Correlated Noise Provably Beats Independent Noise for Differentially Private Learning

Christopher A. Choquette-Choo* **Krishnamurthy (Dj) Dvijotham***
Krishna Pillutla* **Arun Ganesh**
Thomas Steinke **Abhradeep Guha Thakurta**

Google
*Equal contribution; alphabetical ordering.

Abstract

Differentially private learning algorithms inject noise into the learning where the most common private learning algorithm, DP-SGD, adds independent Gaussian noise in each iteration. Motivated by the practical considerations in federated learning, recent work on matrix factorization mechanisms has shown empirically that introducing correlations in the noise can greatly improve their utility. We characterize the asymptotic objective suboptimality for any choice of the correlation function, giving precise analytical bounds for linear regression. We show, using these bounds, how correlated noise provably improves upon vanilla DP-SGD as a function of problem parameters such as the effective dimension and condition number. Moreover, our analytical expression for the near-optimal correlation function circumvents the cubic complexity of the semi-definite program used to optimize the noise correlation in prior work. We validate these theoretical results with experiments on private deep learning in both centralized and federated settings. Our work matches or outperforms prior work while being efficient both in terms of computation and memory.

1 Introduction

The broad adoption of deep learning using sensitive data has led to the increasing popularity of rigorous frameworks for privacy preservation, such as differential privacy [20]. The workhorse of private learning, a differentially private variant of stochastic gradient descent called DP-SGD [2, 9, 48], clips per-example gradients to some ℓ_2 norm and adds *independent* Gaussian noise. DP-SGD has been used in a range of applications from learning with medical images [3] to finetuning large language models with $O(100B)$ parameters [22].

A recent line of work instead proposes a family of algorithms called DP-FTRL that add *correlated* Gaussian noise to each clipped gradient [17, 19, 33, 47]. This work is motivated by the fact that DP-SGD and its federated variant DP-FedAvg [41] require amplification via sampling/shuffling to achieve competitive privacy-utility tradeoffs. This in turn requires uniformly random data samples, an assumption that is violated in federated learning due to complicated client availability [32].

DP-FTRL, on the other hand, has been showing to compete with amplified DP-SGD, and has been used for private federated learning at industrial scale [40, 50]. In fact, by solving an expensive semi-definite program (SDP) to find the noise correlations, Choquette-Choo et al. [16] demonstrated *empirically* that DP-FTRL is never worse and often *much better* than DP-SGD in its privacy-utility tradeoff across multiple modalities.

International Workshop on Federated Learning in the Age of Foundation Models in Conjunction with NeurIPS 2023 (FL@FM-NeurIPS'23). This workshop does not have official proceedings and this paper is non-archival.

Table 1: Comparison of Noisy-FTRL (and Noisy-SGD), the variants of DP-FTRL (and DP-SGD) without gradient clipping. We give their asymptotic suboptimality for linear regression with Gaussian inputs with covariance $\boldsymbol{H}$ and noise multiplier $\sigma_{\mathsf{dp}}^2 = \gamma_\infty(\boldsymbol{\beta})^2/(2\rho)$ based on the limiting sensitivity (defined in §2) in terms of the learning rate η, dimension d, the effective dimension $d_{\mathsf{eff}} = \mathsf{Tr}\left[\boldsymbol{H}\right]/\lambda_{\max}\boldsymbol{H}$, the strong convexity $\mu = \lambda_{\min}\left(\boldsymbol{H}\right)$ and the noise variance ρ^{-1} representing the privacy level. We take the gradient norm $G = 1$ (used to scale the noise) and $\lambda_{\max}\left(\boldsymbol{H}\right) = 1$ w.l.o.g. and only show the term depending on ρ. Since $1 \leq d_{\mathsf{eff}} \leq d$, Noisy-FTRL is better than Noisy-SGD at smaller learning rates η or when the effective dimension d_{eff} is small (e.g., when $\boldsymbol{H}$ is close to low rank).

Algorithm	**Asymptotic Suboptimality** F_∞	**Ratio w/ Lower Bound**	**Remark**
Lower Bound	$\Omega\left(\eta^2\rho^{-1}d_{\mathsf{eff}}\right)$	1	for all $\boldsymbol{\beta}$ with finite $\left\|\boldsymbol{\beta}\right\|_1$
Noisy-SGD	$\Theta\left(\eta\rho^{-1}d\right)$	$\frac{d}{\eta d_{\mathsf{eff}}}$	$\Theta(\cdot)$ denotes matching upper & lower bounds
Noisy-FTRL	$O\left(\eta^2\rho^{-1}d_{\mathsf{eff}}\log^2\frac{1}{\eta\mu}\right)$	$\log^2\frac{1}{\eta\mu}$	The bound is attained for $\boldsymbol{\beta}$ in (7)

However, several questions remain open. Does DP-FTRL **provably improve** over DP-SGD in its expected utility? Further, can we design a more **computationally efficient** procedure to find the noise correlations for DP-FTRL without significantly worsening the privacy-utility tradeoff?

We answer both questions affirmatively by (1) providing a sharp theoretical characterization of the noisy training dynamics of DP-FTRL (summarized in Table 1), and (2) leveraging these analytical tools to circumvent the SDP required in past work. Our experiments on private image classification and language modeling demonstrate that the proposed approach is competitive with the state-of-the-art correlated noise approaches while being significantly more efficient.

2 Problem Setup and Background

Let $\mathcal{D} = \{\boldsymbol{z}_0, \ldots, \boldsymbol{z}_{T-1}\}$ be a dataset of T datapoints, where each datapoint is sampled i.i.d. from an underlying distribution $\mathbb{P}_{\mathsf{data}}$. Our learning objective is to minimize:

$$F\left(\boldsymbol{\theta}\right) = \mathbb{E}_{\boldsymbol{z}\sim\mathbb{P}_{\mathsf{data}}}\left[f\left(\boldsymbol{\theta};\boldsymbol{z}\right)\right] + r\left(\boldsymbol{\theta}\right), \tag{1}$$

where $f(\boldsymbol{\theta};\boldsymbol{z})$ is the loss incurred by model parameters $\boldsymbol{\theta} \in \mathbb{R}^d$ on a datapoint $\boldsymbol{z}$, and $r(\cdot)$ is data-independent regularization. We aim to minimize F while satisfying differential privacy with respect to the dataset $\mathcal{D}$. We assume that F has a unique minimizer denoted $\boldsymbol{\theta}_\star$.

We focus on variants of stochastic gradient descent with a batch size of 1 for data arriving in a stream.[1] DP-FTRL with a noise correlation matrix $\boldsymbol{B} \in \mathbb{R}^{T\times T}$ (which is lower triangular) iterates

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta\left(\mathsf{clip}\left(\nabla f(\boldsymbol{\theta}_t;\boldsymbol{z}_t), G\right) + \nabla r(\boldsymbol{\theta}_t) + \textstyle\sum_{\tau=0}^{t}\boldsymbol{B}_{t,\tau}\boldsymbol{w}_\tau\right) \tag{2}$$

for Gaussian noise $\boldsymbol{w}_t \sim \mathcal{N}(\boldsymbol{0}, \sigma_{\mathsf{dp}}^2 G^2 \boldsymbol{I}_d)$, where $\mathsf{clip}\left(\cdot, G\right)$ denotes projection onto an ℓ_2 ball of radius G. We define Noisy-FTRL to be DP-FTRL without clipping. Taking $\boldsymbol{B} = \boldsymbol{I}$ as the identity matrix recovers DP-SGD (with clipping) and Noisy-SGD (without clipping), and other choices give rise to alternate algorithms. We restate a result from prior work showing that DP-FTRL is differentially private for any choice of $\boldsymbol{B}$, provided the noise multiplier is scaled appropriately.

Theorem 2.1 ([10, 19])**.** *DP-FTRL* (2) *satisfies ρ-zero concentrated differential privacy (zCDP) if the $\sigma_{\mathsf{dp}}^2 = \gamma_T^2(\boldsymbol{B})/(2\rho)$ where $\gamma_T\left(\boldsymbol{B}\right) = \max_{t<T}\|(\boldsymbol{B}^{-1})_{:,t}\|_2$ is the sensitivity of $\boldsymbol{B}^{-1}$.*

Although Noisy-FTRL is not differentially private, it lets us analyze the noise dynamics of DP-FTRL without technicalities associated with clipping. We sharply characterize the asymptotic utility of Noisy-FTRL for linear regression and use this to give bounds for DP-FTRL.

Provable separation between DP-SGD and DP-FTRL: The best-known separation between DP-SGD and DP-FTRL in the literature is due to Kairouz et al. [33]. For G-Lipschitz convex losses, DP-FTRL at a privacy level of ρ-zCDP achieves a suboptimality of $O(Gd^{1/4}/\sqrt{\rho T})$ compared to DP-SGD's $O(Gd^{1/4}/\sqrt{\rho^2 T})$. The only improvement here is in terms of the privacy parameter ρ.

[1] We focus on the centralized setting with a batch size of 1 for simplicity. The analysis extends straightforwardly to larger batch sizes.

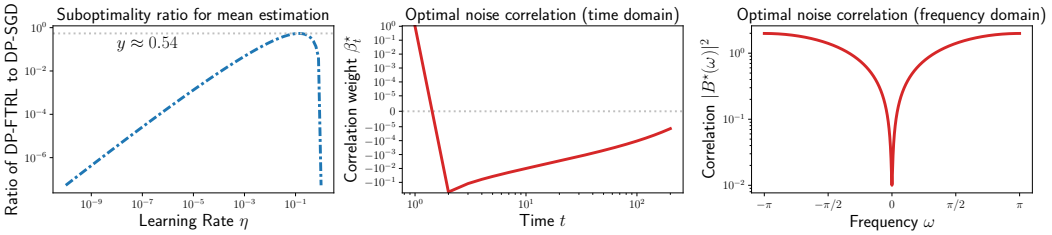

Figure 1: **Left**: The ratio of the asymptotic suboptimalities of DP-FTRL to DP-SGD for mean estimation vs. the learning rate η. DP-FTRL is never worse but is orders of magnitude better at $\eta \to 0$ or 1. **Middle/Right**: Time- and frequency-domain descriptions of the optimal correlations for mean estimation (cf. Theorem 3.1).

This theory fails to reflect the large margin by which DP-FTRL empirically outperforms DP-SGD across the board [16], and a precise characterization is missing.

Computationally efficient DP-FTRL: Prior work on DP-FTRL utilizes the noise correlation matrix $\boldsymbol{B}$ that minimizes the squared error in the *gradient prefix sums* [19, 33]:

$$\varphi(\boldsymbol{B}) = \sum_{t=0}^{T-1} \mathbb{E}\left\| \sum_{\tau=0}^{t} \widetilde{\boldsymbol{g}}_t - \sum_{\tau=0}^{t} \boldsymbol{g}_t \right\|_2^2 \tag{3}$$

where $\boldsymbol{g}_t$ is the clipped gradient applied in iteration t and $\widetilde{\boldsymbol{g}}_t$ is its noisy counterpart (cf. Algorithm 1). This was, in turn, obtained as an upper bound on the regret in an adversarial online learning setting [33, Thm. C.1]. The most potent algorithm from the previous work gave $\boldsymbol{B}$ as the solution of a semidefinite program with matrix variables of size $O(T^2)$, requiring $O(T^3)$ time [19, Eq. 4]. This cost is prohibitive for large learning problems. Moreover, there is a mismatch between the objective (3) used to find the noise correlations and the final learning objective $F(\boldsymbol{\theta}_T)$. In particular, Koloskova et al. [35] give two noise correlation matrices $\boldsymbol{B}_1, \boldsymbol{B}_2$ with equal squared error $\varphi(\boldsymbol{B}_1) = \varphi(\boldsymbol{B}_2)$ such that DP-FTRL with $\boldsymbol{B}_1$ diverges while DP-FTRL with $\boldsymbol{B}_2$ converges.

Our approach: We study the suboptimality in the final objective $\mathbb{E}[F(\boldsymbol{\theta}_T) - F(\boldsymbol{\theta}_\star)]$. We work in the asymptotic $T \to \infty$ regime to allow the use of analytic tools, but also to derive results that apply regardless of the dataset size. Second, we restrict the search over $\boldsymbol{B}$ to *Toeplitz* matrices $\boldsymbol{B}_{t,\tau} = \beta_{t-\tau}$ generated by a sequence $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots)$ of reals, but a stronger motivation is that they are **anytime**, i.e., they do not be recomputed for each value of T and easily apply as $T \to \infty$. Toeplitz $\boldsymbol{B}$ were previously considered for their computational efficiency in learning [17] and for optimal ℓ_2 error (including constants) in linear counting queries [24]. Thus, our goal is to characterize the asymptotic suboptimality

$$F_\infty(\boldsymbol{\beta}) := \lim_{T\to\infty} \mathbb{E}\left[F\left(\boldsymbol{\theta}_T\right) - F(\boldsymbol{\theta}_\star)\right] \tag{4}$$

for $\boldsymbol{\theta}_T$ produced by Noisy-FTRL or DP-FTRL under noise correlation weights $\boldsymbol{\beta}$. We analyze this in the frequency domain using the **discrete-time Fourier transform** $B(\omega) = \sum_{t=0}^{\infty} \boldsymbol{\beta}_t \exp(i\omega t)$, with i the imaginary unit. Further, we define the limiting sensitivity associated with B as the limiting value of γ_T. which, using standard Fourier analysis tools, can be expressed as

$$\gamma_\infty\left(B\right)^2 := \lim_{T\to\infty} \gamma_T\left(B\right)^2 = \tfrac{1}{2\pi} \int_{-\pi}^{\pi} \left|B\left(\omega\right)\right|^{-2} \, \mathrm{d}\omega \,. \tag{5}$$

3 Conceptual Overview: Private Mean Estimation in One Dimension

We begin with the simplest objective function, the squared error for a mean estimation problem on the real line. This setting captures the core intuition and ideas used to derive further results. We emphasize that we do not aim to derive optimal rates for private mean estimation (or later, for private linear regression). Instead, our goal is to demonstrate a separation between DP-FTRL and DP-SGD.

Consider a distribution $\mathbb{P}_{\mathsf{data}}$ with $|z - \mathbb{E}[z]| \le \sigma_{\mathsf{sgd}}$ and $|z| \le 1$ a.s. for $z \sim \mathbb{P}_{\mathsf{data}}$. Our objective now is to estimate the mean by minimizing the objective

$$F\left(\theta\right) = \tfrac{1}{2}\, \mathbb{E}_{z\sim\mathbb{P}_{\mathsf{data}}}(\theta - z)^2 \quad \text{with} \quad f\left(\theta; z\right) = \tfrac{z^2}{2} - z\theta, \quad \text{and} \quad r\left(\theta\right) = \tfrac{\theta^2}{2} \,. \tag{6}$$

We show a strict separation between DP-FTRL and DP-SGD for this simple minimization problem.

Theorem 3.1. *Consider the setting above with learning rate $\eta \leq 1$ and clip norm $G = 1$. Then, the asymptotic suboptimality of a ρ-zCDP sequence $(\theta_t)_{t=0}^{\infty}$ obtained via DP-SGD is $F_\infty(\boldsymbol{\beta}_{\mathsf{dpsgd}}) = \Theta(\eta\rho^{-1} + \eta\sigma_{\mathsf{sgd}}^2)$. Further, the asymptotic suboptimality of any ρ-zCDP sequence $(\theta_t)_{t=0}^{\infty}$ from DP-FTRL is*

$$\inf_{\boldsymbol{\beta}} F_\infty(\boldsymbol{\beta}) = F_\infty(\boldsymbol{\beta}^\star) = \Theta\left(\eta^2\rho^{-1}\log^2(1/\eta) + \eta\sigma_{\mathsf{sgd}}^2\right).$$

The infimum above is attained by $\beta_t^\star = (-1)^t\binom{1/2}{t}(1-\eta)^t$, where $\binom{1/2}{t} = \prod_{k=0}^{t-1}\frac{1/2-k}{t-k}$.

Proof Sketch. Using tools from frequency domain analysis of linear time invariant systems [43], we can show that the asymptotic variance $\mathbb{E}(\theta_t - \mathbb{E}[z])^2$ is an integral of $|B(\omega)|^2$. The sensitivity is an integral of $|B(\omega)|^{-2}$ (cf. (5)) so that F_∞ is a product of these integrals. Its minimizer $B^\star$ can be analytically computed in the Fourier domain (Figure 1, right). An inverse Fourier transform yields the claimed expression for $\boldsymbol{\beta}^\star$ (Figure 1, center). □

The optimal ρ^{-1} coefficient is $\eta^2\log^2(1/\eta)$ improves over DP-SGD's η (Figure 1, left).

ν-DP-FTRL/ν-Noisy-FTRL: Theorem 3.1 gives an analytical expression for the optimal noise correlation weights for DP-FTRL for this simplified setting. Using a parameter $0 < \nu < 1$, we define

$$\hat{\beta}_t^\nu := (-1)^t\binom{1/2}{t}(1-\nu)^t. \tag{7}$$

We analyze this choice theoretically for the setting of Noisy-FTRL and demonstrate near optimality. Later, for our experiments with DP-FTRL, we tune ν as a hyperparameter to tune. We call this approach (with clipping) ν-DP-FTRL and (without clipping) ν-Noisy-FTRL.

4 Analysis for Linear Regression

We now give a precise analysis of F_∞ for linear regression with ν-Noisy-FTRL. We consider (unregularized) linear regression with loss function $f(\boldsymbol{\theta}; (\boldsymbol{x}, y)) = \frac{1}{2}\left(y - \langle\boldsymbol{\theta}, \boldsymbol{x}\rangle\right)^2$

$$F(\boldsymbol{\theta}) = \tfrac{1}{2}\,\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}_{\mathsf{data}}}\left(y - \langle\boldsymbol{\theta}, \boldsymbol{x}\rangle\right)^2. \tag{8}$$

We assume d-dimensional Gaussian covariates $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{H})$ and independent Gaussian residuals $y - \langle\boldsymbol{\theta}_\star, \boldsymbol{x}\rangle \sim \mathcal{N}(0, \sigma_{\mathsf{sgd}}^2)$ where $\boldsymbol{\theta}_\star = \arg\min F$. We make these assumptions for ease of presentation; we state and prove our results under weaker assumptions in the supplement. Further, we assume that F is L-smooth and μ-strongly convex (equivalently, $\mu\boldsymbol{I} \preceq \boldsymbol{H} \preceq L\boldsymbol{I}$ since the input covariance $\boldsymbol{H}$ is also the Hessian of F). We express the bounds on F_∞ in terms of the correlation weights $\boldsymbol{\beta}$ and the problem parameters ρ, G which, for DP-FTRL, denote the target privacy level and the gradient clip norm respectively. See §C for proofs.

Theorem 4.1. *Let c, C_1, C_2 denote universal constants. For $\eta \leq c/\mathsf{Tr}[\boldsymbol{H}]$, we have*

$$\begin{aligned}
&\textbf{\textit{(Noisy-SGD)}} && F_\infty(\boldsymbol{\beta}^{\mathsf{sgd}}) = \Theta\left(\eta d G^2\rho^{-1} + \eta\sigma_{\mathsf{sgd}}^2\mathsf{Tr}[\boldsymbol{H}]\right) && \textit{with } \boldsymbol{\beta}^{\mathsf{sgd}} = (1, 0, \ldots),\\
&\textbf{\textit{(ν-Noisy-FTRL)}} && F_\infty(\hat{\boldsymbol{\beta}}^\nu) \leq C_1\left(\eta^2G^2\rho^{-1}\log^2\tfrac{1}{\nu} + \eta\sigma_{\mathsf{sgd}}^2\right)\mathsf{Tr}[\boldsymbol{H}] && \textit{with } \nu \leq \eta\mu, \textit{ and}\\
&\textbf{\textit{(Lower bound)}} && F_\infty(\boldsymbol{\beta}) \geq C_2\left(\eta^2G^2\rho^{-1} + \eta\sigma_{\mathsf{sgd}}^2\right)\mathsf{Tr}[\boldsymbol{H}] && \textit{for all } \boldsymbol{\beta} \textit{ with } \|\boldsymbol{\beta}\|_1 < \infty.
\end{aligned}$$

Observe that our bounds separate the contributions arising from correlated noise (ρ^{-1} term) and those from the inherent noise in the linear model (σ_{sgd}^2 term). We focus on the effect of correlation because the effect of latter noise is the same across all choices of $\boldsymbol{\beta}$. We plot this in Figure 2.

Exponential separation between Noisy-SGD and Noisy-FTRL: Noisy-SGD's stationary error depends on the ambient dimension d, while the lower bound depends on the *effective dimension* $d_{\mathsf{eff}} = \mathsf{Tr}[\boldsymbol{H}]/\|\boldsymbol{H}\|_2$ of the covariance $\boldsymbol{H}$. We have, $d_{\mathsf{eff}} \leq d$ with equality when all the eigenvalues of $\boldsymbol{H}$ are equal but $d_{\mathsf{eff}} \ll d$ when the eigenvalues of $\boldsymbol{H}$ decay rapidly or it is nearly low rank. This is true particularly for overparameterized models where the features may be highly correlated resulting in an approximately low-rank covariance. For instance, if the eigenvalues of $\boldsymbol{H}$ are $(1, 1/d, \ldots, 1/d)$, then $d_{\mathsf{eff}} \leq 2$. Then, Noisy-FTRL's error of $O(\eta^2\rho^{-1}\log^2(d/\eta))$ is exponentially better than Noisy-SGD's $\Theta(\eta\rho^{-1}d)$. A similar advantage also holds when eigenvalues of $\boldsymbol{H}$ decay

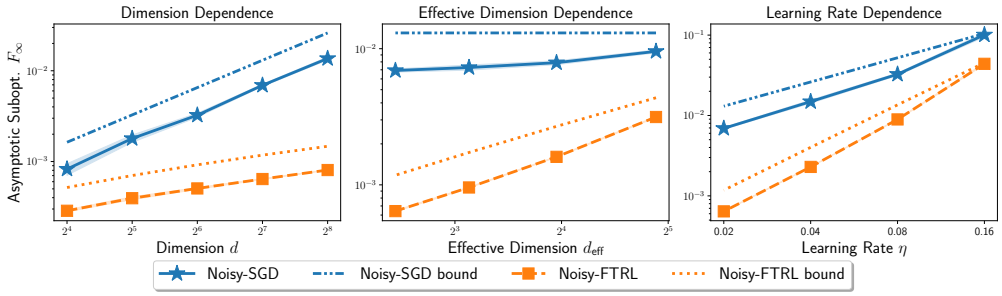

Figure 2: **Linear regression simulations**: We plot the empirically observed asymptotic suboptimality of ν-Noisy-FTRL/Noisy-SGD and their theoretical bounds with $d = 128$ (varied in the left plot) where the Hessian $\boldsymbol{H}$ has eigenvalues $\lambda_k = 1/k$ (varied as $k^{-\alpha}$ for $\alpha \in [0.4, 1]$ in the middle plot), and learning rate $\eta = 0.02$ (varied in the right plot). The **slope** of the corresponding empirical and theoretical lines are nearly equal, showing the **tightness of the theory**. In particular, we observe that Noisy-SGD has a linear dependence on the dimension (slope 1.00) and is nearly constant w.r.t. the effective dimension (slope 0.18) while Noisy-FTRL has a near-linear dependence on the effective dimension (slope 0.94). Noisy-FTRL (slope 2.03) also has a better dependence on the learning rate than Noisy-SGD (slope 1.27).

at various rates (see Table 2 in §C). The learning rate dependence of Noisy-SGD is also suboptimal, similar to §3. This result is also confirmed empirically in Figure 2 (right).

Finite-time privacy-utility analysis: Noisy-FTRL, which we analyzed so far, is not differentially private. Differential privacy requires gradient clipping that significantly complicates the analysis. However, for a finite time horizon T, we can argue using concentration that $\nabla f(\boldsymbol{\theta}; \boldsymbol{z})$ is bounded with high probability and clipping can be avoided. Formal statements and proofs for the finite-time analysis are given in Appendix D.

