_0^2}{b_1^2}\max\{\sqrt{C_u}\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2, \gamma^2/b_1\} + C\frac{C_u^3}{C_l^3}\frac{b_0^3}{b_1^3}\frac{d}{n}b_0\log\left(\frac{T+1-t_1+C_b}{C_b}\right).
\end{aligned}
$$

The proof is then complete.

A.3 Proof of Sequential Learning with Infinite Storage

Same as the previous settings, here the update at t can be characterized as the follows,

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2 = \|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2 - 2\eta_t\langle\boldsymbol{\beta}_t - \boldsymbol{\beta}^*, \mathbf{g}_t\rangle + \eta_t^2\|\mathbf{g}_t\|_2^2. \tag{15}$$

We begin by establishing the regularity properties in this setting, which will be instrumental in proving convergence. According to Lemma 10 in Shen et al. (2023) and Lemma 16, under the event

$$
\mathcal{E}_t := \left\{\sup_{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{R}^d} |f_t(\boldsymbol{\beta}_1) - f_t(\boldsymbol{\beta}_2) - \mathbb{E}\left[f_t(\boldsymbol{\beta}_1) - f_t(\boldsymbol{\beta}_2)\right]| \cdot \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2^{-1}\right.
$$

$$
\left. \leq C\max\{\tau, 1-\tau\}\sqrt{C_u d/\sum_{l=0}^{t} n_l}\right\},
$$

the following properties hold for any $\boldsymbol{\beta} \in \mathbb{R}^d$ and corresponding sub-gradient $\mathbf{g} \in \partial f_t(\boldsymbol{\beta})$:

$$f_t(\boldsymbol{\beta}) - f_t(\boldsymbol{\beta}^*) \geq \frac{1}{4}\sqrt{C_l}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 - \gamma, \quad \|\mathbf{g}\|_2 \leq \sqrt{C_u}, \tag{16}$$

and simultaneously,

$$
\begin{aligned}
f_t(\boldsymbol{\beta}) - f_t(\boldsymbol{\beta}^*) &\geq \frac{1}{12b_0}C_l\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 - C_1\sqrt{C_u d/\sum_{l=0}^{t} n_l}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2,\\
\|\mathbf{g}\|_2 &\leq 2\frac{1}{b_1}C_u\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 + C_2\sqrt{C_u d/\sum_{l=0}^{t} n_l},
\end{aligned}
\tag{17}
$$

where $C > 0$ is a universal constant that is independent of both the ambient dimension and the sample size. Specifically, Proposition 14 shows that

$$\mathbb{P}(\mathcal{E}_t) \geq 1 - \exp(-C_3 d) - \exp\left(-\sqrt{\sum_{l=0}^{t} n_l / \log\left(\sum_{l=0}^{t} n_l\right)}\right).$$

A.3.1 FIRST PHASE

We shall also prove the convergence dynamics via induction. The desired inequality at $\boldsymbol{\beta}_0$ is obvious. Then assume we already have the dynamics at t, namely, $\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2 \leq \left(1 - \frac{1}{100}\frac{C_l}{C_u}\right)^t \cdot D_0$ and we are going to prove the concentration at $t + 1$. For convenience, denote

$$D_t := \left(1 - \frac{1}{100}\frac{C_l}{C_u}\right)^t \cdot D_0, \quad D_{t+1} := \left(1 - \frac{1}{100}\frac{C_l}{C_u}\right)^{t+1} \cdot D_0.$$

Definition of sub-gradient and event $\mathcal{E}_t$ together infer that the intermediate term of (15) could be lower bounded with

$$\langle \boldsymbol{\beta}_t - \boldsymbol{\beta}^*, \mathbf{g}_t \rangle \geq f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*)$$

$$\geq \mathbb{E}\left[f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*)\big|\boldsymbol{\beta}_t\right] - C\max\{\tau, 1 - \tau\}\sqrt{C_u d / \sum_{l=0}^{t} n_l}\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2$$

$$\geq \frac{1}{6}\sqrt{C_l}\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2,$$

where the last line uses expectation calculations, which can be found in the batch learning setting and $n_0 \geq C\max\{\tau^2, (1-\tau)^2\}(C_u/C_l)d$. Moreover, Lemma 16 proves that under $\mathcal{E}_t$, $\|\mathbf{g}_t\|_2 \leq \sqrt{C_u}$. Then (15) could be upper bounded with