Consider DP-FTRL with noise correlation $\hat{\boldsymbol{\beta}}^\nu$ from (7) with $\nu = \eta\mu$ and gradients clipped to any ℓ_2-norm G. As mentioned in §2, the outputs $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_T)$ of DP-FTRL are ρ-zCDP. For an appropriate η, we give utility bounds in terms of the effective dimension d_{eff} and the condition number $\kappa = L/\mu$:

(a) For η small enough, we have with probability at least $1 - p$ that

$$\max_{t<T} \|\boldsymbol{g}_t\|_2 \le c \max\left\{ \mathsf{Tr}\left[\boldsymbol{H}\right] \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_\star\|_2, \sigma_{\mathsf{sgd}} \sqrt{\mathsf{Tr}\left[\boldsymbol{H}\right]} \right\} \mathrm{polylog}\left(T/p\right) =: G\,. \tag{9}$$

Let $\mathcal{E}$ denote this event. If it holds, no gradients are clipped and DP-FTRL coincides with Noisy-FTRL.

(b) For $T \ge \widetilde{\Omega}(\kappa^2 d_{\text{eff}}^2 d/\rho)$, we have (omitting log factors and $o(1/T^2)$ terms and assuming $\|\boldsymbol{H}\|_2 = 1$):

$$\mathbb{E}\left[(F(\boldsymbol{\theta}_t) - F(\boldsymbol{\theta}_\star)) \cdot \mathbb{1}\left(\mathcal{E}\right)\right] \lesssim \begin{cases} \kappa\, d_{\text{eff}} \left(\frac{d d_{\text{eff}} \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_\star\|_2^2}{\rho T} + \frac{d\sigma_{\mathsf{sgd}}^2}{\rho T} + \frac{\sigma_{\mathsf{sgd}}^2}{T} \right) & \text{for DP-SGD,} \\ \kappa d_{\text{eff}} \left(\frac{\kappa d_{\text{eff}}^2 \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_\star\|_2^2}{\rho T^2} + \frac{\kappa d_{\text{eff}} \sigma_{\mathsf{sgd}}^2}{\rho T^2} + \frac{\sigma_{\mathsf{sgd}}^2}{T} \right) & \text{for } \nu\text{-DP-FTRL.} \end{cases}$$

Thus, the dimension d in DP-SGD's bound effectively becomes $\kappa d_{\text{eff}}/T$ for DP-FTRL, leading to a better dimension dependence. While faster $1/(\rho T^2)$ rates are known for DP-SGD-style algorithms for linear regression [39, 49], they require sophisticated adaptive clipping strategies. Our algorithms use a fixed clipping norm G and a fixed noise multiplier σ_{dp} independent of T; the bounds presented above are, to the best of our knowledge, the best known in the literature for DP-SGD in this setting. We leave the exploration of combining adaptive clipping with correlated noise for future work.

Extensions to this analysis: We extend this analysis in several directions:

- We give bounds on the asymptotic suboptimality of other algorithms such as anti-PGD [44] that fall into the Noisy-FTRL framework; see Table 3 in §C.
- We give bounds on the asymptotic suboptimality for general strongly convex functions as the solution to a second-order cone program in §E. These bounds show that ν-DP-FTRL has a better asymptotic error as a function of the condition number than DP-SGD.

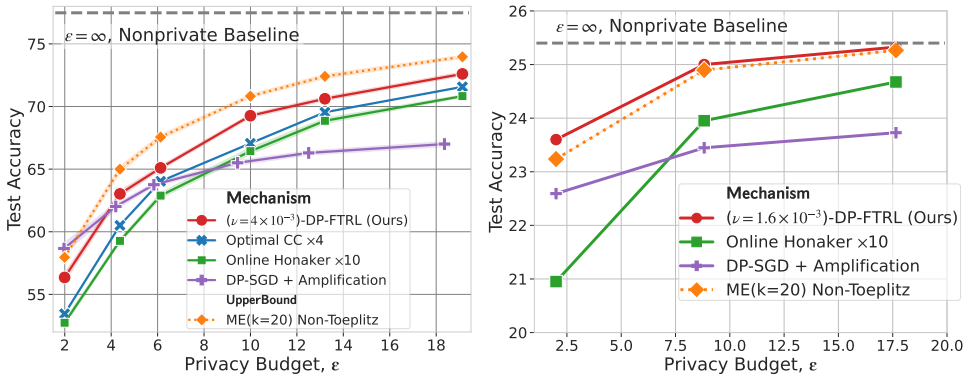

Figure 3: **Left**: Example-level DP on CIFAR-10 (non-federated image classification). **Right**: User-level DP on StackOverflow (federated language modeling). **Our ν-DP-FTRL matrices outperform all other efficient and anytime mechanisms.** They also achieve or slightly outperform state-of-the-art mechanisms that require significantly more compute (cf. Table 4 in Appendix G).

5 Experiments

We demonstrate the practical benefits and versatility of ν-DP-FTRL for two deep learning tasks across modalities and settings: (1) example-level DP for image classification on CIFAR-10 (centralized learning), and (2) federated user-level DP for language modeling with the StackOverflow dataset. The detailed setup is given in §G.

Baselines: In addition to DP-SGD (with amplification), we compare ν-DP-FTRL (per-step cost = T) with the following anytime DP-FTRL baselines: (a) the TreeAgg/Online Honaker method [33] (per-step cost $\log T$), and (b) OptimalCC [24] (per-step cost T)), which corresponds to (7) with $\nu = 0$. In addition, we compare to "ME" the state-of-the-art full DP-FTRL approach [17] where computing the $\boldsymbol{B}$ matrix takes $O(T^3)$ cost and the per-step training cost is $O(T^2)$. This is infeasible[2] as T becomes large in modern models [5, 34], but we use it as the ceiling of DP-FTRL without amplification. Finally, we stamp/restart all baselines as suggested in [17]. This gives the baselines an advantage of an additional tuning parameter (tuned to minimize the squared error (3)), but does not affect their per-step training cost. We denote this by the suffix "$\times S$" for $S > 1$ in the plot.

Main Results: Across both datasets, ν-DP-FTRL outperforms all existing anytime mechanisms by a significant margin (Figure 3, left). We find an average 3pp improvement that grows as ε becomes small. Indeed, the proposed ν-DP-FTRL makes up 30-80% of the gap between previous efficient approaches and the state-of-the-art and computationally intense ME approach. For instance, at $\varepsilon = 10$, we have ν-DP-FTRL at 69.26% nearly matches ME at 70.83%. For StackOverflow, we find that ν-DP-FTRL matches the state-of-the-art ME at $\varepsilon = 8$ and slightly exceeds it at $\varepsilon = 2$ (Figure 3 right; 23.6% for our approach vs. 23.2% at $\varepsilon = 2$).

As ε becomes small, DP-SGD can outperform DP-FTRL due to privacy amplification. We find that ν-DP-FTRL outperforms DP-SGD for $\varepsilon \geq 4$ on CIFAR-10 (63.02% vs. 62.02%) and around $\varepsilon \approx 2$ for StackOverflow (23.6% versus 22.6%), showing its broad applicability. Finally, we observe that that our mechanism achieves near non-private baselines on StackOverflow. A model trained via ν-DP-FTRL gets 24.8% validation accuracy at $\varepsilon = 8$, a mere 0.6% off from the nonprivate baseline.

6 Conclusion

This work shows a clear separation between the noisy training dynamics with uncorrelated (DP-SGD) and correlated noise (DP-FTRL) for linear regression. The matching upper/lower bounds reveal that DP-FTRL has a better dependence than DP-SGD on problem parameters such as the effective dimension and condition number. Inspired by the theory, we propose ν-DP-FTRL and validated its empirical performance on two DP tasks spanning image and language modalities. We

[2]Generating $\boldsymbol{B}$ for $T = 10^4$ takes around 24 hours [17].

found it can compete the state-of-the-art while circumventing the need for any expensive computations like the semi-definite programs used in prior work. This work opens up several exciting directions including leveraging correlated-noise mechanisms for instance-optimal bounds and further improving the computational efficiency to enable private training of foundation models.

Acknowledgements

The authors thank H. Brendan McMahan, Fabian Pedregosa, Ian R. Manchester, Keith Rush, and Rahul Kidambi for fruitful discussions and helpful comments.

References

[1] *NIST Digital Library of Mathematical Functions*. `https://dlmf.nist.gov/`, Release 1.1.10 of 2023-06-15. URL `https://dlmf.nist.gov/`. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds.

[2] Martín Abadi, Andy Chu, Ian J. Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proc. of the 2016 ACM SIGSAC Conf. on Computer and Communications Security (CCS'16)*, pp. 308–318, 2016.

[3] Mohammed Adnan, Shivam Kalra, Jesse C Cresswell, Graham W Taylor, and Hamid R Tizhoosh. Federated learning and differential privacy for medical image analysis. *Scientific reports*, 12(1):1953, 2022.

[4] Rafik Aguech, Eric Moulines, and Pierre Priouret. On a Perturbation Approach for the Analysis of Stochastic Tracking Algorithms. *SIAM J. Control. Optim.*, 39(3):872–899, 2000.

[5] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

[6] Julyan Arbel, Olivier Marchal, and Hien D Nguyen. On strict sub-Gaussianity, optimal proxy variance and symmetry for bounded random variables. *ESAIM: Probability and Statistics*, 24: 39–55, 2020.

[7] Francis R. Bach and Eric Moulines. Non-Strongly-Convex Smooth Stochastic Approximation with Convergence Rate $O(1/n)$. In *NeurIPS*, pp. 773–781, 2013.

[8] Borja Balle, Gilles Barthe, Marco Gaboardi, Justin Hsu, and Tetsuya Sato. Hypothesis Testing Interpretations and Rényi Differential Privacy. In *AISTATS*, pp. 2496–2506, 2020.

[9] Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Proc. of the 2014 IEEE 55th Annual Symp. on Foundations of Computer Science (FOCS)*, pp. 464–473, 2014.

[10] Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pp. 635–658. Springer, 2016.

[11] Paul F Byrd and Morris D Friedman. *Handbook of Elliptic Integrals for Engineers and Scientists*, volume 67. Springer, 2013.

[12] Andrea Caponnetto and Ernesto De Vito. Optimal Rates for the Regularized Least-Squares Algorithm . *Foundations of Computational Mathematics*, 7:331–368, 2007.

[13] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pp. 267–284, 2019.

[14] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

[15] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*, 2022.

[16] Christopher A Choquette-Choo, Arun Ganesh, Ryan McKenna, H Brendan McMahan, Keith Rush, Abhradeep Guha Thakurta, and Zheng Xu. (amplified) banded matrix factorization: A unified approach to private training. *arXiv preprint arXiv:2306.08153*, 2023. URL https://arxiv.org/abs/2306.08153.

[17] Christopher A. Choquette-Choo, Hugh Brendan McMahan, J Keith Rush, and Abhradeep Guha Thakurta. Multi-Epoch Matrix Factorization Mechanisms for Private Machine Learning. In *ICML*, volume 202, pp. 5924–5963, 23–29 Jul 2023.

[18] Alexandre Défossez and Francis R. Bach. Averaged Least-Mean-Squares: Bias-Variance Trade-offs and Optimal Sampling Distributions. In *AISTATS*, volume 38, 2015.

[19] Sergey Denisov, H Brendan McMahan, John Rush, Adam Smith, and Abhradeep Guha Thakurta. Improved Differential Privacy for SGD via Optimal Private Linear Operators on Adaptive Streams. *NeurIPS*, 35:5910–5924, 2022.

[20] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating Noise to Sensitivity in Private Data Analysis. In *Proc. of the Third Conf. on Theory of Cryptography (TCC)*, pp. 265–284, 2006. URL http://dx.doi.org/10.1007/11681878_14.

[21] Hendrik Fichtenberger, Monika Henzinger, and Jalaj Upadhyay. Constant Matters: Fine-grained Error Bound on Differentially Private Continual Observation. In *ICML*, 2023.

[22] Jiyan He, Xuechen Li, Da Yu, Huishuai Zhang, Janardhan Kulkarni, Yin Tat Lee, Arturs Backurs, Nenghai Yu, and Jiang Bian. Exploring the Limits of Differentially Private Deep Learning with Group-wise Clipping. In *ICLR*, 2023.

[23] William Paul Heath and Adrian G Wills. Zames-Falb multipliers for quadratic programming. In *Proceedings of the 44th IEEE Conference on Decision and Control*, pp. 963–968. IEEE, 2005.

[24] Monika Henzinger, Jalaj Upadhyay, and Sarvagya Upadhyay. A unifying framework for differentially private sums under continual observation. *arXiv preprint arXiv:2307.08970*, 2023.

[25] James Honaker. Efficient use of differentially private binary trees. *Theory and Practice of Differential Privacy (TPDP 2015), London, UK*, 2015.

[26] Daniel J. Hsu, Sham M. Kakade, and Tong Zhang. Random Design Analysis of Ridge Regression. *Found. Comput. Math.*, 14(3):569–600, 2014.

[27] Daphne Ippolito, Florian Tramèr, Milad Nasr, Chiyuan Zhang, Matthew Jagielski, Katherine Lee, Christopher A Choquette-Choo, and Nicholas Carlini. Preventing verbatim memorization in language models gives a false sense of privacy. *arXiv preprint arXiv:2210.17546*, 2022.

[28] Palak Jain, Sofya Raskhodnikova, Satchit Sivakumar, and Adam Smith. The Price of Differential Privacy under Continual Observation. In *ICML*, pp. 14654–14678. PMLR, 2023.

[29] Prateek Jain, Sham M. Kakade, Rahul Kidambi, Praneeth Netrapalli, Venkata Krishna Pillutla, and Aaron Sidford. A Markov Chain Theory Approach to Characterizing the Minimax Optimality of Stochastic Gradient Descent (for Least Squares). In *FSTTCS*, volume 93, pp. 2:1–2:10, 2017.

[30] Prateek Jain, Sham M. Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Parallelizing Stochastic Gradient Descent for Least Squares Regression: Mini-batching, Averaging, and Model Misspecification. *J. Mach. Learn. Res.*, 18:223:1–223:42, 2017.

[31] Prateek Jain, Sham M. Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Accelerating Stochastic Gradient Descent for Least Squares Regression. In *COLT*, volume 75, pp. 545–604, 2018.

[32] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaïd Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konecný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning. *CoRR*, abs/1912.04977, 2019. URL `http://arxiv.org/abs/1912.04977`.

[33] Peter Kairouz, Brendan McMahan, Shuang Song, Om Thakkar, Abhradeep Thakurta, and Zheng Xu. Practical and private (deep) learning without sampling or shuffling. In *ICML*, 2021.

[34] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[35] Anastasia Koloskova, Ryan McKenna, Zachary Charles, Keith Rush, and Brendan McMahan. Convergence of Gradient Descent with Linearly Correlated Noise and Applications to Differentially Private Learning. *arXiv Preprint*, 2023.

[36] Dan Kucerovsky, Kaveh Mousavand, and Aydin Sarraf. On some properties of Toeplitz matrices. *Cogent Mathematics*, 3(1):1154705, 2016.

[37] Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, et al. Madlad-400: A multilingual and document-level large audited dataset. *arXiv preprint arXiv:2309.04662*, 2023.

[38] Chao Li, Gerome Miklau, Michael Hay, Andrew McGregor, and Vibhor Rastogi. The matrix mechanism: optimizing linear counting queries under differential privacy. *The VLDB journal*, 24:757–781, 2015.

[39] Xiyang Liu, Prateek Jain, Weihao Kong, Sewoong Oh, and Arun Sai Suggala. Near Optimal Private and Robust Linear Regression. *arXiv preprint arXiv:2301.13273*, 2023.

[40] Brendan McMahan and Abhradeep Thakurta. Federated learning with formal differential privacy guarantees. *Google AI Blog*, 2022.

[41] H Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning Differentially Private Recurrent Language Models. In *ICLR*, 2018.

[42] Kamyar Moshksar. On the Absolute Constant in Hanson-Wright Inequality. *arXiv preprint*, 2021.

[43] Alan V Oppenheim, Alan S Willsky, and Nawab. *Signals and Systems*, volume 2. 1997.

[44] Antonio Orvieto, Hans Kersting, Frank Proske, Francis Bach, and Aurelien Lucchi. Anticorrelated Noise Injection for Improved Generalization. In *ICML*, pp. 17094–17116, 2022.

[45] Krishna Pillutla, Yassine Laguel, Jérôme Malick, and Zaid Harchaoui. Federated learning with superquantile aggregation for heterogeneous data. *Machine Learning*, pp. 1–68, 2023.

[46] Mark Rudelson and Roman Vershynin. Hanson-Wright Inequality and Sub-Gaussian Concentration, 2013.

[47] Adam Smith and Abhradeep Thakurta. (nearly) optimal algorithms for private online learning in full-information and bandit settings. In *Advances in Neural Information Processing Systems*, pp. 2733–2741, 2013.

[48] Shuang Song, Kamalika Chaudhuri, and Anand D Sarwate. Stochastic gradient descent with differentially private updates. In *2013 IEEE Global Conference on Signal and Information Processing*, pp. 245–248. IEEE, 2013.

[49] Prateek Varshney, Abhradeep Thakurta, and Prateek Jain. (nearly) optimal private linear regression via adaptive clipping. *arXiv preprint arXiv:2207.04686*, 2022.

[50] Zheng Xu, Yanxiang Zhang, Galen Andrew, Christopher A Choquette-Choo, Peter Kairouz, H Brendan McMahan, Jesse Rosenstock, and Yuanbo Zhang. Federated Learning of Gboard Language Models with Differential Privacy. *arXiv preprint arXiv:2305.18465*, 2023.

Appendix

Table of Contents

Algorithm 1 The DP-FTRL and Noisy-FTRL algorithms with a noise correlation matrix $\boldsymbol{B} \in \mathbb{R}^{T \times T}$

Input: $\boldsymbol{B} \in \mathbb{R}^{T \times T}$, initial iterate $\boldsymbol{\theta}_0 \in \mathbb{R}^d$, ℓ_2 clip norm G, noise multiplier σ_{dp}, learning rate η, dataset $\mathcal{D}$

1: **for** $t = 0, \ldots, T-1$ **do**

2: Compute the gradient $\boldsymbol{g}_t = \begin{cases} \nabla f(\boldsymbol{\theta}_t; \boldsymbol{z}_t) + \nabla r(\boldsymbol{\theta}) & \text{for Noisy-FTRL,} \\ \mathsf{clip}\left(\nabla f(\boldsymbol{\theta}_t; \boldsymbol{z}_t), G\right) + \nabla r(\boldsymbol{\theta}) & \text{for DP-FTRL} \end{cases}$

3: Sample $\boldsymbol{w}_t \sim \mathcal{N}(0, \sigma_{\mathsf{dp}}^2 G^2 \boldsymbol{I}_d)$ and calculate the correlated noise $\widetilde{\boldsymbol{w}}_t = \sum_{\tau=0}^{t} \boldsymbol{B}_{t,\tau} \boldsymbol{w}_\tau$

4: Update $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \widetilde{\boldsymbol{g}}_t$ for the noisy gradient $\widetilde{\boldsymbol{g}}_t = \boldsymbol{g}_t + \widetilde{\boldsymbol{w}}_t$

Return $\boldsymbol{\theta}_T$

A Further Background on DP-FTRL

In this appendix, we give a more detailed background of DP-FTRL, and its exact notion of DP.

A.1 DP-FTRL: The Matrix Mechanism for Private Learning

The DP-FTRL algorithm [19, 33] is obtained by adapting the matrix mechanism, originally designed for linear counting queries [38], to optimization with a sequence $(\boldsymbol{g}_0, \ldots, \boldsymbol{g}_{T-1})$ of gradient vectors.

Algorithm 1 gives a detailed description of DP-FTRL. We given an alternate description of DP-FTRL with an invertible lower-triangular noise correlation matrix $\boldsymbol{B} \in \mathbb{R}^{T \times T}$. Denoting $\boldsymbol{C} = \boldsymbol{B}^{-1}$, the iterates of DP-FTRL are generated by the update

$$\begin{pmatrix} \boldsymbol{\theta}_1 \\ \vdots \\ \boldsymbol{\theta}_T \end{pmatrix} = \begin{pmatrix} \boldsymbol{\theta}_0 \\ \vdots \\ \boldsymbol{\theta}_{T-1} \end{pmatrix} - \eta \boldsymbol{B} \left(\boldsymbol{C} \begin{pmatrix} \boldsymbol{g}_0 \\ \vdots \\ \boldsymbol{g}_{T-1} \end{pmatrix} + \begin{pmatrix} \boldsymbol{w}_0 \\ \vdots \\ \boldsymbol{w}_{T-1} \end{pmatrix} \right) \tag{10}$$

where η is a learning rate and $\boldsymbol{w}_t \sim \mathcal{N}(\boldsymbol{0}, G^2 \sigma_{\mathsf{dp}}^2 \boldsymbol{I}_d)$ is i.i.d. Gaussian noise with a noise multiplier σ_{dp} and G is the ℓ_2 clip norm. It is common to refer to $\boldsymbol{C}$ as the *encoder*, while $\boldsymbol{B}$ is referred to as the *decoder*.

This privacy of (10) can be seen a postprocessing of a single application of the Gaussian mechanism. Let $\boldsymbol{G}, \boldsymbol{W} \in \mathbb{R}^{T \times d}$ denote the matrix where each row is the gradient $\boldsymbol{g}_t$ (and respectively the noise $\boldsymbol{w}_t$). Then, (10) is effectively the postprocessing of one run one run of the Gaussian mechanism $\boldsymbol{C}\boldsymbol{G} + \boldsymbol{W}$. Under a neighborhood model that can change one row of $\boldsymbol{G}$, it can be seen that the maximum sensitivity of this operation is $\max_t \|\boldsymbol{C}_{:,t}\|_2^2$ [19]. This sensitivity logic also holds for adaptively chosen gradients; we postpone a formal description to Appendix A.2.

Connection to the exposition in prior work: Prior work introduced DP-FTRL differently. Letting $\boldsymbol{A} \in \mathbb{R}^{T \times T}$ denote the lower triangular matrix of all ones, update (10) can also be written as

$$\begin{pmatrix} \boldsymbol{\theta}_1 - \boldsymbol{\theta}_0 \\ \vdots \\ \boldsymbol{\theta}_T - \boldsymbol{\theta}_0 \end{pmatrix} = -\eta \widetilde{\boldsymbol{B}} \left(\boldsymbol{C} \begin{pmatrix} \boldsymbol{g}_0 \\ \vdots \\ \boldsymbol{g}_{T-1} \end{pmatrix} + \begin{pmatrix} \boldsymbol{w}_0 \\ \vdots \\ \boldsymbol{w}_{T-1} \end{pmatrix} \right), \tag{11}$$

where $\widetilde{\boldsymbol{B}} = \boldsymbol{A}\boldsymbol{B}$. The equivalence between (10) and (11) can be seen by multiplying (10) by $\boldsymbol{A}$, which is also equivalent to taking the cumulative sum of the rows of a matrix. In this notation, the objective from (3) used in previous work to find the matrix $\boldsymbol{B}$ can equivalently be written as

$$\varphi(\boldsymbol{B}) = \|\widetilde{\boldsymbol{B}}\|_F^2 = \|\boldsymbol{A}\boldsymbol{B}\|_F^2 .$$

DP-FTRL with Toeplitz matrices: We focus on the class of lower-triangular and Toeplitz matrices $\boldsymbol{B}$. That is, $[\boldsymbol{B}]_{t,t'} = \beta_{t-t'}$ for all $t \geq t'$ where $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{T-1})$ is the first column of $\boldsymbol{B}$.[3] In this case, (10) reduces to this simple update:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \left(\boldsymbol{g}_t + \sum_{\tau=0}^{t} \beta_\tau \boldsymbol{w}_{t-\tau} \right) . \tag{12}$$

[3] This implies that $\boldsymbol{C} = \boldsymbol{B}^{-1}$ is also lower-triangular and Toeplitz [36, Prop. 2.2 & Rem. 2.3].

This lets us study DP-FTRL as a time-invariant stochastic process and characterize its stationary behavior.