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2 \leq \|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2 - \frac{1}{3}\eta_t\sqrt{C_l}\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2 + \eta_t^2 C_u.$$

Then by the induction $\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2 \leq D_t$ and stepsize definition $\eta_t \in \frac{\sqrt{C_l}}{C_u}D_t \cdot [1/8, 5/24]$, and by regarding the right hand side as quadratic function of $\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2$, we have

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2 \leq D_t^2 - \frac{1}{3}\eta_t\sqrt{C_l}D_t + \eta_t^2 C_u \leq \left(1 - \frac{1}{50}\frac{C_l}{C_u}\right)^{t+1} \cdot D_0^2,$$

which yields $\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2 \leq \left(1 - \frac{1}{100}\frac{C_l}{C_u}\right)^{t+1} \cdot D_0$.

A.3.2 SECOND PHASE

In the second phase, the expectation of the loss function could be lower bounded with

$$\mathbb{E}\left[f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*)\big|\boldsymbol{\beta}_t\right] \geq \frac{1}{12b_0}C_l\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2.$$

Furthermore, the definition of sub-gradient and event $\mathcal{E}_t$ imply that

$$f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*) \geq \frac{1}{12b_0}C_l\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2 - C\max\{\tau, 1-\tau\}\sqrt{C_u d/\sum_{l=0}^t n_l} \cdot \|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2$$

$$\geq \frac{1}{24b_0}C_l\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2 - C_1 b_0\frac{C_u}{C_l} \cdot \frac{d}{\sum_{l=0}^t n_l},$$

where Cauchy-Schwarz inequality is used in the last line that

$$\frac{1}{24b_0}C_l\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2 + Cb_0\bar{\tau}^2\frac{C_u}{C_l}\frac{d}{\sum_{l=0}^t n_l} \geq \tilde{C}_1\bar{\tau}\sqrt{C_u d/\sum_{l=0}^t n_l} \cdot \|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2,$$

Lemma 16 shows that

$$\|\mathbf{g}_t\|_2^2 \leq 4\frac{1}{b_1^2}C_u^2\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2 + C_2\bar{\tau}^2 C_u \cdot \frac{d}{\sum_{l=0}^t n_l}.$$

Then equation (15) could be upper bounded with

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2 \leq \|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2 - \frac{1}{12b_0}\eta_t C_l n_l\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2 + \eta_t^2\frac{4}{b_1^2}C_u^2\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2$$

$$+ 2\eta_t C_1 b_0\frac{C_u}{C_l}\frac{d}{\sum_{l=0}^t n_l} + C_2\eta_t^2 C_u\frac{d}{\sum_{l=0}^t n_l}.$$

Plug the stepsize $\eta_t \in [c_1, c_2] \cdot \frac{C_l}{C_u^2} \cdot \frac{b_1^2}{b_0}$ into the above equation and then we have

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2 \leq \left(1 - c\frac{b_1^2}{b_0^2} \cdot \frac{C_l^2}{C_u^2}\right) \cdot \|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2^2 + C\frac{d}{C_u\sum_{l=0}^t n_l} \cdot b_1^2,$$

where c, C are some universal constants. In all, we have

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2 \leq \left(1 - c\frac{b_1^2}{b_0^2} \cdot \frac{C_l^2}{C_u^2}\right)^{t+1-t_1} \cdot \|\boldsymbol{\beta}_{t_1} - \boldsymbol{\beta}^*\|_2^2 + C\frac{d}{C_u} \cdot b_1^2 \cdot \sum_{s=t_1}^t \frac{\left(1 - c\frac{b_1^2}{b_0^2} \cdot \frac{C_l^2}{C_u^2}\right)^{t-s}}{\sum_{l=0}^s n_l}.$$

If the number of arriving samples remains to be n, namely, $n_l = n$ for all $l \geq 1$, then equipped with Lemma 19, the estimation error rate is upper bounded with

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2 \leq \left(1 - 2c\frac{b_1^2}{b_0^2} \cdot \frac{C_l^2}{C_u^2}\right)^{t+1-t_1} \cdot \|\boldsymbol{\beta}_{t_1} - \boldsymbol{\beta}^*\|_2^2$$

$$+ \tilde{C}\frac{C_u}{C_l^2} \cdot \frac{d}{n_0 + tn} \cdot \frac{b_0^2}{|\log(1 - c\frac{b_1^2}{b_0^2} \cdot \frac{C_l^2}{C_u^2})|},$$

where $\tilde{C}$ is some constant irrelevant with t. It is worth noting that when $0 < x < 1$, $|\log(1-x)| \geq x$ holds. Thus the last term of the above equation could be further simplified by

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|_2^2 \leq \left(1 - 2c\frac{b_1^2}{b_0^2} \cdot \frac{C_l^2}{C_u^2}\right)^{t+1-t_1} \cdot \|\boldsymbol{\beta}_{t_1} - \boldsymbol{\beta}^*\|_2^2 + \tilde{C}_1\frac{C_u}{C_l^2} \cdot \frac{d}{n_0 + tn} \cdot b_0^2 \cdot \left(\frac{b_0}{b_1} \cdot \frac{C_u}{C_l}\right)^2,$$

with some constant $\tilde{C}_1$.

Appendix B. Technical Lemmas

The following lemma is derivable through integral calculations and can be found the appendix of Shen et al. (2023).

Lemma 12 *Suppose the predictors $\{\mathbf{X}_i\}_{i=1}^n$ satisfy Assumption 1. Then for any fixed vector $\boldsymbol{\beta} \in \mathbb{R}^d$, we have*

$$\sqrt{\frac{C_l}{2\pi}}\|\boldsymbol{\beta}\|_2 \leq \mathbb{E}\rho_{Q,\tau}(\langle\boldsymbol{\beta}, \mathbf{X}_i\rangle) \leq \sqrt{\frac{C_u}{2\pi}}\|\boldsymbol{\beta}\|_2.$$

Lemma 13 *Suppose the predictors and noise term satisfy Assumption 1 and Assumption 2 (a), respectively. And the loss function is based on one observation, $f(\boldsymbol{\beta}) = \rho_{Q,\tau}(Y - \mathbf{X}^\top\boldsymbol{\beta})$. Then for any fixed $\boldsymbol{\beta}$, we have*

$$\mathbb{E}\{f(\boldsymbol{\beta}) - f(\boldsymbol{\beta}^*)\} \geq \sqrt{\frac{C_l}{2\pi}}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 - \gamma,$$

and

$$\mathbb{E}\{f(\boldsymbol{\beta}) - f(\boldsymbol{\beta}^*)\} \geq \frac{C_l}{12b_0}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2, \quad \text{for all } \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \leq \sqrt{C_u^{-1}}\gamma.$$

Proof Regarding the expectation, we consider the expectation conditional on $\mathbf{X}$ first and then consider the expectation with respect to ξ,

$$\mathbb{E}\{f(\boldsymbol{\beta}) - f(\boldsymbol{\beta}^*)\} = \mathbb{E}\left\{\mathbb{E}\left\{\rho_{Q,\tau}(\xi + \mathbf{X}^\top(\boldsymbol{\beta}^* - \boldsymbol{\beta})) - \rho_{Q,\tau}(\xi)\big|\mathbf{X}\right\}\right\}.$$

Specifically, it's

$$\mathbb{E}\left\{\rho_{Q,\tau}(\xi + \mathbf{X}^\top(\boldsymbol{\beta}^* - \boldsymbol{\beta})) - \rho_{Q,\tau}(\xi)\big|\mathbf{X}\right\} = \int_0^{\langle\mathbf{X},\boldsymbol{\beta}-\boldsymbol{\beta}^*\rangle}(H_\xi(t) - H_\xi(0))\,dt$$

Denote $\sigma := ((\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top\mathbf{\Sigma}(\boldsymbol{\beta} - \boldsymbol{\beta}^*))^{1/2}$ and denote $g(\cdot)$ as the density $\langle\mathbf{X}, \boldsymbol{\beta} - \boldsymbol{\beta}^*\rangle$. First, consider the case when $\sigma \leq \gamma$,

$$\begin{aligned}
\mathbb{E}\{f(\boldsymbol{\beta}) - f(\boldsymbol{\beta}^*)\} &= \int_{-\infty}^{+\infty}\int_0^z g(z)\cdot(H_\xi(t) - H_\xi(0))\,dt\,dz \\
&\geq \int_{-\sigma}^{+\sigma}\int_0^z g(z)\cdot(H_\xi(t) - H_\xi(0))\,dt\,dz \geq \int_{-\sigma}^{+\sigma} g(z)\cdot\frac{1}{2b_0}z^2\,dz \\
&= \frac{1}{2b_0}\int_{-\sigma}^{+\sigma}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{z^2}{2\sigma^2}\right)\cdot z^2\,dz \geq \frac{\sigma^2}{12b_0}.
\end{aligned}$$