A.2 Differential Privacy in Adaptive Streams

Neighboring streams: We consider learning algorithms as operating over streams of gradients $\boldsymbol{g}_0, \boldsymbol{g}_1, \ldots \in \mathbb{R}^d$. We consider differential privacy (DP) under the "zero-out" notion of neighborhood [33]. Two streams $\boldsymbol{G} = (\boldsymbol{g}_0, \ldots, \boldsymbol{g}_{T-1})$ and $\boldsymbol{G}' = (\boldsymbol{g}'_0, \ldots, \boldsymbol{g}'_{T-1})$ of length T are said to be neighbors if $\boldsymbol{g}_\tau = \boldsymbol{g}'_\tau$ for all positions $\tau \leq T-1$ except possibly one position t where one of $\boldsymbol{g}$ or $\boldsymbol{g}'$ is the zero vector.

DP with adaptive continual release: It is customary to formalize DP with adaptive streams as a privacy game between a mechanism $\mathcal{M}$ and a privacy adversary $\mathcal{A}$. This is known as the *adaptive continual release setting* [28]. The game makes a binary choice $b \in \{0, 1\}$ ahead of time — this remains fixed throughout and is not revealed to either $\mathcal{M}$ or $\mathcal{A}$. Each round t consists of four steps:

- $\mathcal{M}$ sends the current model parameters $\boldsymbol{\theta}_t$ to the adversary $\mathcal{A}$;
- $\mathcal{A}$ generates two gradient vectors $\boldsymbol{g}_t, \boldsymbol{g}'_t$ (e.g. as $\nabla f(\boldsymbol{\theta}_t; \boldsymbol{z}_t)$ for $\boldsymbol{z}_t \sim \mathbb{P}_{\mathsf{data}}$ or simply the zero vector);
- the game accepts these inputs if the partial streams $(\boldsymbol{g}_0, \ldots, \boldsymbol{g}_t)$ and $(\boldsymbol{g}'_0, \ldots, \boldsymbol{g}'_t)$ are neighbors;
- $\mathcal{M}$ receives $\boldsymbol{g}_t$ if $b = 0$ else $\boldsymbol{g}'_t$.

DP in this setting requires that the adversary cannot infer the value of b, i.e., the distribution of $\boldsymbol{\theta}_{0:T} | b = 0$ to be "close" to that of $\boldsymbol{\theta}_{0:T} | b = 1$ (where the definition of "closeness" depends on the DP variant). For instance, (ε, δ)-DP [20] requires for each $b \in \{0, 1\}$ and any outcome set S that

$$\mathbb{P}(\boldsymbol{\theta}_{0:T} \in S \,|\, b) \leq \exp(\varepsilon)\, \mathbb{P}(\boldsymbol{\theta}_{0:T} \in S \,|\, 1-b) + \delta\,.$$

Similarly, ρ-zCDP [10] in this setting requires that the Rényi α-divergence between the distribution P_0 of $\boldsymbol{\theta}_{0:T} | b = 0$ and the distribution P_1 of $\boldsymbol{\theta}_{0:T} | b = 1$ are close:

$$D_\alpha(P_0 \| P_1) \leq \rho\alpha$$

for all $\alpha \in (0, \infty)$. Following standard arguments [e.g. 8], ρ-zCDP in this setting implies $(\varepsilon_\delta, \delta)$-DP with

$$\varepsilon_\delta \leq \inf_{\alpha > 1} \left\{ \rho\alpha + \frac{1}{\alpha - 1} \log\left(\frac{1}{\alpha\delta}\right) + \log(1 - \alpha^{-1})\,. \right\}$$

DP-FTRL satisfies a zCDP guarantee as described in Theorem 2.1 in §1. This guarantee is equivalent to the one obtained by interpreting (10) as the postprocessing of one run one run of the Gaussian mechanism $\boldsymbol{C}\boldsymbol{G} + \boldsymbol{W}$.

B Asymptotics of DP-FTRL for Mean Estimation

We now prove Theorem 3.1 on mean estimation.

Proof of Theorem 3.1. Since $|\nabla f(\theta; z)| = |z| \leq 1$ and $G \geq 1$, there is no gradient clipping. Setting $\delta_t = \theta_t - \mu$ and $w_{\mathsf{sgd}} = (z - \mathbb{E}[z])/\sigma_{\mathsf{sgd}}$, the updates (2) can be written:

$$\delta_{t+1} = (1 - \eta)\delta_t + \eta\sigma_{\mathsf{sgd}} w_{\mathsf{sgd}} - \eta\sigma_{\mathsf{dp}} G \sum_{\tau=0}^{t} \beta_\tau w_{t-\tau}\,.$$

We can then use the results from §F.1 (Theorem F.2) to obtain an expression for the steady state error:

$$F_\infty(B) = \frac{1}{2\pi}\eta^2 \int_{-\pi}^{\pi} \frac{|B(\omega)|^2 G^2 \sigma_{\mathsf{dp}}^2 \gamma_\infty(B)^2 + \sigma_{\mathsf{sgd}}^2}{|1 - \eta - \exp(i\omega)|^2}\, \mathrm{d}\omega\,.$$

where we also used the fact that the SGD noise $z - \mathbb{E}[z]$ is iid in each step and independent of the DP noise $\boldsymbol{w}$, so that the power spectral density of the sum of these two noise sources is simply the sum of the power spectral densities of the individual sources, with the power spectral density of

Table 2: Asymptotic suboptimality of Noisy-SGD and Noisy-FTRL for linear regression with Gaussian inputs based on the eigenvalues λ_k of the Hessian $\boldsymbol{H}$. We give the bounds in terms of the learning rate η, dimension d, the effective dimension $d_{\mathsf{eff}} = \mathsf{Tr}\left[\boldsymbol{H}\right]/\|\boldsymbol{H}\|_2$, and the noise variance ρ^{-1} representing the privacy level. We take $G = 1$ and $\|\boldsymbol{H}\|_2 = 1$ w.l.o.g. Noisy-FTRL is always better at large dimension d or small learning rate η.

Eigenvalues of the Hessian $\boldsymbol{H}$	Effective dimension d_{eff}	Noisy-SGD	Noisy-FTRL	Ratio of $\frac{\text{Noisy-FTRL}}{\text{Noisy-SGD}}$
$\lambda_k = 1$	d	$\eta d \rho^{-1}$	$\eta^2 d \rho^{-1} \log^2(\frac{1}{\eta})$	$\eta \log^2(\frac{1}{\eta})$
$\lambda_k = 1/\sqrt{k}$	$\sqrt{d}$	$\eta d \rho^{-1}$	$\eta^2 \sqrt{d} \rho^{-1} \log^2(\frac{d}{\eta})$	$\frac{\eta}{\sqrt{d}} \log^2(\frac{d}{\eta})$
$\lambda_k = 1/k$	$\log d$	$\eta d \rho^{-1}$	$\eta^2 \rho^{-1} \log^3(\frac{d}{\eta})$	$\frac{\eta}{d} \log^3(\frac{d}{\eta})$
$\lambda_k = 1/k^2$	constant	$\eta d \rho^{-1}$	$\eta^2 \rho^{-1} \log^2(\frac{d}{\eta})$	$\frac{\eta}{d} \log^3(\frac{d}{\eta})$

the SGD noise being constant and equal to σ_{sgd}. Furthermore, we also have the expression for the sensitivity.

$$\gamma_\infty\left(B\right)^2 = \frac{1}{2\pi}\left(\int_{-\pi}^{\pi} \left|B\left(\omega\right)\right|^2\right).$$

Thus F_∞ is a product of a linear function of $|B|^2$ and $\frac{1}{|B|^2}$ (through $\gamma\left(B\right)$). By the Cauchy-Schwarz inequality, the produce is minimized when

$$\frac{\left|B\left(\omega\right)\right|^2}{\left|\exp\left(i\omega\right) - \left(1-\eta\right)\right|^2} = \frac{1}{\left|B\left(\omega\right)\right|^2} \implies \left|B^\star\left(\omega\right)\right| = \left|\sqrt{\exp\left(i\omega\right) - \left(1-\eta\right)}\right| = \left|\sqrt{1 - \left(1-\eta\right)\exp\left(-i\omega\right)}\right|$$

The rest of the proof follows by computing standard integrals (See §F.4 and Lemma F.15 for details). □

C Asymptotics of DP-FTRL for Linear Regression

The goal of this section is to prove Theorem 4.1. The proof relies heavily on the following matching upper and lower bounds on the stationary error of Noisy-FTRL with any noise correlations $\boldsymbol{\beta}$ in the frequency domain using its discrete-time Fourier transform (DTFT) B as:

$$F_\infty(B) = \Theta\left(\eta\sigma_{\mathsf{sgd}}^2 \mathsf{Tr}\left[\boldsymbol{H}\right] + \eta^2 G^2 \rho^{-1} \gamma_\infty(B)^2 \int_{-\pi}^{\pi} |B(\omega)|^2 h(\omega)\,\mathrm{d}\omega\right), \tag{13}$$

where the function $h : [-\pi, \pi] \to \mathbb{R}$ depends on the eigenvalues $\lambda_1, \ldots, \lambda_d$ of the input covariance $\boldsymbol{H}$:

$$h(\omega) = \sum_{j=1}^{d} \frac{\lambda_j}{|1 - \exp(i\omega) - \eta\lambda_j|^2}. \tag{14}$$

The outline of the section is

- **Appendix C.1**: Setup, including notation, and assumptions.
- **Appendix C.2**: Proofs of the upper bound of (13), specifically Theorem C.15 (see also Theorem C.14 for the time-domain description).
- **Appendix C.3**: Proofs of the lower bound of (13), specifically Theorem C.17.
- **Appendix C.4**: Asymptotics of ν-Noisy-FTRL.
- **Appendix C.5**: Asymptotics of anti-PGD (see Table 3).
- **Appendix C.6**: Proofs of intermediate technical results.

The separation between Noisy-SGD and ν-Noisy-FTRL is further illustrated in Table 2. Following common practice [e.g. 12], we compare the rates for various regimes of eigenvalue decays for $\boldsymbol{H}$.

Our analysis can also give bounds for other related approaches such as anti-PGD [44].

Table 3: **Comparison to prior work**: We apply our theory to compute F_∞ for linear regression given choices of $\boldsymbol{B}$ used in prior work. Though certain choices of the noise correlation $\boldsymbol{\beta}$ may be optimal for finite linear counting queries [21], our results show that they have $F_\infty = \infty$ because the sensitivity diverges as $T \to \infty$. ν-Noisy-FTRL effectively introduces an additional damping term $(1-\nu)^t$ in the correlations of [21] to achieve near-optimality for linear regression. Damping similarly helps for anti-PGD [44], where the resulting error is the geometric mean of the lower bound and the bound of Noisy-SGD from Theorem 4.1.

Algorithm	**Noise Correlation Weights $\boldsymbol{\beta}$**	**Sensitivity in T steps** $\gamma_T(\boldsymbol{\beta})^2$	**Asymptotic Suboptimality** $F_\infty(\boldsymbol{\beta})$
[21]	Eq. (7) with $\nu = 0$	$\log T$	∞
ν-Noisy-FTRL (Ours)	Eq. (7) with $0 < \nu \le \eta\mu$	$\log(1/\nu)$	$\eta^2 G^2 \rho^{-1} \mathsf{Tr}\left[\boldsymbol{H}\right] \log^2(1/\nu)$
Anti-PGD [44]	$(1, -1, 0, \ldots)$	T	∞
Anti-PGD + Damping	$(1, -(1-\nu), 0, \ldots)$	$1/\nu$	$\eta^{3/2} G^2 \rho^{-1} \sqrt{d \mathsf{Tr}\left[\boldsymbol{H}\right]}$

C.1 Setup, Assumptions, and Notation

C.1.1 Setup

Recall that we wish to minimize the objective

$$F(\boldsymbol{\theta}) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}_{\mathsf{data}}}\left[(y - \langle\boldsymbol{\theta}, \boldsymbol{x}\rangle)^2\right] . \tag{15}$$

Stochastic gradients: Given $(\boldsymbol{x}, y) \sim \mathbb{P}_{\mathsf{data}}$, the vector

$$\boldsymbol{g} := (\boldsymbol{x} \otimes \boldsymbol{x})\,\boldsymbol{\theta} - y\boldsymbol{x} = (\boldsymbol{x} \otimes \boldsymbol{x})(\boldsymbol{\theta} - \boldsymbol{\theta}_\star) - \xi\boldsymbol{x}$$

is a stochastic gradient of F at $\boldsymbol{\theta}$, i.e., $\mathbb{E}[\boldsymbol{g}] = \nabla F(\boldsymbol{\theta})$.

Noisy-FTRL Iterations: We specialize the Noisy-FTRL algorithm with Toeplitz noise correlations. Let T denote the number of iterations and $\boldsymbol{\beta}_{:T} = (\beta_0, \ldots, \beta_{T-1})$ denote the first column of the Toeplitz matrix $\boldsymbol{B} = \mathrm{Toeplitz}(\boldsymbol{\beta}_{:T}) \in \mathbb{R}^{T\times T}$. Starting from a given $\boldsymbol{\theta}_0 \in \mathbb{R}^d$, Noisy-FTRL samples a fresh input-output pair $(\boldsymbol{x}_t, y_t) \sim \mathbb{P}_{\mathsf{data}}$ and noise $\boldsymbol{w}_t$ to set

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta\left((\boldsymbol{x}_t \otimes \boldsymbol{x}_t)\boldsymbol{\theta}_t - y_t\boldsymbol{x}_t\right)) - \eta\sum_{\tau=0}^{t}\beta_\tau \boldsymbol{w}_{t-\tau} . \tag{16}$$

Recall that the sensitivity $\gamma_T(\boldsymbol{\beta})$ equals to the maximum columns norm of $\boldsymbol{B}^{-1} = (\mathrm{Toeplitz}(\boldsymbol{\beta}))^{-1}$:

$$\gamma_T(\boldsymbol{\beta}) = \max_{\tau=0,\ldots,T-1}\left\|\boldsymbol{B}^{-1}\boldsymbol{e}_\tau\right\|_2 , \tag{17}$$

where $\boldsymbol{e}_\tau = \big(\mathbb{I}(j=\tau)\big)_{\tau=0}^{T-1} \in \mathbb{R}^T$ is a standard basis vector. Note that the submatrix $\boldsymbol{B}_{0:m,0:m}$ of the first m rows and columns of $\boldsymbol{B}$ equals $\left(\mathrm{Toeplitz}(\beta_0, \ldots, \beta_{m-1})\right)^{-1}$. Thus, the sensitivity $\gamma_t(\boldsymbol{\beta})$ is an increasing function of t always.

Infinite-time limit of Noisy-FTRL: We study the Noisy-FTRL error under the limit $T \to \infty$ with an infinite sequence $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots)$ of weights.

It is also convenient to re-index time to start from $t = -\infty$ and consider the sequence $(\boldsymbol{\theta})_{t=-\infty}^{\infty}$ produced by analogue of Equation (16), which reads

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta\left((\boldsymbol{x}_t \otimes \boldsymbol{x}_t)\boldsymbol{\theta}_t - y_t\boldsymbol{x}_t\right)) - \eta\sum_{\tau=0}^{\infty}\beta_\tau \boldsymbol{w}_{t-\tau} . \tag{18}$$

Note that this includes a summation over all previous DP noise $(\boldsymbol{w}_\tau)_{\tau=-\infty}^{t}$. For this sum to have finite variance, we require $\sum_{\tau=0}^{\infty}\beta_\tau^2 < \infty$ or that $\boldsymbol{\beta} \in \ell^2$, the space of all square-summable infinite sequences. We will assume this holds throughout.

Sensitivity in the infinite limit: We define the sensitivity $\gamma_\infty(\boldsymbol{\beta})$ by consider the linear operator $\boldsymbol{B} = \mathrm{Toeplitz}(\boldsymbol{\beta})$ as the convolution operator $[\boldsymbol{B}\boldsymbol{w}]_t = \sum_{\tau=0}^{\infty}\beta_\tau \boldsymbol{w}_{t-\tau}$ on input $\boldsymbol{w} = (\boldsymbol{w}_\tau)_{\tau=-\infty}^{\infty}$. Let $\boldsymbol{B}^{-1}$ be the inverse operator to $\boldsymbol{B}$, assuming it exists. Note that the column norms $\left\|\boldsymbol{B}^{-1}\boldsymbol{e}_\tau\right\|_2$

from (17) become equal for all τ as $T \to \infty$. Thus, we get that the limiting sensitivity in the infinite time limit equals

$$\gamma_\infty(\boldsymbol{\beta}) = \left\| \boldsymbol{B}^{-1} \boldsymbol{e}_0 \right\|_2 \tag{19}$$

for $\boldsymbol{B} = \text{Toeplitz}(\boldsymbol{\beta})$ and $\boldsymbol{e}_0 = (\mathbb{1}(\tau = 0))_{\tau=0}^{\infty} \in \ell^2$. If $\boldsymbol{e}_0 \notin \text{Range}(\boldsymbol{B})$, then we take $\gamma_\infty(\boldsymbol{\beta}) = \infty$.

Frequency-domain description: Our analysis relies on the frequency-domain representation $B : [-\pi, \pi] \to \mathbb{C}$ of $\boldsymbol{\beta}$ obtained via a discrete-time Fourier transform (DTFT) and defined as

$$B(\omega) = \sum_{t=0}^{\infty} \beta_t \, \exp(i\omega t) \, . \tag{20}$$

The sequence $\boldsymbol{\beta}$ can be recovered from $B(\omega)$ using the inverse Fourier transform. Note that $\beta \in \ell^2$ is equivalent to $B \in L^2$, the space of square integrable functions, by Parseval's theorem. The sensitivity (19) can be defined in the Fourier domain as follows.

Property C.1. *Let $B(\omega)$ denote the DTFT of $\boldsymbol{\beta} \in \ell^2$. Then, we have*

$$\gamma_\infty(\boldsymbol{\beta})^2 = \gamma_\infty(B)^2 := \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{\mathrm{d}\omega}{|B(\omega)|^2} \, . \tag{21}$$

Proof. Let $\boldsymbol{z} = \boldsymbol{B}^{-1} \boldsymbol{e}_0$ be the solution of the linear system $\boldsymbol{B}\boldsymbol{z} = \boldsymbol{e}_0$. Let $Z(\omega)$ denote the DTFT of $\boldsymbol{z}$. Since the linear operator $\boldsymbol{B}$ is a convolution with the weights of $\boldsymbol{\beta}$, this system can be expressed in the Fourier domain as

$$B(\omega) Z(\omega) = \sum_{\tau=0}^{\infty} [\boldsymbol{e}_0]_\tau \exp(-i\omega\tau) = 1 \, .$$

Thus, $Z(\omega) = 1/B(\omega)$. We complete the proof with Parseval's theorem: $\|\boldsymbol{z}\|_2^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |Z(\omega)|^2 \, \mathrm{d}\omega$. □

C.1.2 Assumptions

We prove the stationary error bounds under a relaxation of the assumptions in §4.

Assumption C.2. *The data distribution $\mathbb{P}_{\mathsf{data}}$ satisfies the following:*

- ***(B1) Input Mean and Covariance**: The inputs have mean $\mathbb{E}[\boldsymbol{x}] = \boldsymbol{0}$ and covariance $\mathbb{E}[\boldsymbol{x} \otimes \boldsymbol{x}] =: \boldsymbol{H}$. Further, $L = \lambda_1 \geq \cdots \geq \lambda_d =: \mu > 0$ are the eigenvalues of $\boldsymbol{H}$.*
- ***(B2) Noise Mean and Variance**: There exists a $\boldsymbol{\theta}_\star \in \mathbb{R}^d$ such that $y = \langle \boldsymbol{\theta}_\star, \boldsymbol{x} \rangle + \xi$ where ξ is independent of $\boldsymbol{x}$ with $\mathbb{E}[\xi] = 0$ and $\mathbb{E}[\xi^2] \leq \sigma_{\mathsf{sgd}}^2$.*
- ***(B3) Input Kurtosis**: There exists $R^2 < \infty$ such that $\mathbb{E}\left[\|\boldsymbol{x}\|_2^2 (\boldsymbol{x} \otimes \boldsymbol{x}) \right] \preceq R^2 \boldsymbol{H}$. Moreover, for every PSD $\boldsymbol{P} \in \mathbb{S}_+^d$ that commutes with $\boldsymbol{H}$ (i.e., $\boldsymbol{P}\boldsymbol{H} = \boldsymbol{H}\boldsymbol{P}$), we have $\mathbb{E}\left[(\boldsymbol{x} \otimes \boldsymbol{x}) \boldsymbol{H}^{-1/2} \boldsymbol{P} \boldsymbol{H}^{-1/2} (\boldsymbol{x} \otimes \boldsymbol{x}) \right] \preceq C_{\mathsf{kurt}} \, \mathsf{Tr}[\boldsymbol{P}] \, \boldsymbol{H}$ for some $C_{\mathsf{kurt}} < \infty$.*

These assumptions are fairly standard in the context of linear regression. Assumption (**B1**) implies that the Hessian matrix of objective $F(\boldsymbol{\theta})$ is $\boldsymbol{H} \succ 0$. Thus, F is L-smooth and μ-strongly convex. Assumption (**B2**) implies that $\boldsymbol{\theta}_\star$ is the unique global minimizer of F and that the linear model is well-specified. The upper bounds we prove continue to hold in the case where the linear model is mis-specified (i.e. ξ is not independent of $\boldsymbol{x}$) but we still have $\mathbb{E}[\xi^2 (\boldsymbol{x} \otimes \boldsymbol{x})] \preceq \sigma_{\mathsf{sgd}}^2 \boldsymbol{H}$.

Assumption (**B3**) is a kurtosis (i.e. 4th moment) assumption on the input distribution; we will momentarily show that it follows with absolute constants when $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{H})$. More generally, by taking a trace, we get from Jensen's inequality that $\mathsf{Tr}[\boldsymbol{H}] \leq R^2$. The case of $\boldsymbol{P} = \boldsymbol{I}$ of the second part of Assumption (**B3**) has a special significance in the literature [e.g. 26, 31] as $C_{\mathsf{kurt}} \mathsf{Tr}[\boldsymbol{I}] = C_{\mathsf{kurt}} d$ is the number of samples that allows the spectral concentration of the empirical covariance to the population covariance $\boldsymbol{H}$.