Then consider another way to find the lower bound, following the triangle inequality and the expectation calculations in Lemma 12,

$$\mathbb{E}\{f(\boldsymbol{\beta}) - f(\boldsymbol{\beta}^*)\} \geq \mathbb{E}\left\{\rho_{Q,\tau}(\mathbf{X}^\top(\boldsymbol{\beta}^* - \boldsymbol{\beta})) - \rho_{Q,\tau}(\xi) - \rho_{Q,\tau}(-\xi)\right\}$$

$$\geq \sqrt{\frac{C_l}{2\pi}} \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| - \gamma,$$

which completes the proof. ∎

Proposition 14 (Proposition 3 in Shen et al. (2023)) *Suppose the loss function is given by the following equation*

$$f(\boldsymbol{\beta}) := \frac{1}{n} \sum_{i=1}^{n} \rho_{Q,\tau}(Y_i - \langle \mathbf{X}_i, \boldsymbol{\beta} \rangle),$$

where $\{\mathbf{X}_i\}_{i=1}^{n}$ follow Assumption 1 and $\{(\mathbf{X}_i, Y_i)\}_{i=1}^{n}$ are independent observations. Then there exist C_1, C_2 such that with probability exceeding $1 - \exp(-C_1 d) - \exp\left(-\sqrt{n/\log n}\right)$, the following holds for all $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{R}^d$,

$$|f(\boldsymbol{\beta}_1) - f(\boldsymbol{\beta}_2) - \mathbb{E}(f(\boldsymbol{\beta}_1) - f(\boldsymbol{\beta}_2))| \leq C_2 \max\{\tau, 1-\tau\} \sqrt{C_u \frac{d}{n}} \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2.$$

Proposition 15 (Concentration with Independence) *Suppose the setting is same as Proposition 14. Then for any fixed $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$, with probability exceeding $1 - 2\exp\left(-C \frac{ns^2}{C_u \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2^2}\right)$ where C is some constant, it has*

$$|f(\boldsymbol{\beta}_1) - f(\boldsymbol{\beta}_2) - \mathbb{E}[f(\boldsymbol{\beta}_1) - f(\boldsymbol{\beta}_2)]| \leq s.$$

Similarly, with probability exceeding $1 - 2\exp\left(-c_1 \frac{ns^2}{C_u \max\{\tau^2, (1-\tau)^2\} \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}^*\|_2^2}\right)$, it has

$$\frac{1}{n} \left| \sum_{i=1}^{n} \langle \boldsymbol{\beta}_1 - \boldsymbol{\beta}^*, \mathbf{X}_i \rangle \cdot g_i - \mathbb{E}\langle \boldsymbol{\beta}_1 - \boldsymbol{\beta}^*, \mathbf{X}_i \rangle \cdot g_i \right| \leq s,$$

where $g_i := -\tau \mathbb{I}_{(\xi_i > \langle \mathbf{X}_i, \boldsymbol{\beta}_1 - \boldsymbol{\beta}^* \rangle)} + (1-\tau) \mathbb{I}_{(\xi_i < \langle \mathbf{X}_i, \boldsymbol{\beta}_1 - \boldsymbol{\beta}^* \rangle)} + \delta \mathbb{I}_{(\xi_i = \langle \mathbf{X}_i, \boldsymbol{\beta}_1 - \boldsymbol{\beta}^* \rangle)}$ is the sub-gradient with any $\delta \in [-\tau, 1-\tau]$.

Proof Note that

$$\left\| \left| |Y_i - \langle \mathbf{X}_i, \boldsymbol{\beta}_1 \rangle| - |Y_i - \langle \mathbf{X}_i, \boldsymbol{\beta}_2 \rangle| \right| \right\|_{\Psi_2} \leq \||\langle \mathbf{X}_i, \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2 \rangle|\|_{\Psi_2} \leq C\sqrt{C_u} \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2.$$

Then by sum of independent sub-Gaussians, we obtain the desired inequality. The second consequence could be obtained in a similar way,

$$\|\langle \boldsymbol{\beta}_1 - \boldsymbol{\beta}^*, \mathbf{X}_i \rangle \cdot g_i\|_{\Psi_2} \leq \max\{\tau, 1-\tau\} \||\langle \boldsymbol{\beta}_1 - \boldsymbol{\beta}^*, \mathbf{X}_i \rangle|\|_{\Psi_2}$$
$$\leq C \max\{\tau, 1-\tau\} \sqrt{C_u} \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2.$$

∎

The following lemma provides two types of upper bound for $\|\mathbf{g}\|_2$ and in the convergence analyses we would choose to use the sharpest one accordingly.

Lemma 16 (Upper Bound of Sub-gradient) *Suppose the predictors and noise term satisfy Assumption 1 and Assumption 2, respectively. And the loss function is based on n independnet observations $f(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\rho_{Q,\tau}(Y_i - \mathbf{X}_i^\top\boldsymbol{\beta})$ with $n \geq Cd$. Suppose*

$$\mathcal{E} := \left\{ \sup_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2\in\mathbb{R}^d} |f(\boldsymbol{\beta}_1) - f(\boldsymbol{\beta}_2) - \mathbb{E}\left[f(\boldsymbol{\beta}_1) - f(\boldsymbol{\beta}_2)\right]|\cdot\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_2\|_2^{-1} \leq C_1\max\{\tau,1-\tau\}\sqrt{\frac{C_u d}{n}} \right\}$$

holds. Then for all $\boldsymbol{\beta}$ and any sub-gradient $\mathbf{g} \in \partial f(\boldsymbol{\beta})$, we have

$$\|\mathbf{g}\|_2 \leq \begin{cases} \sqrt{C_u} \\ 1.5\frac{C_u}{b_1}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 + C_1\max\{\tau,1-\tau\}\sqrt{\frac{C_u d}{n}} \end{cases} .$$

Proof Under the event $\mathcal{E}$ and with sub-gradient definition, for any $\Delta\boldsymbol{\beta} \in \mathbb{R}^d$, we have

$$\langle\Delta\boldsymbol{\beta}, \mathbf{g}\rangle \leq f(\boldsymbol{\beta} + \Delta\boldsymbol{\beta}) - f(\boldsymbol{\beta}) \leq \mathbb{E}f(\boldsymbol{\beta} + \Delta\boldsymbol{\beta}) - \mathbb{E}f(\boldsymbol{\beta}) + C_1\max\{\delta,1-\delta\}\sqrt{C_u\frac{d}{n}}\|\Delta\boldsymbol{\beta}\|_2. \tag{18}$$

We shall provide the two types of upper bounds respectively.

1. Note that the quantile function $\rho_{Q,\tau}(\cdot)$ satisfies $\rho_{Q,\tau}(x_1 + x_2) \leq \rho_{Q,\tau}(x_1) + \rho_{Q,\tau}(x_2)$ and then for any fixed $\boldsymbol{\beta}, \Delta\boldsymbol{\beta}$, we have

$$\mathbb{E}f(\boldsymbol{\beta} + \Delta\boldsymbol{\beta}) - \mathbb{E}f(\boldsymbol{\beta})$$
$$= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\rho_{Q,\tau}(\xi_i - \langle\mathbf{X}_i, \boldsymbol{\beta} + \Delta\boldsymbol{\beta} - \boldsymbol{\beta}^*\rangle) - \rho_{Q,\tau}(\xi_i - \langle\mathbf{X}_i, \boldsymbol{\beta} - \boldsymbol{\beta}^*\rangle)\right]$$
$$\leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\rho_{Q,\tau}(-\langle\mathbf{X}_i, \Delta\boldsymbol{\beta}\rangle).$$

Also notice that $-\langle\mathbf{X}_i, \Delta\boldsymbol{\beta}\rangle$ is a mean zero Gaussian variable and then Assumption 1 implies that $\mathbb{E}\rho_{Q,\tau}(-\langle\mathbf{X}_i, \Delta\boldsymbol{\beta}\rangle) \leq \sqrt{\frac{2}{\pi}}\sqrt{C_u}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2$. Thus together with $n \geq C_3\max\{\tau^2, (1-\tau)^2\}d$ we obtain $\mathbb{E}f(\boldsymbol{\beta} + \Delta\boldsymbol{\beta}) - \mathbb{E}f(\boldsymbol{\beta}) \leq \sqrt{C_u}\|\Delta\boldsymbol{\beta}\|_2$. Then (18) becomes