Property C.3. *if $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{H})$, we have that Assumption (**B3**) holds with $R^2 \leq 3 \, \mathsf{Tr}[\boldsymbol{H}]$ and $C_{\mathsf{kurt}} \leq 3$.*

Proof. Let $\boldsymbol{z} = \boldsymbol{H}^{-1/2}\boldsymbol{x}$ is element-wise independent and distributed as a standard Gaussian. For the first part, denote $\boldsymbol{M} = \boldsymbol{H}^{-1/2}\,\mathbb{E}[\|\boldsymbol{x}\|_2^2\boldsymbol{x}\otimes\boldsymbol{x}]\,\boldsymbol{H}^{-1/2} = \mathbb{E}[\langle\boldsymbol{z},\boldsymbol{H}\boldsymbol{z}\rangle\boldsymbol{z}\otimes\boldsymbol{z}]$. Elementary properties of the standard Gaussian distribution give

$$\mathbb{E}[z_k z_l z_j^2] = \begin{cases} 3\,, & \text{if } k=l=j \\ 1\,, & \text{if } k=l\neq i \\ 0\,, & \text{if } k\neq l\,, \end{cases} \quad \text{and} \quad \mathbb{E}[z_k z_l z_j z_{j'}] = \begin{cases} 1\,, & \text{if } k=j \text{ and } l=j' \\ 1\,, & \text{if } k=j' \text{ and } l=j \\ 0\,, & \text{else} \end{cases}$$

for $j \neq j'$. Thus, we have $\boldsymbol{M} = 2\boldsymbol{H} + \mathsf{Tr}\left[\boldsymbol{H}\right]\boldsymbol{I}$. This gives

$$\mathbb{E}[\|\boldsymbol{x}\|_2^2\boldsymbol{x}\otimes\boldsymbol{x}] = \boldsymbol{H}^{1/2}\boldsymbol{M}\boldsymbol{H}^{1/2} = 2\boldsymbol{H}^2 + \mathsf{Tr}\left[\boldsymbol{H}\right]\boldsymbol{H} \preceq 3\mathsf{Tr}\left[\boldsymbol{H}\right]\boldsymbol{H}\,.$$

For the second part, let $\boldsymbol{H} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ and $\boldsymbol{P} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^\top$ be the eigenvalue decomposition of $\boldsymbol{H}, \boldsymbol{P}$ respectively (since they commute, they are simultaneously diagonalized in the same basis given by the columns of $\boldsymbol{U}$). Since $\boldsymbol{U}^\top\boldsymbol{z}$ has the same distribution as $\boldsymbol{z}$ by the spherical invariance of Gaussians, we have,

$$\boldsymbol{H}^{-1/2}\mathbb{E}\left[(\boldsymbol{x}\otimes\boldsymbol{x})\boldsymbol{H}^{-1/2}\boldsymbol{P}\boldsymbol{H}^{-1/2}(\boldsymbol{x}\otimes\boldsymbol{x})\right]\boldsymbol{H}^{-1/2} = \mathbb{E}\left[(\boldsymbol{z}\otimes\boldsymbol{z})\boldsymbol{P}(\boldsymbol{z}\otimes\boldsymbol{z})\right] = \boldsymbol{U}\,\mathbb{E}\left[(\boldsymbol{z}\otimes\boldsymbol{z})\boldsymbol{\Sigma}(\boldsymbol{z}\otimes\boldsymbol{z})\right]\boldsymbol{U}^\top\,. \tag{22}$$

Each off-diagonal entry of $\mathbb{E}\left[(\boldsymbol{z}\otimes\boldsymbol{z})\boldsymbol{\Sigma}(\boldsymbol{z}\otimes\boldsymbol{z})\right]$ is zero since it involves expected odd powers of Gaussians. Its j^{th} diagonal entry equals (denoting $\sigma_j := [\boldsymbol{\Sigma}]_{j,j}$)

$$\mathbb{E}\left[z_j^2\sum_{k=1}^d \sigma_k z_k^2\right] = \sigma_j\mathbb{E}[z_j^4] + \sum_{k\neq j}\sigma_k\,\mathbb{E}[z_j^2 z_k^2] = 2\sigma_j + \mathsf{Tr}\left[\boldsymbol{\Sigma}\right]\,.$$

This gives $\mathbb{E}\left[(\boldsymbol{z}\otimes\boldsymbol{z})\boldsymbol{\Sigma}(\boldsymbol{z}\otimes\boldsymbol{z})\right] = 2\boldsymbol{\Sigma} + \mathsf{Tr}\left[\boldsymbol{\Sigma}\right]\boldsymbol{I} \preceq 3\mathsf{Tr}\left[\boldsymbol{\Sigma}\right]\boldsymbol{I}$ since $\boldsymbol{\Sigma} \succeq \boldsymbol{0}$. Plugging this back into (22) and rearranging completes the proof. □

C.1.3 Notation

We set up some notation, that we use throughout this section.

- It is convenient to rewrite the Noisy-FTRL recursion in terms of the difference $\boldsymbol{\theta}'_t := \boldsymbol{\theta}_t - \boldsymbol{\theta}_\star$. We can rewrite the Noisy-FTRL recursion (18) as

$$\boldsymbol{\theta}'_{t+1} = \big(\boldsymbol{I} - \eta(\boldsymbol{x}_t\otimes\boldsymbol{x}_t)\big)\boldsymbol{\theta}'_t + \eta\,\xi_t\boldsymbol{x}_t - \eta\sum_{\tau=0}^\infty \beta_\tau\boldsymbol{w}_{t-\tau}\,. \tag{23}$$

  We will analyze this recursion.
- We describe the asymptotic suboptimality in terms of the self-adjoint linear operator $\boldsymbol{T} : \ell^2 \to \ell^2$ defined by

$$[\boldsymbol{T}\boldsymbol{\beta}]_t = \sum_{\tau=0}^\infty \beta_\tau \sum_{j=1}^d (1-\eta\lambda_j)^{|t-\tau|}\,. \tag{24}$$

  This operator is positive semi-definite, as we show in Lemma C.6 below. In the finite time setting, we could represent $\boldsymbol{T}$ by the matrix

$$\boldsymbol{T} = \begin{bmatrix} d & \sum_{j=1}^d(1-\eta\lambda_j) & \sum_{j=1}^d(1-\eta\lambda_j)^2 & \cdots \\ \sum_{j=1}^d(1-\eta\lambda_j) & d & \sum_{j=1}^d(1-\eta\lambda_j) & \cdots \\ \sum_{j=1}^d(1-\eta\lambda_j)^2 & \sum_{j=1}^d(1-\eta\lambda_j) & d & \cdots \\ \vdots & & & \vdots \end{bmatrix}$$

  We only consider step-size $0 < \eta < 1/R^2$, which implies that $1-\eta\lambda_j \in (0,1)$ for all j.
- For $j = 1, \ldots, d$, define $\boldsymbol{T}_j : \ell^2 \to \ell^2$ as the linear operator

$$[\boldsymbol{T}_j\beta]_t = \sum_{\tau=0}^\infty \beta_\tau(1-\eta\lambda_j)^{|t-\tau|}\,. \tag{25}$$

Note that $[\boldsymbol{T}_j\boldsymbol{\beta}]_t < \infty$ always since

$$\sum_{\tau=0}^{\infty} \beta_\tau (1-\eta\lambda_j)^{|t-\tau|} \le \frac{2\|\boldsymbol{\beta}\|_\infty}{\eta\lambda_j} < \infty\,,$$

since $0 < \eta\lambda < 1$. Thus, we have that $\boldsymbol{T} = \sum_{j=1}^d \boldsymbol{T}_j$ by the bounded convergence theorem. Further, we show in the upcoming Lemma C.6 that each $\boldsymbol{T}_j$ are PSD.

- Define $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}, \boldsymbol{P}_{\boldsymbol{\beta}} \in \mathbb{S}^d$ as

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}} := \mathsf{diag}\left((\langle \boldsymbol{\beta}, \boldsymbol{T}_j\boldsymbol{\beta}\rangle)_{j=1}^d\right), \quad \text{and} \quad \boldsymbol{P}_{\boldsymbol{\beta}} = \boldsymbol{U}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\boldsymbol{U}^\top\,, \tag{26}$$

where $\boldsymbol{U}$ is the eigen-basis of $\boldsymbol{H} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$. By definition, $\boldsymbol{P}_{\boldsymbol{\beta}}$ commutes with $\boldsymbol{H}$ since $\boldsymbol{P}_{\boldsymbol{\beta}}\boldsymbol{H} = \boldsymbol{H}\boldsymbol{P}_{\boldsymbol{\beta}} = \boldsymbol{U}(\Lambda\boldsymbol{\Sigma}_{\boldsymbol{\beta}})\boldsymbol{U}^\top$. Further, since each $\boldsymbol{T}_j$ is PSD (Lemma C.6), we have that $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ and $\boldsymbol{P}_{\boldsymbol{\beta}}$ are PSD as well. We also have

$$\mathsf{Tr}\left[\boldsymbol{P}_{\boldsymbol{\beta}}\right] = \mathsf{Tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\right] = \langle \boldsymbol{\beta}, \boldsymbol{T}\boldsymbol{\beta}\rangle\,. \tag{27}$$

- Define the matrix $\boldsymbol{M}_\omega \in \mathbb{C}^{d\times d}$ as

$$\boldsymbol{M}_\omega = \left((1-\exp(i\omega))\boldsymbol{I} - \eta\boldsymbol{H}\right)^{-1}. \tag{28}$$

Throughout, we assume that Assumption C.2 holds.

Preliminary lemmas: This lemma helps us move back and forth between the time-domain and frequency-domain representations. See Appendix C.6 for a proof.

Lemma C.4. *Consider $\boldsymbol{\beta} \in \ell^2$ and its DTFT $B(\omega)$. If $0 < \eta < 1/\lambda_j$, we have*

$$\frac{1}{2}\langle \boldsymbol{\beta}, \boldsymbol{T}_j\boldsymbol{\beta}\rangle \le \frac{\eta\lambda_j}{2\pi}\int_{-\pi}^{\pi} \frac{|B(\omega)|^2\,\mathrm{d}\omega}{|1-\eta\lambda_j-\exp(i\omega)|^2} \le \langle \boldsymbol{\beta}, \boldsymbol{T}_j\boldsymbol{\beta}\rangle\,.$$

Setting $B(\omega) = 1$ and $\boldsymbol{\beta} = (1, 0, \ldots)$ gives the next corollary.

Corollary C.5. *If $0 < \eta < 1/\lambda_j$, we have,*

$$\frac{1}{2} \le \frac{\eta\lambda_j}{2\pi}\int_{-\pi}^{\pi} \frac{\mathrm{d}\omega}{|1-\eta\lambda_j-\exp(i\omega)|^2} \le 1\,.$$

Lemma C.6. *The operators $\boldsymbol{T}_j$ defined in* (25) *and $\boldsymbol{T}$ defined in* (24) *are both positive semi-definite for $\eta < 1/\max_{j\in[d]}\lambda_j$.*

Proof. Consider any $\boldsymbol{\beta} \in \ell^2$ and its DTFT $B(\omega)$. We have from Lemma C.4 that

$$0 \le \int_{-\pi}^{\pi} \frac{|B(\omega)|^2\,\mathrm{d}\omega}{|1-\eta\lambda_j-\exp(i\omega)|^2} \le \frac{2\pi}{\eta\lambda_j}\langle \boldsymbol{\beta}, \boldsymbol{T}_j\boldsymbol{\beta}\rangle\,,$$

or that $\langle \boldsymbol{\beta}, \boldsymbol{T}_j\boldsymbol{\beta}\rangle \ge 0$. □

C.2 Proof of the Upper Bound on the Asymptotic Suboptimality

The key tool in the analysis is the use of linear time-invariant (LTI) input-output systems to relate the output covariance to the input covariance using its *transfer function* (see Appendix F.1 for a summary). The Noisy-FTRL recursion is not trivial to characterize in this manner because the update (18) is not LTI. Instead, we decompose it into an infinite sequence of LTI systems and carefully analyze the error propagation.

This consists of the following steps:

Part 1: Decompose the Noisy-FTRL recursion into a sequence of LTI systems.
Part 2: Compute the transfer function of each LTI system.
Part 3: Compute the stationary covariance for each LTI system from the previous one.
Part 4: Combine the stationary covariances to get the stationary error of the original iterate.

C.2.1 Part 1: Decomposition into a Sequence of LTI Systems

A challenge in analyzing the stationary error of Equation (23) in the frequency domain is that it is not an LTI system. Replacing $\boldsymbol{x}_t \otimes \boldsymbol{x}_t$ by $\boldsymbol{H}$ in Equation (23) results in a LTI update; this system is quite similar to fixed design linear regression. However, this leads to an error in the general case, which satisfies a recursion of the same form as (23). We can repeat the same technique of replacing $\boldsymbol{x}_t \otimes \boldsymbol{x}_t$ by $\boldsymbol{H}$ and repeat this process indefinitely. This proof technique has been used in [4] to analyze stochastic tracking algorithms and [7] to analyze iterate-averaged SGD for linear regression. We adopt this technique to analyze the stationary covariance of DP mechanisms with correlated noise.

We define sequences $(\boldsymbol{\theta}_t^{(r)})_{t=-\infty}^{\infty}$ and $(\boldsymbol{\delta}_t^{(r)})_{t=-\infty}^{\infty}$ for $r \geq 0$ as follows:

$$\begin{aligned}
\boldsymbol{\theta}_{t+1}^{(0)} &= (\boldsymbol{I} - \eta \boldsymbol{H})\boldsymbol{\theta}_t^{(0)} + \eta \xi_t \boldsymbol{x}_t - \eta \sum_{\tau=0}^{\infty} \beta_\tau \boldsymbol{w}_{t-k}\,, \\
\boldsymbol{\theta}_{t+1}^{(r)} &= (\boldsymbol{I} - \eta \boldsymbol{H})\boldsymbol{\theta}_t^{(r)} + \eta(\boldsymbol{H} - \boldsymbol{x}_t \otimes \boldsymbol{x}_t)\boldsymbol{\theta}_t^{(r-1)} \text{ for } r > 0\,, \\
\boldsymbol{\delta}_{t+1}^{(r)} &= (\boldsymbol{I} - \eta \boldsymbol{x}_t \otimes \boldsymbol{x}_t)\boldsymbol{\delta}_t^{(r)} + \eta(\boldsymbol{H} - \boldsymbol{x}_t \otimes \boldsymbol{x}_t)\boldsymbol{\theta}_t^{(r)}\,.
\end{aligned} \tag{29}$$

These recursions are assumed to start at $t = -\infty$ from $\boldsymbol{\theta}_t^{(0)} = \boldsymbol{\theta}'_t$, $\boldsymbol{\delta}_t^{(r)} = \boldsymbol{0}$ for $r \geq 0$ and $\boldsymbol{\theta}_t^{(r)} = \boldsymbol{0}$ for $r > 0$. These recursions are a decomposition of (23) as we define below.

Property C.7. *For each iteration t and any integer $m \geq 0$, we have $\boldsymbol{\theta}'_t = \sum_{r=0}^{m} \boldsymbol{\theta}_t^{(r)} + \boldsymbol{\delta}_t^{(m)}$.*

Proof. We prove this by induction. The base case at $t = -\infty$ holds by definition. Assume that this is true for some integer t. Then, we have

$$\begin{aligned}
\sum_{r=0}^{m} \boldsymbol{\theta}_{t+1}^{(r)} + \boldsymbol{\delta}_{t+1}^{(m)} &= (\boldsymbol{I} - \eta \boldsymbol{x}_t \otimes \boldsymbol{x}_t)\left(\sum_{r=0}^{m} \boldsymbol{\theta}_t^{(r)} + \boldsymbol{\delta}_t^{(m)}\right) + \eta \xi_t \boldsymbol{x}_t - \eta \sum_{\tau=0}^{\infty} \beta_\tau \boldsymbol{w}_{t-\tau} \\
&= (\boldsymbol{I} - \eta \boldsymbol{x}_t \otimes \boldsymbol{x}_t)\boldsymbol{\theta}'_t + \eta \xi_t \boldsymbol{x}_t - \eta \sum_{\tau=0}^{\infty} \beta_\tau \boldsymbol{w}_{t-\tau} = \boldsymbol{\theta}'_{t+1}\,.
\end{aligned}$$

□

The idea behind the proof is to show that $\mathbb{E}\left[\boldsymbol{\delta}_0^{(m)} \otimes \boldsymbol{\delta}_0^{(m)}\right] \to \boldsymbol{0}$ as $m \to \infty$. Then, we can use the triangle inequality to bound

$$\|\boldsymbol{\theta}'_t\| \leq \sum_{r=0}^{\infty} \left\|\boldsymbol{\theta}_t^{(r)}\right\|,$$

where the stationary error of the right side can be obtained from analyzing the LTI systems defined in (29).

C.2.2 Part 2: Characterize the Transfer Function of each LTI System

There are two LTI systems. First, $\boldsymbol{\theta}_t^{(r)}$ for $r > 0$ is an LTI system

$$\boldsymbol{z}_{t+1} = (\boldsymbol{I} - \eta \boldsymbol{H})\boldsymbol{z}_t + \eta \boldsymbol{u}_t \tag{30}$$

with input $\boldsymbol{u}_t \in \mathbb{R}^d$ and output $\boldsymbol{z}_t \in \mathbb{R}^d$. Second, $\boldsymbol{\theta}_t^{(0)}$ satisfies satisfies an LTI system

$$\boldsymbol{z}_{t+1} = (\boldsymbol{I} - \eta \boldsymbol{H})\boldsymbol{z}_t + \eta \boldsymbol{u}_t - \eta \sum_{\tau=0}^{\infty} \beta_t \boldsymbol{w}_{t-\tau} \tag{31}$$

with inputs $(\boldsymbol{u}_t, \boldsymbol{w}_t) \in \mathbb{R}^d \times \mathbb{R}^d$ and output $\boldsymbol{z}_t \in \mathbb{R}^d$ where the weights $\boldsymbol{\beta} \in \ell^2$ is assumed given.

We now characterize the transfer functions of these LTI systems; see Appendix F.1 for a review.

Property C.8. *The LTI sytem* (30) *is $\boldsymbol{G}(\omega) = -\eta \boldsymbol{M}_\omega \in \mathbb{C}^{d \times d}$, where $\boldsymbol{M}_\omega$ is defined in Equation* (28)*. Moreover, this system is asymptotically stable as long as $\boldsymbol{0} \prec \eta \boldsymbol{H} \prec \boldsymbol{I}$.*

Proof. Let $\boldsymbol{U}(\omega) \in \mathbb{C}^d$ and $\boldsymbol{Z}(\omega) \in \mathbb{C}^d$ be the Fourier transforms of $\boldsymbol{u}_t$ and $\boldsymbol{z}_t$ respectively. The transfer function must hold for any input-output sequences, so we can choose some sequences and solve for the transfer functions. It is convenient to consider the delta spike on a standard basis (upto scaling), i.e., $\boldsymbol{U} = 2\pi\delta_\omega \boldsymbol{e}_j$, where δ_ω is the Dirac delta at ω, and $\boldsymbol{e}_j$ is the j^{th} standard basis vector in $\mathbb{R}^d$. This gives $\boldsymbol{Z} = 2\pi \boldsymbol{g}_j \delta_\omega$ where $\boldsymbol{g}_j(\cdot)$ is the j^{th} column of $\boldsymbol{G}(\cdot)$.

To move back to time domain, we take an inverse Fourier transform to get $\boldsymbol{u}_t = \exp(i\omega t)\boldsymbol{e}_j$ and $\boldsymbol{z}_t = \boldsymbol{g}_j(\omega)\exp(i\omega t)$. Plugging this into the update (30) gives and solving for $\boldsymbol{g}_j(\omega)$ gives $\boldsymbol{g}_j(\omega) = -\eta \boldsymbol{M}_\omega \boldsymbol{e}_j$. Stacking these into a matrix gives the expression.

If $\boldsymbol{u}_t \equiv \boldsymbol{0}$ for all t, then $\|\boldsymbol{z}_{t+1}\|_2 \le \|\boldsymbol{I} - \eta \boldsymbol{H}\|_2 \|\boldsymbol{z}_t\|_2 < \|\boldsymbol{z}_t\|_2$ since $\|\boldsymbol{I} - \eta\boldsymbol{H}\|_2 < 1$. Hence, $\|\boldsymbol{z}_t\|_2 \to 0$ giving the asymptotic stability of the system. □

Property C.9. *The transfer function of the LTI system* (31) *is*

$$\widetilde{\boldsymbol{G}}(\omega) = [\boldsymbol{G}(\omega) \quad \boldsymbol{G}'(\omega)] \in \mathbb{C}^{d \times 2d}$$

where $\boldsymbol{G}(\omega) = -\eta \boldsymbol{M}_\omega$ *and* $\boldsymbol{G}'(\omega) = \eta B(\omega)\boldsymbol{M}_\omega$ *with* $B(\omega)$ *as the DTFT of* $\boldsymbol{\beta}$*. Moreover, this system is asymptotically stable as long as* $\boldsymbol{0} \prec \eta \boldsymbol{H} \prec \boldsymbol{I}$*.*

Proof. The expression for $\boldsymbol{G}(\omega)$ is the same as in Property C.8. To find $\boldsymbol{G}'$, we set the Fourier transforms $\boldsymbol{U} \equiv \boldsymbol{0}$, $\boldsymbol{W} = 2\pi\delta_\omega \boldsymbol{e}_j$ so that $\boldsymbol{Z} = 2\pi\delta_\omega \boldsymbol{g}'_j$, where $\boldsymbol{g}'_j(\cdot)$ is the j^{th} column of $\boldsymbol{G}'(\cdot)$.

An inverse Fourier transforms gives the time domain versions $\boldsymbol{w}_t = \exp(i\omega t)$, $\boldsymbol{u}_t \equiv \boldsymbol{0}$, $\boldsymbol{z}_t = \exp(i\omega t)\boldsymbol{g}'_j(\omega)$. Plugging these into (31) and plugging in the definition of $B(\omega)$ gives the expression for the transfer function. Its asymptotic stability holds similar to Property C.8. □

C.2.3 Part 3: Compute the Stationary Covariance of each LTI System

The stationary covariance of an LTI system driven by white noise. A sequence $(\boldsymbol{u}_t)$ is said to be a white noise process if it is mean zero and $\mathbb{E}[\boldsymbol{u}_t \boldsymbol{u}_\tau] = \boldsymbol{0}$ for $t \neq \tau$. This is true for both $\boldsymbol{\theta}_t^{(0)}$ as well $\boldsymbol{\theta}_t^{(r)}$ for $r > 0$. Since we care about the stationary distribution and we start at $t = -\infty$, we have reached the steady state at $t = 0$. So, we compute $\mathbb{E}[\boldsymbol{\theta}_0^{(r)} \otimes \boldsymbol{\theta}_0^{(r)}]$.

Stationary covariance of the base recursion: We first start with $\boldsymbol{\theta}_t^{(0)}$.

Proposition C.10. *We have that* $\mathbb{E}\left[\boldsymbol{\theta}_t^{(0)} \otimes \boldsymbol{\theta}_t^{(0)}\right]$ *is equal for all* $t > -\infty$ *and is bounded as*

$$\mathbb{E}\left[\boldsymbol{\theta}_t^{(0)} \otimes \boldsymbol{\theta}_t^{(0)}\right] \preceq \eta\sigma_{\mathsf{sgd}}^2 \boldsymbol{I} + \eta\sigma^2\, \boldsymbol{H}^{-1/2} \boldsymbol{P}_{\boldsymbol{\beta}} \boldsymbol{H}^{-1/2}\,,$$

where $\boldsymbol{P}_{\boldsymbol{\beta}}$ *is defined in Equation* (26) *and we denote* $\sigma^2 = G^2\gamma_\infty(\boldsymbol{\beta})^2/(2\rho)$*.*

Proof. The input $(\xi_t \boldsymbol{x}_t, \boldsymbol{w}_t)$ forms a white noise sequence, since for $t \neq \tau$, we have $\mathbb{E}[\xi_t \boldsymbol{x}_t \xi_\tau \boldsymbol{x}_\tau] = \mathbb{E}[\xi_t \boldsymbol{x}_t]\,\mathbb{E}[\xi_\tau \boldsymbol{x}_\tau] = \boldsymbol{0}$ (since $\xi_t \boldsymbol{x}_t$ for each t is i.i.d.) and $\mathbb{E}[\boldsymbol{w}_t \boldsymbol{w}_\tau] = \boldsymbol{0}$. The covariance of the input is

$$\mathbb{E}[(\xi_t \boldsymbol{x}_t, \boldsymbol{w}_t) \otimes (\xi_t \boldsymbol{x}_t, \boldsymbol{w}_t)] = \begin{bmatrix} \mathbb{E}[\xi_t^2 \boldsymbol{x}_t \boldsymbol{x}_t] & \boldsymbol{0} \\ \boldsymbol{0} & \mathbb{E}[\boldsymbol{w}_t \otimes \boldsymbol{w}_t] \end{bmatrix} = \mathbb{E}[(\xi_\tau \boldsymbol{x}_\tau, \boldsymbol{w}_\tau) \otimes (\xi_\tau \boldsymbol{x}_\tau, \boldsymbol{w}_\tau)]$$

for all t, τ. This is further bounded by Assumption (**B1**) as

$$\mathbb{E}[(\xi_t \boldsymbol{x}_t, \boldsymbol{w}_t) \otimes (\xi_t \boldsymbol{x}_t, \boldsymbol{w}_t)] \preceq \begin{bmatrix} \sigma_{\mathsf{sgd}}^2 \boldsymbol{H} & \boldsymbol{0} \\ \boldsymbol{0} & \sigma^2 \boldsymbol{I} \end{bmatrix}$$

The output covariance of the asymptotically stable LTI system (31) can be given in terms of the transfer function $\widetilde{\boldsymbol{G}}(\omega) = [\boldsymbol{G}(\omega) \quad \boldsymbol{G}'(\omega)]$ characterized in Property C.9 using Theorem F.2. This gives that $\mathbb{E}\left[\boldsymbol{\theta}_t^{(0)} \otimes \boldsymbol{\theta}_t^{(0)}\right]$ is equal for each $t > -\infty$ and is bounded as

$$\mathbb{E}\left[\boldsymbol{\theta}_t^{(0)} \otimes \boldsymbol{\theta}_t^{(0)}\right] \preceq \frac{1}{2\pi} \int_{-\pi}^{\pi} \left(\eta^2 \sigma_{\mathsf{sgd}}^2 \boldsymbol{M}_\omega \boldsymbol{H} \boldsymbol{M}_\omega^* + \eta^2 \sigma^2 |B(\omega)|^2\, \boldsymbol{M}_\omega \boldsymbol{M}_\omega^*\right) \mathrm{d}\omega\,. \tag{32}$$

With the eigenvalue decomposition $\boldsymbol{H} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$, we get $\boldsymbol{M}_\omega = \boldsymbol{U}\big((1-\exp(i\omega))\boldsymbol{I} - \eta\boldsymbol{\Lambda}\big)^{-1}\boldsymbol{U}^\top$. This gives

$$\boldsymbol{M}_\omega \boldsymbol{H} \boldsymbol{M}_\omega^* = \boldsymbol{U}\,\mathsf{diag}\left(\left(\lambda_j/|1-\exp(i\omega)-\eta\lambda_j|^2\right)_{j=1}^d\right)\boldsymbol{U}^\top\,.$$

We invoke Corollary C.5 to say

$$\begin{aligned}\int_{-\pi}^{\pi} \boldsymbol{M}_\omega \boldsymbol{H} \boldsymbol{M}_\omega^* \mathrm{d}\omega &= \boldsymbol{U}\,\mathsf{diag}\left(\left(\int_{-\pi}^{\pi}\mathrm{d}\omega\,\lambda_j/|1-\exp(i\omega)-\eta\lambda_j|^2\right)_{j=1}^d\right)\boldsymbol{U}^\top \\ &\preceq \boldsymbol{U}\,\mathsf{diag}\left(\left(2\pi/\eta\right)_{j=1}^d\right)\boldsymbol{U}^\top = \frac{2\pi}{\eta}\boldsymbol{I}\,. \end{aligned} \tag{33}$$

Similarly, we invoke Lemma C.4 to compute

$$\begin{aligned}\int_{-\pi}^{\pi} |B(\omega)|^2 \boldsymbol{M}_\omega \boldsymbol{M}_\omega^* \mathrm{d}\omega &= \boldsymbol{U}\,\mathsf{diag}\left(\left(\int_{-\pi}^{\pi}\mathrm{d}\omega\,|B(\omega)|^2/|1-\exp(i\omega)-\eta\lambda_j|^2\right)_{j=1}^d\right)\boldsymbol{U}^\top \\ &\preceq \boldsymbol{U}\,\mathsf{diag}\left(\left(2\pi\langle\boldsymbol{\beta},\boldsymbol{T}_j\boldsymbol{\beta}\rangle/(\eta\lambda_j)\right)_{j=1}^d\right)\boldsymbol{U}^\top \\ &= \frac{2\pi}{\eta}\boldsymbol{U}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{U}^\top = \frac{2\pi}{\eta}\boldsymbol{H}^{-1/2}\boldsymbol{P}_{\boldsymbol{\beta}}\boldsymbol{H}^{-1/2}\,, \end{aligned} \tag{34}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ and $\boldsymbol{P}_{\boldsymbol{\beta}}$ are defined in (26). Plugging in (33) and (33) into (32) completes the proof of the upper bound. □

Stationary covariance of the higher-order recursion: Next, we turn to $\boldsymbol{\theta}_t^{(r)}$.