$$\langle\Delta\boldsymbol{\beta}, \mathbf{g}\rangle \leq \sqrt{C_u}\|\Delta\boldsymbol{\beta}\|_2,$$

holds for all $\boldsymbol{\beta}, \Delta\boldsymbol{\beta}$. Then insert $\Delta\boldsymbol{\beta} = \mathbf{g}$ into the above equation and we acquire

$$\|\mathbf{g}\|_2 \leq \sqrt{C_u}.$$

2. According to the first two equations in the proof of Lemma 11 in Shen et al. (2023), it holds that

$$\mathbb{E}f(\boldsymbol{\beta} + \Delta\boldsymbol{\beta}) - \mathbb{E}f(\boldsymbol{\beta}) \leq \frac{1}{2b_1}C_u\|\Delta\boldsymbol{\beta}\|_2^2 + \frac{1}{b_1}C_u\|\Delta\boldsymbol{\beta}\|_2\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2.$$

Inserting it into (18), and taking $\Delta\boldsymbol{\beta} = \frac{b_1}{4C_u}\mathbf{g}$, we obtain

$$\|\mathbf{g}\|_2 \leq 1.5\frac{C_u}{b_1}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 + C_1\max\{\delta,1-\delta\}\sqrt{\frac{C_u d}{n}}.$$

Lemma 17 (Lemma 6 in Shen et al. (2023)) *Suppose the predictors and noise term satisfy Assumption 1 and Assumption 2(b), respectively. Let $f(\boldsymbol{\beta}) = (1/n)\sum_{i=1}^{n} \rho_{Q,\tau}(Y_i - \mathbf{X}_i^{\top}\boldsymbol{\beta})$ for $n \geq 1$. Then, for any fixed $\boldsymbol{\beta}$,*

$$\mathbb{E}\left[f(\boldsymbol{\beta}) - f(\boldsymbol{\beta}^*)\right] \leq \frac{C_u}{b_1}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2.$$

Lemma 18 *Under Assumptions 1 and 2(b), for any $\boldsymbol{\beta} \in \mathbb{R}^d$ we have*

$$\mathbb{E}\langle\boldsymbol{\beta} - \boldsymbol{\beta}^*, \mathbf{X}\rangle \cdot \partial\rho_{Q,\tau}(\xi - \langle\mathbf{X}, \boldsymbol{\beta} - \boldsymbol{\beta}^*\rangle) \leq \frac{1}{b_1}C_u\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2,$$

where $\partial\rho_{Q,\tau}(u) = -\tau\mathbb{I}_{(u\geq 0)} + (1-\tau)\mathbb{I}_{(u<0)} + \delta\mathbb{I}_{(u=0)}$ is the sub-gradient with $\delta \in [-\tau, 1-\tau]$.

Proof First consider the conditional expectation,

$$\mathbb{E}\left[\langle\boldsymbol{\beta} - \boldsymbol{\beta}^*, \mathbf{X}\rangle \cdot \partial\rho_{Q,\tau}(\xi - \langle\mathbf{X}, \boldsymbol{\beta} - \boldsymbol{\beta}^*\rangle)\big|\mathbf{X}\right] = \langle\boldsymbol{\beta} - \boldsymbol{\beta}^*, \mathbf{X}\rangle \cdot \mathbb{E}\left[\partial\rho_{Q,\tau}(\xi - \langle\mathbf{X}, \boldsymbol{\beta} - \boldsymbol{\beta}^*\rangle)\big|\mathbf{X}\right].$$

Then the following equation can be obtained via some trivial calculations,

$$\mathbb{E}\left[\partial\rho_{Q,\tau}(\xi - \langle\mathbf{X}, \boldsymbol{\beta} - \boldsymbol{\beta}^*\rangle)\big|\mathbf{X}\right] = \int_0^{\langle\mathbf{X}, \boldsymbol{\beta} - \boldsymbol{\beta}^*\rangle} h_\xi(x)\ dx.$$