Proposition C.11. *For any $r \geq 1$, we have*

$$\mathbb{E}\left[\boldsymbol{\theta}_0^{(r)} \otimes \boldsymbol{\theta}_0^{(r)}\right] \preceq \eta\left(\eta R^2\right)^r\left(\sigma_{\mathsf{sgd}}^2 + \frac{C_{\mathsf{kurt}}\sigma^2}{R^2}\langle\boldsymbol{\beta},\boldsymbol{T}\boldsymbol{\beta}\rangle\right)\,.$$

Proof. Follows from combining Proposition C.10 with the more general Lemma C.12 below. □

Lemma C.12. *For some $r \geq 1$, suppose that $\mathbb{E}\left[\boldsymbol{\theta}_t^{(r-1)} \otimes \boldsymbol{\theta}_t^{(r-1)}\right]$ is equal for each t and is bounded as $\mathbb{E}\left[\boldsymbol{\theta}_t^{(r-1)} \otimes \boldsymbol{\theta}_t^{(r-1)}\right] \preceq a\boldsymbol{I} + b\boldsymbol{H}^{-1/2}\boldsymbol{P}_{\boldsymbol{\beta}}\boldsymbol{H}^{-1/2}$ for some scalars $a, b \geq 0$. Then, we have the following.*

(a) We have that $\boldsymbol{\zeta}_t^{(r)} := (\boldsymbol{H} - \boldsymbol{x}_t \otimes \boldsymbol{x}_t)\,\boldsymbol{\theta}_t^{(r-1)}$ is a white-noise process with

$$\mathbb{E}\left[\boldsymbol{\zeta}_t^{(r)} \otimes \boldsymbol{\zeta}_t^{(r)}\right] \preceq \left(aR^2 + bC_{\mathsf{kurt}}\langle\boldsymbol{\beta},\boldsymbol{T}\boldsymbol{\beta}\rangle\right)\boldsymbol{H}\,.$$

(b) We have that $\mathbb{E}\left[\boldsymbol{\theta}_t^{(r)} \otimes \boldsymbol{\theta}_t^{(r)}\right]$ is equal for each t and is bounded as

$$\mathbb{E}\left[\boldsymbol{\theta}_t^{(r)} \otimes \boldsymbol{\theta}_t^{(r)}\right] \preceq \eta\left(aR^2 + bC_{\mathsf{kurt}}\langle\boldsymbol{\beta},\boldsymbol{T}\boldsymbol{\beta}\rangle\right)\boldsymbol{I}\,.$$

Proof. Note that $\mathbb{E}\left[\boldsymbol{\zeta}_t^{(r)} \otimes \boldsymbol{\zeta}_\tau^{(r)}\right] = \boldsymbol{0}$ for $t \neq \tau$ since $\boldsymbol{x}_t$ is independent of $\boldsymbol{x}_\tau$ and $\mathbb{E}[\boldsymbol{x}_t \otimes \boldsymbol{x}_t] = \boldsymbol{H}$. Since $\boldsymbol{x}_t$ is independent of $\boldsymbol{\theta}_t^{(r-1)}$, we get from the tower rule of expectations that

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{\zeta}_t^{(r)} \otimes \boldsymbol{\zeta}_t^{(r)}\right] &= \mathbb{E}\left[(\boldsymbol{H} - \boldsymbol{x}_t \otimes \boldsymbol{x}_t)\,\left(\boldsymbol{\theta}_t^{(r-1)} \otimes \boldsymbol{\theta}_t^{(r-1)}\right)\,(\boldsymbol{H} - \boldsymbol{x}_t \otimes \boldsymbol{x}_t)\right] \\ &= \mathbb{E}\left[(\boldsymbol{H} - \boldsymbol{x}_t \otimes \boldsymbol{x}_t)\,\mathbb{E}\left[\boldsymbol{\theta}_t^{(r-1)} \otimes \boldsymbol{\theta}_t^{(r-1)}\right]\,(\boldsymbol{H} - \boldsymbol{x}_t \otimes \boldsymbol{x}_t)\right]\,, \end{aligned}$$

or that $(\boldsymbol{\zeta}_t^{(r)})$ is a white noise process. Its covariance can further be bounded as

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{\zeta}_t^{(r)} \otimes \boldsymbol{\zeta}_t^{(r)}\right] &\preceq \mathbb{E}\left[(\boldsymbol{H} - \boldsymbol{x}_t \otimes \boldsymbol{x}_t)\,\left(a\boldsymbol{I} + b\boldsymbol{H}^{-1/2}\boldsymbol{P}_{\boldsymbol{\beta}}\boldsymbol{H}^{-1/2}\right)\,(\boldsymbol{H} - \boldsymbol{x}_t \otimes \boldsymbol{x}_t)\right] \\ &\preceq a\,\mathbb{E}\left[\|\boldsymbol{x}_t\|_2^2\,(\boldsymbol{x}_t \otimes \boldsymbol{x}_t)\right] + b\,\mathbb{E}\left[(\boldsymbol{x}_t \otimes \boldsymbol{x}_t)\boldsymbol{H}^{-1/2}\boldsymbol{P}_{\boldsymbol{\beta}}\boldsymbol{H}^{-1/2}(\boldsymbol{x}_t \otimes \boldsymbol{x}_t)\right) \\ &\preceq aR^2\boldsymbol{H} + bC_{\mathsf{kurt}}\,\mathsf{Tr}\left[\boldsymbol{P}_{\boldsymbol{\beta}}\right]\boldsymbol{H}\,, \end{aligned}$$

where the last inequality followed from Assumption (**B3**). Further, note that $\mathsf{Tr}\left[\boldsymbol{P}_{\boldsymbol{\beta}}\right] = \langle \boldsymbol{\beta}, \boldsymbol{T}\boldsymbol{\beta}\rangle$ from (27).

The output covariance of the asymptotically stable LTI system (30) can be given in terms of the transfer function $\boldsymbol{G}(\omega) = -\eta \boldsymbol{M}_\omega$ using Theorem F.2 as

$$\mathbb{E}\left[\boldsymbol{\theta}_t^{(r)} \otimes \boldsymbol{\theta}_t^{(r)}\right] \preceq \frac{\eta^2\left(aR^2 + bC_{\mathsf{kurt}}\langle \boldsymbol{\beta}, \boldsymbol{T}\boldsymbol{\beta}\rangle\right)}{2\pi}\int_{-\pi}^{\pi} \boldsymbol{M}_\omega \boldsymbol{H}\boldsymbol{M}_\omega^* \,\mathrm{d}\omega \overset{(33)}{\preceq} \eta\left(aR^2 + bC_{\mathsf{kurt}}\langle \boldsymbol{\beta}, \boldsymbol{T}\boldsymbol{\beta}\rangle\right)\boldsymbol{I}\,.$$

□

Remainder Term: It remains to show that the remainder term $\boldsymbol{\delta}_t$ can be neglected by taking $m \to \infty$.

Proposition C.13. *We have* $\lim_{m\to\infty} \mathbb{E}\left[\boldsymbol{\delta}_t^{(m)} \otimes \boldsymbol{\delta}_t^{(m)}\right] = \mathbf{0}$.

Proof. Let $\boldsymbol{\zeta}_t^{(m+1)} := (\boldsymbol{H} - \boldsymbol{x}_t \otimes \boldsymbol{x}_t)\,\boldsymbol{\theta}_t^{(m)}$. By Lemma C.12 and Proposition C.11, we have $\boldsymbol{\zeta}_t$ is a white-noise process with

$$\mathbb{E}\left[\boldsymbol{\zeta}_t^{(m+1)} \otimes \boldsymbol{\zeta}_t^{(m+1)}\right] \preceq (\eta R^2)^{m+1}\left(\sigma_{\mathsf{sgd}}^2 + \frac{C_{\mathsf{kurt}}\sigma^2}{R^2}\langle \boldsymbol{\beta}, \boldsymbol{T}\boldsymbol{\beta}\rangle\right)\boldsymbol{H} \to \mathbf{0}$$

as $m \to \infty$ since $\eta < 1/R^2$. Note that the update for $\boldsymbol{\delta}_t^{(m)}$ exactly matches that of SGD (without added DP noise), and the noise covariance is $\mathbf{0}$. The statement of this result is equivalent to showing that the stationary covariance of SGD with zero residuals is zero. This observation is formalized in Lemma 4 of [29] (see also Theorem F.3 of Appendix F), which gives for any t that

$$\mathbf{0} \preceq \mathbb{E}[\boldsymbol{\delta}_t^{(m)} \otimes \boldsymbol{\delta}_t^{(m)}] \preceq \frac{\eta}{1-\eta R^2}\left[(\eta R^2)^{m+1}\left(\sigma_{\mathsf{sgd}}^2 + \frac{C_{\mathsf{kurt}}\sigma^2}{R^2}\langle \boldsymbol{\beta}, \boldsymbol{T}\boldsymbol{\beta}\rangle\right)\right]\boldsymbol{I} \to \mathbf{0}$$

as $m \to \infty$. □

C.2.4 Part 4: Combining the Errors

Time-domain description: We now state and prove a time-domain description of the upper bound of Equation (13).

Theorem C.14. *Suppose Assumption C.2 holds. Consider the sequence* $(\boldsymbol{\theta}_t)_{t=-\infty}^{\infty}$ *produced by the Noisy-FTRL update in Equation* (18) *with some given weights* $\boldsymbol{\beta} \in \ell^2$ *and noise variance* $\boldsymbol{w}_t \sim \mathcal{N}(\mathbf{0}, G^2\gamma_\infty(\boldsymbol{\beta})^2/(2\rho)\boldsymbol{I})$. *If the learning rate satisfies* $\eta < 1/R^2$, *we have*

$$F_\infty(\boldsymbol{\beta}) \le \left(1 + \left(1 - \sqrt{\eta R^2}\right)^{-2}\right)\eta R^2\sigma_{\mathsf{sgd}}^2 + \left(1 + C_{\mathsf{kurt}}\left(1 - \sqrt{\eta R^2}\right)^{-2}\right)\frac{\eta G^2\gamma_\infty(\boldsymbol{\beta})^2}{2\rho}\langle \boldsymbol{\beta}, \boldsymbol{T}\boldsymbol{\beta}\rangle\,.$$

Proof. We use shorthand $\sigma^2 = \frac{G^2\gamma_\infty(\boldsymbol{\beta})^2}{2\rho}$. First, note that $\eta < 1/R^2$ also implies that $\eta\lambda_j < 1$ for each eigenvalue λ_j of $\boldsymbol{H}$. The right side is well-defined since Lemma F.17 gives

$$|\langle \boldsymbol{\beta}, \boldsymbol{T}\boldsymbol{\beta}\rangle| \le \sum_{j=1}^{d}\left|\sum_{t=0}^{\infty}\sum_{\tau=0}^{\infty}\beta_t\beta_\tau(1-\eta\lambda_j)^{|t-\tau|}\right| \le \|\boldsymbol{\beta}\|_2^2\sum_{j=1}^{d}\frac{2}{\eta\lambda_j} < \infty \tag{35}$$

for $\beta \in \ell^2$. Next, using Proposition C.10, $\mathsf{Tr}\left[\boldsymbol{H}\right] \le R^2$, and $\mathsf{Tr}\left[\boldsymbol{P}_{\boldsymbol{\beta}}\right] = \langle \boldsymbol{\beta}, \boldsymbol{T}\boldsymbol{\beta}\rangle$, we get

$$\mathbb{E}\left\|\boldsymbol{\theta}_0^{(0)}\right\|_{\boldsymbol{H}}^2 = \mathsf{Tr}\left[\boldsymbol{H}\mathbb{E}\left[\boldsymbol{\theta}_0^{(0)} \otimes \boldsymbol{\theta}_0^{(0)}\right]\right] \le \eta R^2\sigma_{\mathsf{sgd}}^2 + \eta\sigma^2\langle \boldsymbol{\beta}, \boldsymbol{T}\boldsymbol{\beta}\rangle\,. \tag{36}$$

Similarly, using Proposition C.11, we get for $r \ge 1$ that

$$\mathbb{E}\left\|\boldsymbol{\theta}_0^{(r)}\right\|_{\boldsymbol{H}}^2 \le (\eta R^2)^{r+1}\left(\sigma_{\mathsf{sgd}}^2 + \frac{C_{\mathsf{kurt}}\sigma^2}{R^2}\langle \boldsymbol{\beta}, \boldsymbol{T}\boldsymbol{\beta}\rangle\right)\,.$$

We can ignore the remainder term since $\mathbb{E}\left\|\boldsymbol{\delta}_t^{(m)}\right\|_{\boldsymbol{H}}^2 \to 0$ as $m \to \infty$, from Proposition C.13. Thus, we get using Property C.7 and the triangle inequality on the norm $\boldsymbol{u} \mapsto \sqrt{\mathbb{E}\langle \boldsymbol{u}, \boldsymbol{H}\boldsymbol{u}\rangle}$ of a random vector $\boldsymbol{u}$ to get

$$\sqrt{\mathbb{E}\|\boldsymbol{\theta}_0'\|_{\boldsymbol{H}}^2} \leq \sum_{r=0}^{\infty} \sqrt{\mathbb{E}\left\|\boldsymbol{\theta}_0^{(r)}\right\|_{\boldsymbol{H}}^2}.$$

To complete the proof, we plug in Equations (35) and (36) and sum up the infinite series. We simplify the result using $\|\boldsymbol{x}+\boldsymbol{y}\|_{\boldsymbol{H}}^2 \leq 2\|\boldsymbol{x}\|_{\boldsymbol{H}}^2 + 2\|\boldsymbol{y}\|_{\boldsymbol{H}}^2$ and use $F(\boldsymbol{\theta}) - F(\boldsymbol{\theta}_\star) = (1/2)\|\boldsymbol{\theta}-\boldsymbol{\theta}_\star\|_{\boldsymbol{H}}^2$. □

Frequency-domain description: We now state and prove the frequency domain description of the upper bound (13).

Theorem C.15. *Consider the setting of Theorem C.14. If $B \in L^2$, i.e., $\int_{-\pi}^{\pi} |B(\omega)|^2 \,\mathrm{d}\omega < \infty$, we have*

$$\begin{aligned} F_\infty(B) \leq & \left(1 + \left(1 - \sqrt{\eta R^2}\right)^{-2}\right) \eta R^2 \sigma_{\mathsf{sgd}}^2 \\ & + \left(1 + C_{\mathsf{kurt}}\left(1 - \sqrt{\eta R^2}\right)^{-2}\right) \frac{\eta^2 G^2 \gamma_\infty(B)^2}{2\pi\rho} \int_{-\pi}^{\pi} |B(\omega)|^2 \, h(\omega) \,\mathrm{d}\omega \,. \end{aligned}$$

Proof. We again use the shorthand $\sigma^2 = \frac{G^2\gamma_\infty(\boldsymbol{\beta})^2}{2\rho}$. First note that

$$h(\omega) \leq \sum_{j=1}^{d} \frac{\lambda_j}{1 + (1-\eta\lambda_j)^2 - 2(1-\eta\lambda_j)} = \sum_{j=1}^{d} \frac{1}{\eta^2\lambda_j} = \frac{\mathsf{Tr}\left[\boldsymbol{H}^{-1}\right]}{\eta^2} \,.$$

Thus, the right-side is well-defined since

$$\int_{-\pi}^{\pi} |B(\omega)|^2 \, h(\omega) \mathrm{d}\omega \leq \frac{\mathsf{Tr}\left[\boldsymbol{H}^{-1}\right]}{\eta^2} \int_{-\pi}^{\pi} |B(\omega)|^2 \,\mathrm{d}\omega < \infty$$

by assumption. We use Lemma C.4 to get

$$\langle \boldsymbol{\beta}, \boldsymbol{T}\boldsymbol{\beta}\rangle = \sum_{j=1}^{d} \langle \boldsymbol{\beta}, \boldsymbol{T}_j\boldsymbol{\beta}\rangle \leq \sum_{j=1}^{d} \frac{\eta\lambda_j}{\pi} \int_{-\pi}^{\pi} \frac{|B(\omega)|^2 \mathrm{d}\omega}{|1 - \exp(i\omega) - \eta\lambda_j|^2} = \frac{\eta}{\pi} \int_{-\pi}^{\pi} |B(\omega)|^2 \, h(\omega) \,\mathrm{d}\omega \,.$$

□

C.3 Proofs of Lower Bounds on the Asymptotic Suboptimality

We now state and prove the lower bound part of (13) on the asymptotic suboptimality.

Assumption C.16. *In addition to Assumption C.2, the data distribution $\mathbb{P}_{\mathsf{data}}$ satisfies the following:*

- ***(A2') Worst-Case Residuals**: For $(\boldsymbol{x}, y) \sim \mathbb{P}_{\mathsf{data}}$, the residual $\xi := y - \langle \boldsymbol{\theta}_\star, \boldsymbol{x}\rangle$ has variance $\mathbb{E}[\xi^2] = \sigma_{\mathsf{sgd}}^2$.*

Note that the variance of ξ^2 holds with equality under Assumption C.16.

Theorem C.17. *Suppose Assumption C.16 holds. Consider the sequence $(\boldsymbol{\theta}_t)_{t=-\infty}^{\infty}$ produced by the Noisy-FTRL update in Equation* (18) *with some given weights $\boldsymbol{\beta} \in \ell^1$. If the learning rate satisfies $\eta < 1/R^2$, we have*

$$F_\infty(\boldsymbol{\beta}) \geq \frac{\eta\sigma_{\mathsf{sgd}}^2}{2}\mathsf{Tr}\left[\boldsymbol{H}\right] + \frac{\eta^2 G^2 \gamma_\infty(B)^2}{4\pi\rho} \int_{-\pi}^{\pi} |B(\omega)|^2 \, h(\omega) \,\mathrm{d}\omega \geq \frac{\eta\sigma_{\mathsf{sgd}}^2}{2}\mathsf{Tr}\left[\boldsymbol{H}\right] + \frac{\eta G^2 \gamma_\infty(\boldsymbol{\beta})^2}{4\rho} \langle \boldsymbol{\beta}, \boldsymbol{T}\boldsymbol{\beta}\rangle \,,$$

where $h(\omega)$ is defined in (14) *and $\boldsymbol{T}$ is defined in* (24)*. Furthermore, the minimal stationary error over all choices of $\boldsymbol{\beta}$ is bounded as*

$$\inf_{\boldsymbol{\beta}} \; F_\infty(\boldsymbol{\beta}) \geq \frac{1}{4}\left(2\eta\sigma_{\mathsf{sgd}}^2 + \frac{\eta^2 G^2}{2\rho}\right)\mathsf{Tr}\left[\boldsymbol{H}\right]$$

where the infimum is attained by $\boldsymbol{\beta}_\star$ whose DTFT $B_\star$ verifies $|B_\star(\omega)|^2 = 1/\sqrt{h(\omega)}$.

Note that we assume $\boldsymbol{\beta} \in \ell^1$, i.e., $\|\boldsymbol{\beta}\|_1 = \sum_{\tau=0}^{\infty} |\beta_\tau| < \infty$ for technical reasons. This implies that $\boldsymbol{\beta} \in \ell^2$, which we assumed for the upper bounds.

The key idea behind the proof is that the variance of $\boldsymbol{\theta}'_t$ is no smaller than that of an LTI system with $\boldsymbol{x}_t \otimes \boldsymbol{x}_t$ replaced by its expectation $\boldsymbol{H}$. We can quantify this latter covariance with equality under Assumption C.16. We set up some notation and develop some preliminary results before proving this theorem.

Formally, consider the sequences $(\boldsymbol{\theta}_t^{(0)})_{t=-\infty}^{\infty}$ and $(\boldsymbol{\delta}_t^{(0)})_{t=-\infty}^{\infty}$ as defined in (29) (cf. §C.2.1). They start at $t = -\infty$ from $\boldsymbol{\theta}_t^{(0)} = \boldsymbol{\theta}'_t$ and $\boldsymbol{\delta}_t^{(0)} = \mathbf{0}$. By Property C.7, we these satisfy $\boldsymbol{\theta}'_t = \boldsymbol{\theta}_t^{(0)} + \boldsymbol{\delta}_t^{(0)}$.

We use a technical result that $\boldsymbol{\theta}_t^{(0)}$ and $\boldsymbol{\delta}_t$ are uncorrelated. It is proved at the end of this section.

Proposition C.18. *Consider the setting of Theorem C.17. We have for all t that*

$$\mathbb{E}\left[\boldsymbol{\theta}_t^{(0)} \otimes \boldsymbol{\delta}_t^{(0)}\right] = \mathbf{0}\,.$$

We now give the proof of Theorem C.17.

Proof of Theorem C.17. We use shorthand $\sigma^2 = \frac{G^2 \gamma_\infty(\boldsymbol{\beta})^2}{2\rho}$. Since $\boldsymbol{\theta}'_t = \boldsymbol{\theta}_t^{(0)} + \boldsymbol{\delta}_t^{(0)}$, we have

$$\mathbb{E}\left[\boldsymbol{\theta}'_t \otimes \boldsymbol{\theta}'_t\right] = \mathbb{E}\left[\boldsymbol{\theta}_t^{(0)} \otimes \boldsymbol{\theta}_t^{(0)}\right] + \mathbb{E}\left[\boldsymbol{\delta}_t^{(0)} \otimes \boldsymbol{\delta}_t^{(0)}\right] \succeq \mathbb{E}\left[\boldsymbol{\theta}_t^{(0)} \otimes \boldsymbol{\theta}_t^{(0)}\right] \tag{37}$$

where the cross terms disappear from Proposition C.18 for the first equality. We can get an expression for this term by following the proof of Proposition C.10: under Assumption C.16, we have that Equation (32) holds with equality. Thus, we get for all $t > -\infty$ that

$$\begin{aligned} F_\infty(B) &= \mathsf{Tr}\left[\boldsymbol{H}\,\mathbb{E}\left[\boldsymbol{\theta}'_t \otimes \boldsymbol{\theta}'_t\right]\right] \succeq \mathsf{Tr}\left[\boldsymbol{H}\,\mathbb{E}\left[\boldsymbol{\theta}_t^{(0)} \otimes \boldsymbol{\theta}_t^{(0)}\right]\right] \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \left(\eta^2\sigma_{\mathsf{sgd}}^2 \mathsf{Tr}\left[\boldsymbol{H}^{1/2}\boldsymbol{M}_\omega \boldsymbol{H} \boldsymbol{M}_\omega^* \boldsymbol{H}^{1/2}\right] + \eta^2\sigma^2 |B(\omega)|^2\, \mathsf{Tr}\left[\boldsymbol{H}^{1/2}\boldsymbol{M}_\omega \boldsymbol{M}_\omega^* \boldsymbol{H}^{1/2}\right]\right) \mathrm{d}\omega\,. \end{aligned} \tag{38}$$

We invoke Corollary C.5 to obtain

$$\begin{aligned} \int_{-\pi}^{\pi} \mathsf{Tr}\left[\boldsymbol{H}^{1/2}\boldsymbol{M}_\omega \boldsymbol{H} \boldsymbol{M}_\omega^* \boldsymbol{H}^{1/2}\right] \mathrm{d}\omega &= \sum_{j=1}^{d} \lambda_j^2 \int_{-\pi}^{\pi} \frac{\mathrm{d}\omega}{|1 - \exp(i\omega) - \eta\lambda_j|^2} \\ &\geq \sum_{j=1}^{d} \frac{\pi\lambda_j}{\eta} = \frac{\pi}{\eta}\mathsf{Tr}\left[\boldsymbol{H}\right]\,. \end{aligned}$$

Similarly, we invoke Lemma C.4 to compute

$$\begin{aligned} \int_{-\pi}^{\pi} |B(\omega)|^2\, \mathsf{Tr}\left[\boldsymbol{H}^{1/2}\boldsymbol{M}_\omega \boldsymbol{M}_\omega^* \boldsymbol{H}^{1/2}\right] \mathrm{d}\omega &= \int_{-\pi}^{\pi} \left(\sum_{j=1}^{d} |B(\omega)|^2 \frac{\lambda_j}{|1 - \exp(i\omega) - \eta\lambda_j|^2}\right) \mathrm{d}\omega \\ &= \int_{-\pi}^{\pi} |B(\omega)|^2\, h(\omega)\, \mathrm{d}\omega \geq \frac{\pi}{\eta}\langle \boldsymbol{\beta}, \boldsymbol{T}\boldsymbol{\beta}\rangle\,. \end{aligned}$$

This establishes the lower bound for specific choices of $\boldsymbol{\beta}$.

Now, we turn to the universal lower bound. Using the expression for $\gamma_\infty(B)$ from Property C.1, we get that the lower bound from the theorem statement is

$$F_\infty(B) \geq \frac{\eta\sigma_{\mathsf{sgd}}^2}{2}\mathsf{Tr}\left[\boldsymbol{H}\right] + \frac{\eta^2 G^2}{8\pi^2\rho}\left(\int_{-\pi}^{\pi} \frac{\mathrm{d}\omega}{|B(\omega)|^2}\right)\left(\int_{-\pi}^{\pi} |B(\omega)|^2 h(\omega)\right)\,. \tag{39}$$

The Cauchy-Schwarz inequality gives us that

$$\left(\int_{-\pi}^{\pi} \frac{\mathrm{d}\omega}{|B(\omega)|^2}\right)\left(\int_{-\pi}^{\pi} |B(\omega)|^2 h(\omega)\right) \geq \left(\int_{-\pi}^{\pi} \sqrt{h(\omega)}\,\mathrm{d}\omega\right)^2\,,$$

with equality attained for $|B(\omega)|^2 = 1/\sqrt{h(\omega)}$. This gives the universal lower bound on (39) over all possible choices of B (or equivalently, all possible choices of $\boldsymbol{\beta}$). To further lower bound this, we use $\cos(\omega) \geq -1$ to get

$$h(\omega) = \sum_{j=1}^{d} \frac{\lambda_j}{1 + (1-\eta\lambda_j)^2 - 2(1-\eta\lambda_j)\cos(\omega)} \geq \sum_{j=1}^{d} \frac{\lambda_j}{(2-\eta\lambda_j)^2} \geq \frac{1}{4}\sum_{j=1}^{d} \lambda_j = \frac{\mathsf{Tr}\left[\boldsymbol{H}\right]}{4}\,.$$

Thus, we get that (39) can be further lower bounded as

$$F_\infty(B) \geq \frac{\eta\sigma_{\mathsf{sgd}}^2}{2}\mathsf{Tr}\left[\boldsymbol{H}\right] + \frac{\eta^2 G^2}{8\pi^2\rho}\left(\int_{-\pi}^{\pi} \frac{\sqrt{\mathsf{Tr}\left[\boldsymbol{H}\right]}}{2}\,\mathrm{d}\omega\right)^2 = \frac{\eta\sigma_{\mathsf{sgd}}^2}{2}\mathsf{Tr}\left[\boldsymbol{H}\right] + \frac{\eta^2 G^2}{8\rho}\mathsf{Tr}\left[\boldsymbol{H}\right]\,.$$

□

Missing technical proofs in the lower bound: We now give the proof of Proposition C.18, which first relies on the following intermediate result.

Proposition C.19. *Consider the setting of Theorem C.17. We have for all t, τ that*

$$\mathbb{E}\left[\boldsymbol{w}_\tau \otimes \boldsymbol{\delta}_t^{(0)}\right] = \mathbf{0}\,.$$

Proof. For this proof, we start the sequences at $t = 0$ rather than $t = -\infty$. We drop the superscript to write $\boldsymbol{\delta}_t^{(0)}$ as $\boldsymbol{\delta}_t$. Define shorthand $\boldsymbol{Q}_t := \boldsymbol{I} - \eta\boldsymbol{x}_t \otimes \boldsymbol{x}_t$ and $\boldsymbol{R}_t := \boldsymbol{H} - \boldsymbol{x}_t \otimes \boldsymbol{x}_t$. We expand out the recursion to get

$$\begin{aligned}
\boldsymbol{\delta}_t &= \boldsymbol{Q}_{t-1}\boldsymbol{\delta}_{t-1} + \eta\boldsymbol{R}_{t-1}\boldsymbol{\theta}_{t-1}^{(0)} \\
&= \boldsymbol{Q}_{t-1}(\boldsymbol{Q}_{t-2}\boldsymbol{\delta}_{t-2} + \eta\boldsymbol{R}_{t-2}\boldsymbol{\theta}_{t-2}^{(0)}) + \eta\boldsymbol{R}_{t-1}\boldsymbol{\theta}_{t-1}^{(0)} \\
&= \boldsymbol{Q}_{t-1}\boldsymbol{Q}_{t-2}\cdots\boldsymbol{Q}_0\boldsymbol{\delta}_0 + \eta\left(\boldsymbol{R}_{t-1}\boldsymbol{\theta}_{t-1}^{(0)} + \boldsymbol{Q}_{t-1}\boldsymbol{R}_{t-2}\boldsymbol{\theta}_{t-2}^{(0)} + \cdots + \boldsymbol{Q}_{t-1}\cdots\boldsymbol{Q}_1\boldsymbol{R}_0\boldsymbol{\theta}_0^{(0)}\right)\,.
\end{aligned}$$

The first term is zero because $\boldsymbol{\delta}_0 = \mathbf{0}$ at initialization. Since $\boldsymbol{R}_\tau$ is mean zero and independent of $\boldsymbol{\theta}_\tau^{(0)}$ and $\boldsymbol{R}_t$ for $t > \tau$, we have

$$\begin{aligned}
\frac{1}{\eta}\mathbb{E}[\boldsymbol{\delta}_t \otimes \boldsymbol{w}_\tau] &= \mathbb{E}[\boldsymbol{R}_{t-1}]\mathbb{E}\left[\boldsymbol{\theta}_{t-1}^{(0)} \otimes \boldsymbol{w}_\tau\right] \\
&\qquad + \mathbb{E}[\boldsymbol{Q}_{t-1}]\,\mathbb{E}[\boldsymbol{R}_{t-2}]\mathbb{E}\left[\boldsymbol{\theta}_{t-2}^{(0)} \otimes \boldsymbol{w}_\tau\right] + \cdots + \mathbb{E}[\boldsymbol{Q}_{t-1}\cdots\boldsymbol{Q}_1]\,\mathbb{E}[\boldsymbol{R}_0]\mathbb{E}\left[\boldsymbol{\theta}_0^{(0)} \otimes \boldsymbol{w}_\tau\right] \\
&= \mathbf{0}\,,
\end{aligned}$$

giving us the desired result. □

Proof of Proposition C.18. We drop the superscript to write $\boldsymbol{\delta}_t^{(0)}$ as $\boldsymbol{\delta}_t$. We prove the claim by induction. At initialization, we have $\boldsymbol{\delta}_{-\infty} = \mathbf{0}$ so the hypothesis holds. Now assume that it holds at time t, i.e., $\mathbb{E}\left[\boldsymbol{\theta}_t^{(0)} \otimes \boldsymbol{\delta}_t\right] = \mathbf{0}$.

Next, we expand out $\mathbb{E}\left[\boldsymbol{\theta}_{t+1}^{(0)} \otimes \boldsymbol{\delta}_{t+1}\right]$ using their respective recursions. Note that $\boldsymbol{w}_t$, $\boldsymbol{H} - \boldsymbol{x}_t \otimes \boldsymbol{x}_t$ and ξ_t are each zero mean and independent of all quantities appearing up to iteration t (formally, they are independent of the σ-algebra generated by ($\boldsymbol{\theta}_t^{(0)}$ and $\boldsymbol{\delta}_t$). This gives

$$\mathbb{E}\left[\boldsymbol{\theta}_{t+1}^{(0)} \otimes \boldsymbol{\delta}_{t+1}\right] = (\boldsymbol{I} - \eta\boldsymbol{H})\mathbb{E}\left[\boldsymbol{\theta}_t^{(0)} \otimes \boldsymbol{\delta}_t\right](\boldsymbol{I} - \eta\boldsymbol{H}) - \eta\mathbb{E}\left[\sum_{\tau=0}^{\infty}\beta_\tau\left(\boldsymbol{w}_{t-\tau} \otimes \delta_t^{(0)}\right)\right](\boldsymbol{I} - \eta\boldsymbol{H})\,. \tag{40}$$

The first term is zero by the induction hypothesis. For the second term, we can interchange the expectation and the infinite sum by the Fubini-Tonelli theorem since

$$\sum_{\tau=0}^{\infty}|\beta_\tau|\;\mathbb{E}\left|\left\langle \boldsymbol{w}_{t-\tau}, \boldsymbol{\delta}_t^{(0)}\right\rangle\right| \leq \|\boldsymbol{\beta}\|_1 \max_{\tau=0,\ldots,\infty}\mathbb{E}\left|\left\langle \boldsymbol{w}_{t-\tau}, \boldsymbol{\delta}_t^{(0)}\right\rangle\right| < \infty$$

since $\boldsymbol{\beta}_1 \in \ell^1$ and $\mathbb{E}\left|\left\langle \boldsymbol{w}_{t-\tau}, \boldsymbol{\delta}_t^{(0)} \right\rangle\right| < \infty$ because

$$\mathbb{E}\left\langle \boldsymbol{w}_{t-\tau}, \boldsymbol{\delta}_t^{(0)} \right\rangle = \mathsf{Tr}\left[\mathbb{E}\left[\boldsymbol{w}_{t-\tau} \otimes \boldsymbol{\delta}_t^{(0)}\right]\right] = 0$$

by Proposition C.19. By Proposition C.19 again, we thus get

$$\mathbb{E}\left[\sum_{\tau=0}^{\infty} \beta_\tau \left(\boldsymbol{w}_{t-\tau} \otimes \delta_t^{(0)}\right)\right] = \sum_{\tau=0}^{\infty} \beta_\tau \, \mathbb{E}\left[\left(\boldsymbol{w}_{t-\tau} \otimes \delta_t^{(0)}\right)\right] = \mathbf{0}\,.$$

□

C.4 Asymptotics of ν-Noisy-FTRL

We now state and prove the upper bound for ν-Noisy-FTRL. Note that ν-Noisy-FTRL can be described in the frequency domain as $|\hat{B}^\nu(\omega)|^2 = |1 - \nu - \exp(i\omega)|$.

Proposition C.20. *Consider the setting of Theorem C.15 with $\sigma_{\mathsf{sgd}}^2 = 0$. Then, ν-Noisy-FTRL with $\nu \le \eta\mu$ satisfies*

$$F_\infty(\hat{\boldsymbol{\beta}}^\nu) \le C \, \max\{1, C_{\mathsf{kurt}}\} \, \eta^2 G^2 \rho^{-1} \, \mathsf{Tr}\left[\boldsymbol{H}\right] \, \log^2\left(\frac{8}{\nu}\right) + \widetilde{O}(\eta^3 R^2 \mu G^2 \rho^{-1})\,,$$

for a universal constant $C > 0$, and $\widetilde{O}(\cdot)$ suppresses polylogarithmic terms in the problem parameters.

Proof. We use C to denote a universal constant that can change from line to line. Denote $\mathcal{I}(a, b)$ for $0 < a, b < 1$ as the integral

$$\mathcal{I}(a, b) := \int_{-\pi}^{\pi} \frac{|1 - a - \exp(i\omega)|}{|1 - b - \exp(i\omega)|^2} \, \mathrm{d}\omega\,.$$

We can express the bound of Theorem C.15 with our specific choice of $B(\omega)$ as

$$F_\infty(\hat{B}^\nu) \le C \, \max\{1, C_{\mathsf{kurt}}\} \, \mathcal{I}(\nu, \nu) \sum_{j=1}^{d} \lambda_j \mathcal{I}(\nu, \eta\lambda_j)\,. \tag{41}$$

The strategy is to reduce each $\mathcal{I}$ term to standard elliptic integrals and leverage their well-studied properties to get the result. We start with the first term. We use Lemma F.15 to rewrite in terms of the elliptic integral of the first kind $K(k) = \int_0^{\pi/2} \mathrm{d}\omega / \sqrt{1 - k^2 \sin^2(\omega)}$ (denoted as (a)). Then, we use Property F.10 which says that $K(k) = O(-\log\sqrt{1 - k^2})$ (denoted as (b)). This gives,

$$\mathcal{I}(\nu, \nu) \overset{\text{(a)}}{=} \frac{4}{2 - \nu} K\left(\frac{\sqrt{1 - \nu}}{1 - \nu/2}\right) \overset{\text{(b)}}{\le} \frac{5}{2 - \nu} \log\left(\frac{4}{\nu}(2 - \nu)\right) \le 5 \log\left(\frac{8}{\nu}\right)\,. \tag{42}$$

Similarly, we can express the second integral in terms of the elliptic integral of the third kind $\Pi(\alpha^2, k)$, whose definition is given in (88). From Lemma F.16, we have for $a, b \in (0, 1)$ that

$$\mathcal{I}(a, b) = \frac{2a^2}{b^2(1 - a/2)} \Pi(\alpha^2, k) \quad \text{where} \quad \alpha^2 = \frac{b^2(1 - a) - a^2(1 - b)}{b^2(1 - a/2)^2}$$

and $k = \sqrt{1 - a}/(1 - a/2)$. We invoke Property F.11 to bound the behavior of $\Pi(\alpha^2, k)$ as $k \to 1^-$ (i.e. $a \to 0^+$) to get

$$\begin{aligned}
\mathcal{I}(a, b) &\le \frac{2a^2}{b^2(1 - a/2)} \frac{1}{\sqrt{1 - \alpha^2}} \log\frac{4}{\sqrt{1 - k^2}} \left(1 + O(a)\right) \\
&= \frac{2(1 - a/2)}{(1 - b/2)^2} \log\left(\frac{4}{a}(2 - a)\right) \left(1 + O(a)\right) \le \frac{128}{49} \log(8/a) \left(1 + O(a)\right)\,,
\end{aligned}$$

where the last inequality holds for $a \le b \le 1/4$. We plug in $a = \nu$ and $b = \eta\lambda_j$ so these conditions are satisfied to get

$$\mathcal{I}(\eta\mu, \eta\lambda_j) \le C \log\left(\frac{8}{\nu}\right) \left(1 + O(\nu)\right)\,. \tag{43}$$

The last term is $O(\nu) \le O(\eta\mu)$. Plugging in (42) and (43) into (41) and using $\mathsf{Tr}\left[\boldsymbol{H}\right] = \sum_{j=1}^{n} \lambda_j \le R^2$ completes the proof. □

C.5 Asymptotics of Anti-PGD

As we discussed in Table 3, anti-PGD [44] is a special case of Noisy-FTRL with $\boldsymbol{\beta} = (1, -1, 0, \ldots)$. Then, we have that $(\mathrm{Toeplitz}(\boldsymbol{\beta}))^{-1}$ is the lower triangular matrix of all ones, so we have $\gamma_T(\boldsymbol{\beta}) = T$, or that its limiting sensitivity is infinite.

We can circumvent the infinity by damping $\boldsymbol{\beta} = (1, -(1-\nu), 0, \ldots)$ for some $0 < \nu < 1$ to be decided later. In this case, we have $B(\omega) = 1 - (1-\nu)\exp(-i\omega)$, so that $|B(\omega)|^2 = |1 - \nu - \exp(i\omega)|^2$, which is the analogue of ν-Noisy-FTRL with a square.

Proposition C.21. *Consider the setting of Theorem C.15 with $\sigma^2_{\mathsf{sgd}} = 0$ and $\boldsymbol{\beta} = (1, -(1-\eta\lambda), 0, \ldots)$ for some $\lambda \in (0, 1/\eta]$. Then, we have,*

$$F_\infty(\boldsymbol{\beta}) = \Theta\left(\eta G^2 \rho^{-1}\left(\nu d + \frac{\eta \mathsf{Tr}\left[\boldsymbol{H}\right]}{\nu}\right)\right).$$

Further, if the learning rate satisfies $\eta = c/\mathsf{Tr}\left[\boldsymbol{H}\right]$ and we take $\boldsymbol{\beta} = (1, -(1-\sqrt{1/d}), \ldots)$, we get

$$F_\infty(\boldsymbol{\beta}) = \Theta\left((c^{1/2} + c^{-1/2})\eta^{3/2}\sigma^2\sqrt{d\,\mathsf{Tr}\left[\boldsymbol{H}\right]}\right).$$

Proof. Let $\sigma^2 = G^2/(2\rho)$. From Theorems C.15 and C.17, we get that

$$F_\infty(\boldsymbol{\beta}) = \Theta\left(\eta^2\sigma^2\left(\int_{-\pi}^{\pi}\frac{\mathrm{d}\omega}{|1-\nu-\exp(i\omega)|^2}\right)\left(\sum_{j=1}^{d}\lambda_j\int_{-\pi}^{\pi}\frac{|1-\nu-\exp(i\omega)|^2}{|1-\eta\lambda_j-\exp(i\omega)|^2}\,\mathrm{d}\omega\right)\right). \tag{44}$$

Using Lemma F.12, we have

$$\int_{-\pi}^{\pi}\frac{\mathrm{d}\omega}{|1-\nu-\exp(i\omega)|^2} = \frac{2\pi}{\nu(2-\nu)} = \Theta\left(\frac{1}{\nu}\right).$$

For the second integral, we expand out the numerator and invoke Lemma F.12 again to get

$$\begin{aligned}\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{|1-\nu-\exp(i\omega)|^2}{|1-\eta\lambda_j-\exp(i\omega)|^2}\,\mathrm{d}\omega &= \frac{1+(1-\nu)^2}{\eta\lambda_j(2-\eta\lambda_j)} - 2(1-\nu)\frac{1-\eta\lambda_j}{\eta\lambda_j(2-\eta\lambda_j)}\\ &= \Theta\left(\frac{\nu^2}{\eta\lambda_j}+1\right),\end{aligned}$$

where we use $1 \le 2-\nu \le 2$ and the same for λ_j instead of λ. Plugging the two integrals back into (44) completes the proof. □

C.6 Proofs of Technical Lemmas

We now prove Lemma C.4.

Proof of Lemma C.4. Denote

$$I = \int_{-\pi}^{\pi}\frac{|B(\omega)|^2\,\mathrm{d}\omega}{|1-\eta\lambda_j-\exp(i\omega)|^2}.$$

The denominator is simply

$$|1-\exp(i\omega)-\eta\lambda_j|^2 = 1 + (1-\eta\lambda_j)^2 - 2(1-\eta\lambda_j)\cos\omega. \tag{45}$$

We expand the numerator as

$$\begin{aligned}|B(\omega)|^2 &= \sum_{t=0}^{\infty}\beta_t^2 + \sum_{t=0}^{\infty}\sum_{\tau=0}^{t-1}\beta_t\beta_\tau\big(\exp(i\omega(t-\tau)) + \exp(-i\omega(\tau-t))\big)\\ &= \sum_{t=0}^{\infty}\beta_t^2 + 2\sum_{t=0}^{\infty}\sum_{\tau=0}^{t-1}\beta_t\beta_\tau\cos(\omega(t-\tau))\\ &= \sum_{t=0}^{\infty}\sum_{\tau=0}^{\infty}\beta_t\beta_\tau\cos(\omega(t-\tau)).\end{aligned} \tag{46}$$

This is bounded since the Cauchy-Schwarz inequality gives

$$|B(\omega)|^2 \le \|\boldsymbol{\beta}\|_2^2 < \infty\,.$$

Thus, we can apply Fubini's theorem to exchange the sum and integral to give

$$\begin{aligned} I &= \sum_{t=0}^{\infty}\sum_{\tau=0}^{\infty}\beta_t\beta_\tau \int_{-\pi}^{\pi} \frac{\cos(\omega(t-\tau))\mathrm{d}\omega}{1+(1-\eta\lambda_j)^2-2(1-\eta\lambda_j)\cos(\omega)} \\ &= \sum_{t=0}^{\infty}\sum_{\tau=0}^{\infty}\frac{2\pi}{1-(1-\eta\lambda_j)^2}(1-\eta\lambda_j)^{|t-\tau|} = \frac{2\pi\langle\boldsymbol{\beta},\boldsymbol{T}_j\boldsymbol{\beta}\rangle}{\eta\lambda_j(2-\eta\lambda_j)}\,, \end{aligned}$$

where we evaluated the integral using Lemma F.12. We use $1 \le 2-\eta\lambda_j \le 2$ to complete the proof. □

D Finite-Time Privacy-Utility Tradeoffs for Linear Regression

The goal of this section is to establish the finite time convergence of DP-FTRL. The key idea of the proof is to establish high probability bounds on the ℓ_2 norm of the iterates of Noisy-FTRL and use that to deduce a clip norm that does not clip any gradients with high probability.

Outline: The outline of the rest of this section is as follows:

- **Appendix D.1**: Preliminaries, including setup, notation and assumptions.
- **Appendix D.2**: High probability bounds the iterates of Noisy-FTRL.
- **Appendix D.3**: Expected bounds on the iterates of Noisy-FTRL.
- **Appendix D.4**: Connecting DP-FTRL to Noisy-FTRL for the final bound privacy-utility bounds (Corollary D.14 for DP-SGD and Corollary D.15 for DP-FTRL).