Denote $z := \langle\mathbf{X}, \boldsymbol{\beta} - \boldsymbol{\beta}^*\rangle$, which follows mean zero Gaussian distributions with variance $(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^{\top}\boldsymbol{\Sigma}(\boldsymbol{\beta} - \boldsymbol{\beta}^*) \leq C_u\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2$. Let $f_z(\cdot)$ be its density. Thus we obtain

$$\mathbb{E}\langle\boldsymbol{\beta} - \boldsymbol{\beta}^*, \mathbf{X}\rangle \cdot \partial\rho_{Q,\tau}(\xi - \langle\mathbf{X}, \boldsymbol{\beta} - \boldsymbol{\beta}^*\rangle) = \mathbb{E}\left[\langle\boldsymbol{\beta} - \boldsymbol{\beta}^*, \mathbf{X}\rangle \cdot \mathbb{E}\left[\partial\rho_{Q,\tau}(\xi - \langle\mathbf{X}, \boldsymbol{\beta} - \boldsymbol{\beta}^*\rangle)\big|\mathbf{X}\right]\right]$$

$$= \int_{-\infty}^{+\infty}\int_0^y y f_z(y) h_\xi(x)\ dx\ dy \leq \frac{1}{b_1}C_u\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2,$$

where the last equation uses the upper bound of the density characterized in Assumption 2. ∎

Lemma 19 *Suppose $0 < a < 1$ is some fixed constant. Define the sequence with $a_n := \sum_{k=1}^{n} \frac{a^{n-k}}{k+m}$, where $m > 0$ is some integer. Then the following holds for each n,*

$$(m+n)a_n \leq \frac{9}{1-a}\frac{1}{|\log a|}.$$

Proof First consider the function $h_a(x) := xa^x$ with the domain $x > 0$. Its first-order derivative is

$$h_a'(x) = a^x - xa^x|\log a| = (1 - |\log a|x)a^x.$$

It infers that $h_a(x)$ achieves the maximum at $x = \frac{1}{|\log a|}$ and $h_a(x)$ is monotone decreasing in the region $x \in (\frac{1}{|\log a|}, +\infty)$. And the $\max_{x>0} h_a(x) = \frac{1}{|\log a|} \exp(\frac{1}{|\log a|} \log a) \leq \frac{3}{|\log a|}$. In this way we have

$$na^{[\frac{n}{2}]} \leq \frac{6}{|\log a|}, \quad \text{for all } n \in \mathbb{Z}_+. \tag{19}$$

Then consider the sequence $(m+n)a_n$,

$$
\begin{aligned}
(m+n)a_n &= 1 + \frac{m+n}{m+n-1}a + \cdots + \frac{m+n}{m+1}a^{n-1} \\
&= \underbrace{1 + \frac{m+n}{m+n-1}a + \cdots + \frac{m+n}{m+n-[\frac{n}{2}]}a^{[\frac{n}{2}]}}_{A_1} \\
&\quad + \underbrace{\frac{m+n}{m+n-[\frac{n}{2}]-1}a^{[\frac{n}{2}]+1} + \cdots + \frac{m+n}{m+1}a^{n-1}}_{A_2}.
\end{aligned}
$$

As for term A_1, when $k \leq [\frac{n}{2}]$ it has $\frac{m+n}{m+n-k} \leq 2$ and then we obtain

$$A_1 \leq 2\left(1 + a + \cdots a^{[\frac{n}{2}]}\right) \leq \frac{2}{1-a}.$$

And A_2 could be bounded with

$$
\begin{aligned}
A_2 &\leq \frac{m+n}{m}a^{[\frac{n}{2}]+1} + \cdots + \frac{m+n}{m}a^{n-1} = \frac{m+n}{m}\left(a^{[\frac{n}{2}]+1} + \cdots + a^{n-1}\right) \\
&\leq \frac{1}{1-a}\left(1 + \frac{n}{m}\right)a^{[\frac{n}{2}]+1} \leq 7\frac{1}{1-a}\frac{1}{|\log a|},
\end{aligned}
$$

where equation (19) is used. Thus altogether, we have $(m+n)a_n \leq \frac{9}{1-a}\frac{1}{|\log a|}$. ∎

The following lemma provides the concentration of Gaussian vectors, an immediate result of Bernstein inequality (Theorem 2.8.1 of Vershynin (2018)).

Lemma 20 *Suppose the random vector $\mathbf{X} \in \mathbb{R}^d$ satisfies Assumption 1. Then with probability exceeding $1 - \exp(-cd)$, its ℓ_2-norm is bounded by $\|\mathbf{X}\|_2^2 \leq 2C_u d$.*