D.1 Setup, Assumptions and Notation

In this section, we make precise the assumptions, notation, and give some preliminary results.

D.1.1 Assumptions

We make the following assumptions throughout this section.

Assumption D.1. *The data distribution $\mathbb{P}_{\mathsf{data}}$ satisfies the following:*

*(**A1**) **Input Distribution**: The inputs have mean $\mathbb{E}[\boldsymbol{x}] = \boldsymbol{0}$ and covariance $\mathbb{E}[\boldsymbol{x}\otimes\boldsymbol{x}] =: \boldsymbol{H}$. We have $\mu\boldsymbol{I} \preceq \boldsymbol{H} \preceq L\boldsymbol{I}$ for $\mu, L > 0$. Further, $\boldsymbol{H}^{-1/2}\boldsymbol{x}$ is element-wise independent and sub-Gaussian with variance proxy 1, e.g. $\boldsymbol{H}^{-1/2}\boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{I})$.*

*(**A2**) **Noise Distribution**: There exists a $\boldsymbol{\theta}_\star \in \mathbb{R}^d$ such that $y = \langle\boldsymbol{\theta}_\star, \boldsymbol{x}\rangle + \xi$, where ξ is independent of $\boldsymbol{x}$ and is zero-mean sub-Gaussian with variance proxy σ_{sgd}^2, e.g. $\xi \sim \mathcal{N}(0, \sigma_{\mathsf{sgd}}^2)$.*

Not under Assumption (**A2**) that $\boldsymbol{\theta}_\star$ is the global minimizer of $F(\boldsymbol{\theta}) = (1/2)\mathbb{E}(\langle\boldsymbol{\theta},\boldsymbol{x}\rangle - y)^2$.

D.1.2 Notation

- As in Assumption C.2, we denote R^2 as the smallest number such that the fourth moment of $\boldsymbol{x}$ is bounded as

$$\mathbb{E}\left[\|\boldsymbol{x}\|_2^2\,\boldsymbol{x}\otimes\boldsymbol{x}\right] \preceq R^2\boldsymbol{H}\,. \tag{47}$$

  Under Assumption (**A1**), we have $R^2 = \Theta(\mathsf{Tr}\left[\boldsymbol{H}\right])$ always. While $\mathsf{Tr}\left[\boldsymbol{H}\right] \le R^2$ directly follows from (47) using Jensen's inequality, we show that $R^2 \le 3\mathsf{Tr}\left[\boldsymbol{H}\right]$ in Property C.3 in Appendix C.1.
- It is convenient to rewrite the Noisy-FTRL recursion (16) in terms of the difference $\boldsymbol{\theta}'_t := \boldsymbol{\theta}_t - \boldsymbol{\theta}_\star$ as

$$\boldsymbol{\theta}'_{t+1} = \big(\boldsymbol{I} - \eta(\boldsymbol{x}_t\otimes\boldsymbol{x}_t)\big)\boldsymbol{\theta}'_t + \eta\,\xi_t\boldsymbol{x}_t - \eta\sum_{\tau=0}^{t}\beta_\tau\boldsymbol{w}_{t-\tau}\,. \tag{48}$$

We will show in the upcoming Property D.2 that $\boldsymbol{\theta}'_t = \hat{\boldsymbol{\theta}}_t + \widetilde{\boldsymbol{\theta}}^{\mathsf{sgd}} + \widetilde{\boldsymbol{\theta}}^{\mathsf{dp}}$, where $\hat{\boldsymbol{\theta}}_t$ captures the effect of the initial iterate, $\widetilde{\boldsymbol{\theta}}^{\mathsf{sgd}}$ captures the effect of the SGD noise, and $\widetilde{\boldsymbol{\theta}}^{\mathsf{dp}}$ captures the effect of the additive DP noise. We will define these quantities now and state and prove Property D.2 later. Note that these recursions are defined for the same sequences of input realizations $(\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots)$ drawn from $\mathbb{P}_{\mathsf{data}}$, linear model noise realizations $(\xi_0, \xi_1, \ldots)$, and DP noise realizations $(\boldsymbol{w}_0, \boldsymbol{w}_1, \ldots)$.

- We define the noise-free version of the DP-FTRL recursion as $\hat{\boldsymbol{\theta}}_0 = \boldsymbol{\theta}'_0$ and

$$\hat{\boldsymbol{\theta}}_{t+1} = \big(\boldsymbol{I} - \eta(\boldsymbol{x}_t \otimes \boldsymbol{x}_t)\big)\hat{\boldsymbol{\theta}}_t \,. \tag{49}$$

- The effect of the SGD noise in the Noisy-FTRL process can be quantified by creating a process starting from $\widetilde{\boldsymbol{\theta}}_0^{\mathsf{sgd}} = \boldsymbol{0}$ with no DP noise (i.e. $\boldsymbol{w}_\tau \equiv \boldsymbol{0}$):

$$\widetilde{\boldsymbol{\theta}}_{t+1}^{\mathsf{sgd}} = \big(\boldsymbol{I} - \eta(\boldsymbol{x}_t \otimes \boldsymbol{x}_t)\big)\widetilde{\boldsymbol{\theta}}_t^{\mathsf{sgd}} + \eta\,\xi_t \boldsymbol{x}_t \,. \tag{50}$$

- The effect of the DP noise in the Noisy-FTRL process can be quantified by creating a process starting from $\widetilde{\boldsymbol{\theta}}_0^{\mathsf{dp}} = \boldsymbol{0}$ with no SGD noise (i.e., $\xi_t \equiv 0$):

$$\widetilde{\boldsymbol{\theta}}_{t+1}^{\mathsf{dp}} = \big(\boldsymbol{I} - \eta(\boldsymbol{x}_t \otimes \boldsymbol{x}_t)\big)\widetilde{\boldsymbol{\theta}}_t^{\mathsf{dp}} - \eta \sum_{\tau=0}^{t} \beta_\tau \boldsymbol{w}_{t-\tau} \,. \tag{51}$$

- For an input $\boldsymbol{x}_t$ drawn from $\mathbb{P}_{\mathsf{data}}$ We define the matrix

$$\boldsymbol{Q}_t := \boldsymbol{I} - \eta \boldsymbol{x}_t \otimes \boldsymbol{x}_t \,. \tag{52}$$

  Note that $\mathbb{E}[\boldsymbol{Q}_t] = \boldsymbol{I} - \eta \boldsymbol{H}$.
- Define the linear operator $\mathcal{P} : \mathbb{S}_+^d \to \mathbb{S}_+^d$ that operates on the cone of PSD matrices given by

$$\mathcal{P}\boldsymbol{M} = \mathbb{E}[(\boldsymbol{I} - \eta \boldsymbol{x} \otimes \boldsymbol{x})\boldsymbol{M}(\boldsymbol{I} - \eta \boldsymbol{x} \otimes \boldsymbol{x})] \,, \tag{53}$$

  where $\boldsymbol{x}$ is an input drawn from $\mathbb{P}_{\mathsf{data}}$. By definition, we have $\mathbb{E}[\boldsymbol{Q}_t \boldsymbol{M} \boldsymbol{Q}_t] = \mathcal{P}\boldsymbol{M}$ and by independence,

$$\mathbb{E}[\boldsymbol{Q}_t \boldsymbol{Q}_{t-1} \boldsymbol{M} \boldsymbol{Q}_{t-1} \boldsymbol{Q}_t] = \mathcal{P}(\mathcal{P}\boldsymbol{M}) = \mathcal{P}^2 \boldsymbol{M} \,. \tag{54}$$

  This extends to higher powers of $\mathcal{P}$ as well. Finally, we will heavily use the fact that $\mathsf{Tr}\left[\mathcal{P}\boldsymbol{M}\right] \le (1 - \eta\mu)\mathsf{Tr}\left[\boldsymbol{M}\right]$ for PSD matrices $\boldsymbol{M}$ (see Lemma F.18 for a proof).
- For each iteration t, we define the PSD matrix $\boldsymbol{\Sigma}_t^{\mathsf{sgd}}$ as

$$\boldsymbol{\Sigma}_t^{\mathsf{sgd}} = \boldsymbol{x}_{t-1} \otimes \boldsymbol{x}_{t-1} + \boldsymbol{Q}_{t-1}(\boldsymbol{x}_{t-2} \otimes \boldsymbol{x}_{t-2})\boldsymbol{Q}_{t-1} + \cdots + \boldsymbol{Q}_{t-1} \cdots \boldsymbol{Q}_1 (\boldsymbol{x}_0 \otimes \boldsymbol{x}_0) \boldsymbol{Q}_1 \cdots \boldsymbol{Q}_{t-1} \,, \tag{55}$$

- For each iteration t, we define the PSD matrix $\boldsymbol{\Sigma}_t^{\mathsf{dp}}$ as

$$\begin{aligned} \boldsymbol{\Sigma}_t^{\mathsf{dp}} &= \sum_{\tau=0}^{t-1} \boldsymbol{V}_{t,\tau} \boldsymbol{V}_{t,\tau}^\top \quad \text{where} \\ \boldsymbol{V}_{t,\tau} &= \begin{cases} \beta_\tau \boldsymbol{I} + \beta_{\tau-1} \boldsymbol{Q}_{t-1} + \cdots + \beta_0 \boldsymbol{Q}_{t-1} \cdots \boldsymbol{Q}_{t-\tau} \,, & \text{if } 1 \le \tau \le t-1 \,, \\ \beta_0 \boldsymbol{I} \,, & \text{if } \tau = 0 \,. \end{cases} \end{aligned} \tag{56}$$

D.1.3 Preliminary Results

The first result is a decomposition of the Noisy-FTRL process in to three processes: (a) gradient descent without additive noise, (b) a noise process with only noise from the linear model, and (c) a noise process with only the DP noise.

Property D.2. *For the sequences $\boldsymbol{\theta}'_t, \hat{\boldsymbol{\theta}}_t, \widetilde{\boldsymbol{\theta}}^{\mathsf{sgd}}_t, \widetilde{\boldsymbol{\theta}}^{\mathsf{dp}}_t$ defined in Equations* (48) *to* (51)*, we have the following:*

$$\boldsymbol{\theta}'_t = \hat{\boldsymbol{\theta}}_t + \widetilde{\boldsymbol{\theta}}^{\mathsf{sgd}}_t + \widetilde{\boldsymbol{\theta}}^{\mathsf{dp}}_t \tag{57}$$

$$\hat{\boldsymbol{\theta}}_t = \ \boldsymbol{Q}_t \cdots \boldsymbol{Q}_0 \boldsymbol{\theta}'_0 \tag{58}$$

$$\widetilde{\boldsymbol{\theta}}^{\mathsf{sgd}}_t = \eta \left(\boldsymbol{x}_t \xi_t + \boldsymbol{Q}_t \boldsymbol{x}_{t-1} \xi_{t-1} + \cdots + \boldsymbol{Q}_t \cdots \boldsymbol{Q}_1 \boldsymbol{x}_0 \xi_0 \right) \tag{59}$$

$$\begin{aligned} \widetilde{\boldsymbol{\theta}}^{\mathsf{dp}}_t &= -\eta \left(\sum_{\tau=0}^{t} \beta_\tau \boldsymbol{w}_{t-\tau} + \boldsymbol{Q}_t \sum_{\tau=0}^{t-1} \beta_\tau \boldsymbol{w}_{t-1-\tau} + \cdots + \boldsymbol{Q}_t \cdots \boldsymbol{Q}_1 (\beta_0 \boldsymbol{w}_0) \right) \\ &= -\eta \Big(\beta_0 \boldsymbol{w}_{t-1} + (\beta_1 \boldsymbol{I} + \beta_0 \boldsymbol{Q}_{t-1}) \boldsymbol{w}_{t-2} + \cdots + (\beta_{t-1} \boldsymbol{I} + \beta_{t-2} \boldsymbol{Q}_{t-1} + \cdots + \beta_0 \boldsymbol{Q}_{t-1} \cdots \boldsymbol{Q}_1) \boldsymbol{w}_0 \Big) . \end{aligned} \tag{60}$$

Proof. The expressions follow from unrolling their respective updates. By unrolling the DP-FTRL update (48), we get,

$$\begin{aligned} \boldsymbol{\theta}'_{t+1} &= \boldsymbol{Q}_t \boldsymbol{\theta}'_t + \eta \boldsymbol{x}_t \xi_t - \eta \sum_{\tau=0}^{t} \beta_\tau \boldsymbol{w}_{t-\tau} \\ &= \boldsymbol{Q}_t \boldsymbol{Q}_{t-1} \boldsymbol{\theta}'_{t-1} + \eta \left(\boldsymbol{x}_t \xi_t + \boldsymbol{Q}_t \boldsymbol{x}_{t-1} \xi_{t-1} \right) - \eta \left(\sum_{\tau=0}^{t} \beta_\tau \boldsymbol{w}_{t-\tau} + \boldsymbol{Q}_t \sum_{\tau=0}^{t-1} \beta_\tau \boldsymbol{w}_{t-1-\tau} \right) \\ &= \boldsymbol{Q}_t \cdots \boldsymbol{Q}_0 \boldsymbol{\theta}'_0 + \eta \left(\boldsymbol{x}_t \xi_t + \boldsymbol{Q}_t \boldsymbol{x}_{t-1} \xi_{t-1} + \cdots + \boldsymbol{Q}_t \cdots \boldsymbol{Q}_1 \boldsymbol{x}_0 \xi_0 \right) \\ &\quad - \eta \left(\sum_{\tau=0}^{t} \beta_\tau \boldsymbol{w}_{t-\tau} + \boldsymbol{Q}_t \sum_{\tau=0}^{t-1} \beta_\tau \boldsymbol{w}_{t-1-\tau} + \cdots + \boldsymbol{Q}_t \cdots \boldsymbol{Q}_1 (\beta_0 \boldsymbol{w}_0) \right) . \end{aligned}$$

Unrolling Equations (49) to (51) respectively gives Equations (58) to (60), and comparing them with the expression above gives Equation (57). □

D.2 High-Probability Bounds on Noisy-FTRL

The goal of this subsection is to prove a high probability bound on norms of the iterates of Noisy-FTRL. We require a technical convergence condition on the weights $\boldsymbol{\beta}$.

Definition D.3. *A sequence $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots)$ is said to satisfy Half-Expo Decay with parameter $\nu \in (0, 1)$ if for all nonnegative integers τ, we have*

$$|\beta_0|(1-\nu)^{\tau/2} + |\beta_1|(1-\nu)^{(\tau-1)/2} + \cdots + |\beta_\tau| \leq C(1-\nu)^{\tau/2} \tag{61}$$

for a universal constant $C > 0$.

Theorem D.4. *Fix a constant $0 < p < 1$ and suppose the Assumption D.1 holds. Consider the sequence $(\boldsymbol{\theta}_t)_{t=0}^{T-1}$ of iterates and the sequence $(\boldsymbol{g}_t)_{t=0}^{T-1}$ of gradients when running Noisy-FTRL for T iterations with noise coefficients $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{T-1})$, DP noise $\boldsymbol{w}_t \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$ of a given variance[4] σ^2, a learning rate $\eta \leq \big(cR^2 \log(T/p)\big)$ for a universal constant $c \geq 1$. Further, suppose that $\boldsymbol{\beta}$ satisfies Half-Expo Decay with parameter ν for some $\nu \leq \eta\mu$. Then, with probability at least $1 - p$, we have*

$$\|\boldsymbol{\theta}'_t\|_2^2 \leq C \left(\|\boldsymbol{\theta}'_0\|_2^2 + \frac{\eta R^2 \sigma_{\mathsf{sgd}}^2}{\mu} + \frac{\eta^2 \sigma^2 d \, \|\boldsymbol{\beta}\|_1^2}{\nu} \right) \log^3 \left(\frac{T}{p} \right) \quad \textit{and}$$

$$\|\boldsymbol{g}_t\|_2^2 \leq C R^4 \left(\|\boldsymbol{\theta}'_0\|_2^2 + \frac{\eta R^2 \sigma_{\mathsf{sgd}}^2}{\mu} + \frac{\sigma_{\mathsf{sgd}}^2}{R^2} + \frac{\eta^2 \sigma^2 d \|\boldsymbol{\beta}\|_1^2}{\nu} \right) \log^5 \left(\frac{T}{p} \right) .$$

for a universal constant C.

We prove this theorem over a sequence of intermediate results.

[4] In the context of this paper, we have $\sigma^2 = G^2 \gamma(\boldsymbol{\beta})^2 / (2\rho)$.

D.2.1 Proof Setup: Definition of Events

The proof strategy relies on defining some events (that hold with high probability from concentration of measure) and prove the required boundedness under those events. Consider $0 < p < 1$ and a universal constant C from statement of Theorem D.4. We define the following events.

- Define the event where the inputs are bounded in norm as:

$$\mathcal{E}_1 := \bigcap_{t=0}^{T-1} \left\{ \|\boldsymbol{x}_t\|_2^2 \le CR^2 \log\left(\frac{T}{p}\right) \right\} . \tag{62}$$

- Define an event where the noise in the linear model is bounded as:

$$\mathcal{E}_2 := \bigcap_{t=0}^{T-1} \left\{ |\xi_t|^2 \le 2\sigma_{\mathsf{sgd}}^2 \log\left(\frac{2T}{p}\right) \right\} . \tag{63}$$

- Define the event where the norm of $\widetilde{\boldsymbol{\theta}}^{\mathsf{sgd}}$ defined in (50) is bounded

$$\mathcal{E}_1^{\mathsf{dp}} := \bigcap_{t=0}^{T-1} \left\{ \left\| \widetilde{\boldsymbol{\theta}}^{\mathsf{sgd}} \right\|_2^2 \le C\eta^2\sigma_{\mathsf{sgd}}^2 \, \mathsf{Tr}\left[\boldsymbol{\Sigma}_t^{\mathsf{sgd}}\right] \log\left(\frac{T}{p}\right) \right\} , \tag{64}$$

  where we define the random matrix $\boldsymbol{\Sigma}_t^{\mathsf{sgd}} = \boldsymbol{x}_{t-1} \otimes \boldsymbol{x}_{t-1} + \boldsymbol{Q}_{t-1}(\boldsymbol{x}_{t-2} \otimes \boldsymbol{x}_{t-2})\boldsymbol{Q}_{t-1} + \cdots + \boldsymbol{Q}_{t-1} \cdots \boldsymbol{Q}_1(\boldsymbol{x}_0 \otimes \boldsymbol{x}_0)\boldsymbol{Q}_1 \cdots \boldsymbol{Q}_{t-1}$ (see also (55)). When this event holds, we have that $\boldsymbol{0} \preceq \boldsymbol{Q}_t \preceq \boldsymbol{I}$ for $t = 0, \ldots, T-1$ as long as $\eta \le 1/\left(CR^2 \log(T/p)\right)$. Indeed, in this case, we have

$$\boldsymbol{I} - \eta \boldsymbol{x}_t \otimes \boldsymbol{x}_t \succeq \left(1 - \eta \|\boldsymbol{x}_t\|_2^2\right) \boldsymbol{I} \succeq \boldsymbol{0} . \tag{65}$$

- The components of the sum defining $\boldsymbol{\Sigma}_t^{\mathsf{sgd}}$ are the PSD matrices $\boldsymbol{W}_{t,\tau}$, defined for $\tau \le t-1$ as

$$\boldsymbol{W}_{t,\tau} = \begin{cases} \boldsymbol{Q}_{t-1} \cdots Q_{\tau+1}(\boldsymbol{x}_\tau \otimes \boldsymbol{x}_\tau)\boldsymbol{Q}_{\tau+1} \cdots \boldsymbol{Q}_{t-1} , & \text{if } \tau < t-1, \\ \boldsymbol{x}_{t-1} \otimes \boldsymbol{x}_{t-1} , & \text{if } \tau = t-1 . \end{cases} \tag{66}$$

  Define the event where these are bounded in trace as

$$\mathcal{E}_2^{\mathsf{sgd}} := \bigcap_{t=0}^{T-1} \bigcap_{\tau=0}^{t-1} \left\{ \mathsf{Tr}\left[\boldsymbol{W}_{t,\tau}\right] \le \frac{T^2R^2}{p}(1-\eta\mu)^{t-1-\tau} \right\} . \tag{67}$$

- Define the event where the norm of $\widetilde{\boldsymbol{\theta}}^{\mathsf{dp}}$ defined in (51) is bounded as

$$\mathcal{E}_1^{\mathsf{dp}} := \bigcap_{t=0}^{T-1} \left\{ \left\| \widetilde{\boldsymbol{\theta}}_t^{\mathsf{dp}} \right\|_2^2 \le C\eta^2\sigma^2 \, \mathsf{Tr}\left[\boldsymbol{\Sigma}_t^{\mathsf{dp}}\right] \log\left(\frac{T}{p}\right) \right\} , \tag{68}$$

  where $\boldsymbol{\Sigma}_t^{\mathsf{dp}}$ is defined in (56).
- Define the event where the matrix $\boldsymbol{V}_{t,\tau}$ defined in (56) is bounded in trace:

$$\mathcal{E}_2^{\mathsf{dp}} := \bigcap_{t=0}^{T-1} \bigcap_{\tau=0}^{t-1} \left\{ \mathsf{Tr}\left[\boldsymbol{V}_{t,\tau}\boldsymbol{V}_{t,\tau}^\top\right] \le \frac{T^2 d}{p} \left(\sum_{k=0}^{\tau} |\beta_k| (1-\eta\mu)^{(\tau-k)/2} \right) \right\} . \tag{69}$$

We show that all these events hold with high probability.

Proposition D.5. *Consider the setting of Theorem D.4. We have,*

$$\mathbb{P}\left(\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_1^{\mathsf{sgd}} \cap \mathcal{E}_2^{\mathsf{sgd}} \cap \mathcal{E}_1^{\mathsf{dp}} \cap \mathcal{E}_2^{\mathsf{dp}}\right)) \ge 1 - 6p .$$

Proof. We will show that each of the events holds with probability at least $1 - p$ and a union bound gives the desired result.

Event $\mathcal{E}_1$: Since $\boldsymbol{z}_t = \boldsymbol{H}^{-1/2}\boldsymbol{x}_t$ is element-wise independent and 1-sub-Gaussian, we have from the Hanson-Wright inequality (Lemma F.6) that

$$\mathbb{P}(\|\boldsymbol{x}_t\|_2^2 > C\mathsf{Tr}\left[\boldsymbol{H}\right]\log(1/p)) = \mathbb{P}(\langle \boldsymbol{z}_t, \boldsymbol{H}\boldsymbol{z}_t\rangle > C\mathsf{Tr}\left[\boldsymbol{H}\right]\log(1/p)) \le p\,.$$

Taking a union bound over $t = 0, 1, \ldots, T-1$ gives that $\mathbb{P}(\mathcal{E}_1) \ge 1 - p$.

Event $\mathcal{E}_2$: Since ξ_t is sub-Gaussian with mean zero and variance proxy σ_{sgd}^2, we have,

$$\mathbb{P}(|\xi_t| > s) \le 2\exp\left(-\frac{s^2}{2\sigma_{\mathsf{sgd}}^2}\right).$$

Setting the right side equal to p/T and taking a union bound over $t = 0, 1, \ldots, T-1$ gives $\mathbb{P}(\mathcal{E}_2) \ge 1 - p$.

Event $\mathcal{E}_1^{\mathsf{sgd}}$: From the expression for $\widetilde{\boldsymbol{\theta}}_t^{\mathsf{sgd}}$ from (59), we can say that $\widetilde{\boldsymbol{\theta}}_t^{\mathsf{sgd}}$ conditioned on $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{t-1}$ is mean zero and satisfies

$$\widetilde{\boldsymbol{\theta}}_t^{\mathsf{sgd}} = \eta \underbrace{\begin{bmatrix}\boldsymbol{x}_{t-1} & \boldsymbol{Q}_{t-1}\boldsymbol{x}_{t-1} & \cdots & (\boldsymbol{Q}_{t-1}\cdots\boldsymbol{Q}_1\boldsymbol{x}_0)\end{bmatrix}}_{=:\boldsymbol{M}_t} \begin{bmatrix}\xi_{t-1}\\ \vdots \\ \xi_0\end{bmatrix}.$$

Using the assumption that each ξ_τ is independent and sub-Gaussian with variance proxy σ_{sgd}^2, we get from the Hanson-Wright inequality (Lemma F.6) again that

$$\mathbb{P}\left(\left\|\widetilde{\boldsymbol{\theta}}_t^{\mathsf{sgd}}\right\|_2^2 > C\eta^2\sigma_{\mathsf{sgd}}^2\,\mathsf{Tr}\left[\boldsymbol{M}_t\boldsymbol{M}_t^\top\right]\log(1/p)\right) = \mathbb{P}\left(\left\langle \boldsymbol{\xi}_{:t}, \boldsymbol{M}_t\boldsymbol{M}_t^\top\boldsymbol{\xi}_{:t}\right\rangle > C\eta^2\sigma_{\mathsf{sgd}}^2\,\mathsf{Tr}\left[\boldsymbol{M}_t\boldsymbol{M}_t^\top\right]\log(1/p)\right) \le p\,.$$

Next, we confirm that

$$\mathsf{Tr}\left[\boldsymbol{M}_t\boldsymbol{M}_t^\top\right] = \|\boldsymbol{x}_{t-1}\|_2^2 + \|\boldsymbol{Q}_{t-1}\boldsymbol{x}_{t-1}\|_2^2 + \cdots + \|\boldsymbol{Q}_{t-1}\cdots\boldsymbol{Q}_1\boldsymbol{x}_0\|_2^2 = \mathsf{Tr}\left[\boldsymbol{\Sigma}_t^{\mathsf{sgd}}\right].$$

Finally, a union bound over $t = 0, 1, \ldots, T-1$ gives that $\mathbb{P}(\mathcal{E}_1^{\mathsf{sgd}}) \ge 1 - p$.

Event $\mathcal{E}_2^{\mathsf{sgd}}$: Markov's inequality gives

$$\mathbb{P}\left(\mathsf{Tr}\left[\boldsymbol{W}_{t,\tau}\right] > s\right) \le \frac{1}{s}\mathbb{E}\left[\boldsymbol{W}_{t,\tau}\right] \le (1-\eta\mu)^{t-1-\tau}\frac{R^2}{s}$$

where the calculations for the expected bound are deferred to Lemma D.9. Taking a union bound over all $T(T+1)/2 \le T^2$ choices of (t, τ) gives $\mathbb{P}(\mathcal{E}_2^{\mathsf{sgd}}) \ge 1 - p$.

Event $\mathcal{E}_1^{\mathsf{dp}}$: From the expression for $\widetilde{\boldsymbol{\theta}}_t^{\mathsf{dp}}$ from (60), we deduce that

$$\widetilde{\boldsymbol{\theta}}_t^{\mathsf{dp}} \,|\, \boldsymbol{x}_0, \ldots, \boldsymbol{x}_{t-1} \;\sim\; \mathcal{N}(\boldsymbol{0}, \eta^2\sigma^2\boldsymbol{\Sigma}_t^{\mathsf{dp}})\,.$$

Invoking the Hanson-Wright inequality (Lemma F.6) and union bounding over $t = 0, \ldots, T-1$ gives $\mathbb{P}(\mathcal{E}_1^{\mathsf{dp}}) \ge 1 - p$.

Event $\mathcal{E}_2^{\mathsf{dp}}$: Markov's inequality gives

$$\mathbb{P}\left(\mathsf{Tr}\left[\boldsymbol{V}_{t,\tau}\boldsymbol{V}_{t,\tau}^\top\right] > s\right) \le \frac{1}{s}\mathbb{E}\left[\boldsymbol{V}_{t,\tau}\boldsymbol{V}_{t,\tau}^\top\right] \le \left(\sum_{k=0}^{\tau}|\beta_k|(1-\eta\mu)^{(\tau-k)/2}\right)\frac{d}{s}$$

where we defer the technical calculations involved in bounding the expectation above to Lemma D.10. Taking a union bound over all $T(T+1)/2 \le T^2$ choices of (t, τ) gives $\mathbb{P}(\mathcal{E}_2^{\mathsf{dp}}) \ge 1 - p$. □

D.2.2 High Probability Bounds on Component Recursions

Bound on the noise-less iterates: We start with $\hat{\boldsymbol{\theta}}_t$ from (49).

Proposition D.6. *Under event $\mathcal{E}_1$ and if $\eta \leq (CR^2 \log(T/p))^{-1}$, we have that $\left\|\hat{\boldsymbol{\theta}}_t\right\|_2 \leq \|\boldsymbol{\theta}'_0\|_2$.*

Proof. Using the fact that $\mathbf{0} \preceq \boldsymbol{Q}_t \preceq \boldsymbol{I}$ under $\mathcal{E}_1$ (cf. Equation (65)), we get

$$\left\|\hat{\boldsymbol{\theta}}_t\right\|_2 = \|\boldsymbol{Q}_{t-1} \cdots \boldsymbol{Q}_0 \boldsymbol{\theta}'_0\|_2 \leq \|\boldsymbol{Q}_{t-1}\|_2 \cdots \|\boldsymbol{Q}_0\|_2 \|\boldsymbol{\theta}'_0\|_2 \leq \|\boldsymbol{\theta}'_0\|_2 .$$

□

Bound on $\widetilde{\boldsymbol{\theta}}_t^{\mathsf{sgd}}$: We turn to $\widetilde{\boldsymbol{\theta}}_t^{\mathsf{sgd}}$ from (50).

Proposition D.7. *Under events $\mathcal{E}_1, \mathcal{E}_1^{\mathsf{sgd}}, \mathcal{E}_2^{\mathsf{sgd}}$, and $\eta \leq (CR^2 \log(T/p))^{-1}$, we have*

$$\left\|\widetilde{\boldsymbol{\theta}}_t^{\mathsf{sgd}}\right\|_2^2 \leq C \left(\frac{\eta R^2}{\mu}\right) \log^3\left(\frac{T}{p}\right) .$$

Proof. Under $\mathcal{E}_1^{\mathsf{sgd}}$, we have

$$\left\|\widetilde{\boldsymbol{\theta}}^{\mathsf{sgd}}\right\|_2^2 \leq C\eta^2 \sigma_{\mathsf{sgd}}^2 \, \mathsf{Tr}\left[\boldsymbol{\Sigma}_t^{\mathsf{sgd}}\right] \log\left(\frac{T}{p}\right) . \tag{70}$$

We bound $\mathsf{Tr}\left[\boldsymbol{\Sigma}_t\right] = \sum_{\tau=0}^{t-1} \mathsf{Tr}\left[\boldsymbol{W}_{t,\tau}\right]$ for $\boldsymbol{W}_{t,\tau}$ defined in (66). We have two bounds for $\mathsf{Tr}\left[\boldsymbol{W}_{t,\tau}\right]$:

(a) Using $\mathbf{0} \preceq \boldsymbol{Q}_t \preceq \boldsymbol{I}$ under $\mathcal{E}_1$ (cf. Equation (65)), we bound

$$\mathsf{Tr}\left[\boldsymbol{W}_{t,\tau}\right] = \|\boldsymbol{Q}_{t-1} \cdots \boldsymbol{Q}_{\tau+1} \boldsymbol{x}_\tau\|_2^2 \leq \|\boldsymbol{Q}_{t-1}\|_2^2 \cdots \|\boldsymbol{Q}_{\tau+1}\|_2^2 \|\boldsymbol{x}_\tau\|_2^2 \leq CR^2 \log(T/p) .$$

(b) Under event $\mathcal{E}_2^{\mathsf{sgd}}$, we have the bound

$$\mathsf{Tr}\left[\boldsymbol{W}_{t,\tau}\right] \leq \frac{T^2 R^2}{p} (1 - \eta\mu)^{t-1-\tau} .$$

Using the first bound for the last $\tau \leq t-1$ iterations and the second bound for the rest, we get

$$\begin{aligned}
\mathsf{Tr}\left[\boldsymbol{\Sigma}_t^{\mathsf{sgd}}\right] &\leq \sum_{k=0}^{t-\tau-1} \frac{T^2 R^2}{p} (1-\eta\mu)^{t-1-\tau} \mathbb{1}\left(\tau < t-1\right) + \tau \left(CR^2 \log(T/p)\right) \\
&\leq \frac{T^2 R^2}{p} (1-\eta\mu)^{\tau} \sum_{k=0}^{t-\tau-1} (1-\eta\mu)^{k} \mathbb{1}\left(\tau < t-1\right) + \tau \left(CR^2 \log(T/p)\right) \\
&\leq \frac{T^2 R^2}{p} \frac{\exp(-\eta\mu\tau)}{\eta\mu} \mathbb{1}\left(\tau < t-1\right) + \tau \left(CR^2 \log(T/p)\right) .
\end{aligned}$$

Choosing $\tau = \min\left\{t-1, \frac{1}{\eta\mu} \log\left(\frac{T^2}{Cp \log(T/p)}\right)\right\}$ as per Lemma F.20 gives

$$\mathsf{Tr}\left[\boldsymbol{\Sigma}_t^{\mathsf{sgd}}\right] \leq \frac{CR^2 \log(T/p)}{\eta\mu} \left(1 + \log\left(\frac{T^2}{p \log(T/p)}\right)\right) \leq \frac{C' R^2}{\eta\mu} \log^2(T/p)$$

for some absolute constants C, C'. Plugging this back into (70) completes the proof. □

Bound on $\widetilde{\boldsymbol{\theta}}_t^{\mathsf{dp}}$: We turn to $\widetilde{\boldsymbol{\theta}}_t^{\mathsf{dp}}$ from (51).

Proposition D.8. *Consider the setting of Theorem D.4. Under events $\mathcal{E}_1, \mathcal{E}_1^{\mathsf{dp}}, \mathcal{E}_2^{\mathsf{dp}}$, and $\eta \leq (CR^2 \log(T/p))^{-1}$, we have*

$$\left\|\widetilde{\boldsymbol{\theta}}_t^{\mathsf{sgd}}\right\|_2^2 \leq C \left(\frac{\eta R^2}{\mu}\right) \log^3\left(\frac{T}{p}\right) .$$

Proof. Based on the bound on $\left\|\widetilde{\boldsymbol{\theta}}_t^{\mathsf{dp}}\right\|_2$ from $\mathcal{E}_1^{\mathsf{dp}}$, we bound $\mathsf{Tr}\left[\boldsymbol{\Sigma}_t^{\mathsf{dp}}\right] = \sum_{\tau=0}^{t-1} \mathsf{Tr}\left[\boldsymbol{V}_{t,\tau}\boldsymbol{V}_{t,\tau}^\top\right]$. We bound each trace on the right side in two ways:

(a) We have $\mathsf{Tr}\left[\boldsymbol{V}_{t,\tau}\boldsymbol{V}_{t,\tau}^\top\right] \le \|\boldsymbol{\beta}\|_1^2 d$ from Lemma D.10.

(b) Under $\mathcal{E}_2^{\mathsf{dp}}$ and the assumption $(*)$ of Half-Expo Decay of β with parameter $\nu \le \eta\mu$, we also have

$$\begin{aligned}\mathsf{Tr}\left[\boldsymbol{V}_{t,\tau}\boldsymbol{V}_{t,\tau}^\top\right] &\le \frac{T^2 d}{p}\left(\sum_{\tau=0}^{\tau}|\beta_k|(1-\eta\mu)^{(\tau-k)/2}\right)^2 \\ &\le \frac{T^2 d}{p}\left(\sum_{\tau=0}^{\tau}|\beta_k|(1-\nu)^{(\tau-k)/2}\right)^2 \\ &\overset{(*)}{\le} \frac{CT^2 d}{p}(1-\nu)^\tau\,.\end{aligned}$$

Using the first bound for the first τ iterations and the second bound for the rest, we get

$$\begin{aligned}\mathsf{Tr}\left[\boldsymbol{\Sigma}_t^{\mathsf{dp}}\right] &\le \tau\left(\|\boldsymbol{\beta}\|_1^2 d\right) + \sum_{k=\tau}^{t-1}\frac{CT^2 d}{p}(1-\nu)^k \mathbb{1}\left(\tau > t-1\right) \\ &\le \tau\left(\|\boldsymbol{\beta}\|_1^2 d\right) + \frac{CT^2 d}{p}(1-\nu)^\tau \sum_{k=0}^{\infty}(1-\nu)^k \mathbb{1}\left(\tau > t-1\right) \\ &\le \tau\left(\|\boldsymbol{\beta}\|_1^2 d\right) + \frac{CT^2 d\exp(-\nu\tau)}{p\nu}\mathbb{1}\left(\tau > t-1\right)\,.\end{aligned}$$

Choosing $\tau \le \left\{t-1, \frac{1}{\nu}\log(CT^2/p\|\boldsymbol{\beta}\|_1^2)\right\}$ as per Lemma F.20, we get,

$$\mathsf{Tr}\left[\boldsymbol{\Sigma}_t^{\mathsf{dp}}\right] \le \frac{\|\boldsymbol{\beta}\|_1^2 d}{\nu}\left(1+\log\left(\frac{CT^2}{p\|\boldsymbol{\beta}\|_1^2}\right)\right) \le C'\frac{\|\boldsymbol{\beta}\|_1^2 d}{\nu}\log\left(\frac{T}{p}\right)\,,$$

where we used $\|\boldsymbol{\beta}\|_1 \ge |\beta_0| = 1$ and C, C' are some universal constants. Combining this with the bound on $\left\|\widetilde{\boldsymbol{\theta}}_t^{\mathsf{dp}}\right\|_2$ asserted by $\mathcal{E}_1^{\mathsf{dp}}$ completes the proof. □

D.2.3 Completing the Proof of the High Probability Bounds

We are now ready to prove Theorem D.4.

Proof of Theorem D.4. Under events $\mathcal{E}_1, \mathcal{E}_1^{\mathsf{sgd}}, \mathcal{E}_2^{\mathsf{sgd}}, \mathcal{E}_1^{\mathsf{dp}}, \mathcal{E}_2^{\mathsf{dp}}$, we have bounds on the norms of $\hat{\boldsymbol{\theta}}_t, \widetilde{\boldsymbol{\theta}}_t^{\mathsf{sgd}}, \widetilde{\boldsymbol{\theta}}_t^{\mathsf{dp}}$ respectively from Propositions D.6 to D.8. We combine them together with the triangle inequality and Equation (57) of Property D.2 to the claimed bound on $\|\boldsymbol{\theta}_t'\|_2$.

Next, for the gradients, we use the triangle and Cauchy-Schwarz inequalities on the definition $\boldsymbol{g}_t = \boldsymbol{x}_t\langle\boldsymbol{x}_t, \boldsymbol{\theta}_t'\rangle - \boldsymbol{x}_t\xi_t$ to get

$$\|\boldsymbol{g}_t\|_2^2 \le 2\|\boldsymbol{x}_t\|_2^4\|\boldsymbol{\theta}_t'\|_2^2 + 2\|\boldsymbol{x}_t\|_2^2|\xi_t|_2^2\,.$$

Plugging in the bounds on $\|\boldsymbol{x}_t\|_2$ and $|\xi|_t$ from $\mathcal{E}_1$ and $\mathcal{E}_2$ respectively gives the claimed bound on $\|\boldsymbol{g}_t\|_2^2$.

Finally, all the events above hold with probability at least $1-6p$ from Proposition D.5. Substituting $p/6$ for p and adjusting the constants completes the proof. □

D.2.4 Helper Lemmas

Lemma D.9. *Consider the setting of Theorem D.4 and consider the PSD matrices $\boldsymbol{W}_{t,\tau}$, defined for $\tau \le t-1$ as*

$$\boldsymbol{W}_{t,\tau} = \begin{cases}\boldsymbol{Q}_{t-1}\cdots Q_{\tau+1}(\boldsymbol{x}_\tau\otimes\boldsymbol{x}_\tau)\boldsymbol{Q}_{\tau+1}\cdots\boldsymbol{Q}_{t-1}\,, & \text{if } \tau < t-1,\\ \boldsymbol{x}_{t-1}\otimes\boldsymbol{x}_{t-1}\,, & \text{if } \tau = t-1\,.\end{cases}$$

We have that $\mathbb{E}[\mathsf{Tr}\left[\boldsymbol{W}_{t,\tau}\right]] \le R^2(1-\eta\mu)^{t-1-\tau}$.

Proof. For $\tau = t-1$, we have $\mathbb{E}[\boldsymbol{W}_{t,t-1}] = \mathsf{Tr}\left[\boldsymbol{H}\right] \leq R^2$. For $\tau < t-1$, we have by independence of each $\boldsymbol{x}_t$ that

$$\begin{aligned}\mathsf{Tr}\left[\mathbb{E}[\boldsymbol{W}_{t,\tau}]\right] &= \mathsf{Tr}\left[\mathbb{E}[\boldsymbol{Q}_{t-1}\cdots\boldsymbol{Q}_{\tau+1}\boldsymbol{H}\boldsymbol{Q}_{\tau+1}\cdots\boldsymbol{Q}_{t-1}]\right] = \mathsf{Tr}\left[\mathbb{E}[\boldsymbol{Q}_{t-1}\cdots\boldsymbol{Q}_{\tau}(\mathcal{P}\boldsymbol{H})\boldsymbol{Q}_{\tau}\cdots\boldsymbol{Q}_{t-1}]\right] = \cdots \\ &= \mathsf{Tr}\left[\mathcal{P}^{t-1-\tau}\boldsymbol{H}\right] .\end{aligned}$$

Recursively bounding $\mathsf{Tr}\left[\mathcal{P}^{\tau}H\right] = \mathsf{Tr}\left[\mathcal{P}(\mathcal{P}^{\tau-1}\boldsymbol{H})\right] \leq (1-\eta\mu)\mathsf{Tr}\left[\mathcal{P}^{\tau-1}\boldsymbol{H}\right]$ from Lemma F.18 completes the proof. □

Lemma D.10. *Consider $\boldsymbol{V}_{t,\tau}$ as defined in* (56)*. We have that*

$$\mathbb{E}\left[\mathsf{Tr}\left[\boldsymbol{V}_{t,\tau}\boldsymbol{V}_{t,\tau}^{\top}\right]\right] \leq d\left(\sum_{k=0}^{\tau}|\beta_k|(1-\eta\mu)^{(\tau-k)/2}\right) .$$

Further, if the event $\mathcal{E} = \cap_{\tau=1}^{t}\{\boldsymbol{Q}_t \succeq \boldsymbol{0}\}$ holds, then we also have

$$\mathsf{Tr}\left[\boldsymbol{V}_{t,\tau}\boldsymbol{V}_{t,\tau}^{\top}\right] \leq d\left(\sum_{k=0}^{\tau}|\beta_k|\right)^2 .$$

Proof. Since t is fixed throughout, we simply write $\boldsymbol{V}_{t,\tau}$ as $\boldsymbol{V}_{\tau}$. We define a sequence of matrices $\boldsymbol{A}_0, \ldots, \boldsymbol{A}_{\tau}$ as $\boldsymbol{A}_0 = \beta_0\boldsymbol{I}$ and

$$\boldsymbol{A}_{k+1} = \beta_{k+1}\boldsymbol{I} + \boldsymbol{Q}_{t-\tau+k}\boldsymbol{A}_k$$

for $k = 0, \ldots, \tau-1$. We first prove the expected bound followed by the absolute bound.

Expected bound: Then, we successively deduce the following.

(a) We have $\boldsymbol{A}_k = \beta_k\boldsymbol{I} + \beta_{k-1}\boldsymbol{Q}_{t-\tau+k-1} + \cdots + \beta_0\boldsymbol{Q}_{t-\tau+k-1}\ldots\boldsymbol{Q}_{t-\tau}$ by simply unrolling the recursions.

(b) We immediately recognize that $\boldsymbol{V}_{\tau} = \boldsymbol{A}_{\tau}$.

(c) By independence of each $\boldsymbol{Q}_t$, taking an expectation of the expression in (a) gives

$$\mathbb{E}[\boldsymbol{A}_k] = \sum_{l=0}^{k}\beta_l(\boldsymbol{I}-\eta\boldsymbol{H})^{k-l} .$$

(d) We establish a recursion

$$\mathbb{E}\mathsf{Tr}\left[\boldsymbol{A}_{k+1}\boldsymbol{A}_{k+1}^{\top}\right] \leq d\beta_{k+1}^2 + 2d|\beta_{k+1}|\sum_{l=0}^{k}|\beta_l|(1-\eta\mu)^{k-l+1} + (1-\eta\mu)\mathbb{E}\mathsf{Tr}\left[\boldsymbol{A}_k\boldsymbol{A}_k^{\top}\right] .$$

Indeed, by expanding out the square of the recursion and using the independence of the $\boldsymbol{x}_t$'s, we get

$$\begin{aligned}\mathbb{E}\mathsf{Tr}\left[\boldsymbol{A}_{k+1}\boldsymbol{A}_{k+1}^{\top}\right] &= \beta_{k+1}^2\mathsf{Tr}\left[\boldsymbol{I}\right] + 2\beta_{k+1}\mathsf{Tr}\left[(\boldsymbol{I}-\eta\boldsymbol{H})\mathbb{E}[\boldsymbol{A}_k]\right] + \mathsf{Tr}\left[\mathcal{P}(\mathbb{E}[\boldsymbol{A}_k\boldsymbol{A}_k^{\top}])\right] \\ &\leq d\beta_{k+1}^2 + 2|\beta_{k+1}|\sum_{l=0}^{k}|\beta_l|\ \mathsf{Tr}\left[(\boldsymbol{I}-\eta\boldsymbol{H})^{k-l+1}\right] + (1-\eta\mu)\mathbb{E}\mathsf{Tr}\left[\boldsymbol{A}_k\boldsymbol{A}_k^{\top}\right] ,\end{aligned}$$

where we plugged in the expression for $\mathbb{E}[\boldsymbol{A}_k]$ from item (c) and used Lemma F.18 to bound the last term. Using $\boldsymbol{0} \preceq \boldsymbol{I}-\eta\boldsymbol{H} \preceq (1-\eta\mu)\boldsymbol{I}$ gives the claimed expression.

(e) Using induction and the recursion from part (d), we prove that

$$\mathbb{E}\mathsf{Tr}\left[\boldsymbol{A}_k\boldsymbol{A}_k^{\top}\right] \leq d\left(\sum_{l=0}^{k}|\beta_l|(1-\eta\mu)^{(k-l)/2}\right)^2 .$$

Together with part (b), this gives the desired result.

Indeed, the base case holds because $\mathbb{E}\mathsf{Tr}\left[\boldsymbol{A}_0\boldsymbol{A}_0^\top\right] = \beta_0^2 d$. Supposing the induction hypothesis holds for some $k < \tau - 1$, we use the recursion of item (d) to get

$$\begin{aligned}\frac{1}{d}\,\mathbb{E}\mathsf{Tr}\left[\boldsymbol{A}_{k+1}\boldsymbol{A}_{k+1}^\top\right] &\le \beta_{k+1}^2 + 2|\beta_{k+1}|\sum_{l=0}^{k}|\beta_l|(1-\eta\mu)^{k-l+1} + \left(\sum_{l=0}^{k}|\beta_l|(1-\eta\mu)^{\frac{k-l+1}{2}}\right)^2 \\ &\le \beta_{k+1}^2 + 2|\beta_{k+1}|\sum_{l=0}^{k}|\beta_l|(1-\eta\mu)^{\frac{k-l+1}{2}} + \left(\sum_{l=0}^{k}|\beta_l|(1-\eta\mu)^{\frac{k-l+1}{2}}\right)^2 \\ &= \left(\sum_{l=0}^{k+1}|\beta_l|(1-\eta\mu)^{\frac{k-l+1}{2}}\right)^2,\end{aligned}$$

where the second inequality used $1 - \eta\mu \le 1$.

Absolute bound: Next, we prove the absolute bound, assuming that $\mathcal{E}$ holds. Again, we successively deduce:

(a) We starting with $\boldsymbol{A}_k = \beta_k\boldsymbol{I} + \beta_{k-1}\boldsymbol{Q}_{t-\tau+k-1} + \cdots + \beta_0\boldsymbol{Q}_{t-\tau+k-1}\cdots\boldsymbol{Q}_{t-\tau}$.

(b) Then, we get

$$|\mathsf{Tr}\left[\boldsymbol{A}_k\right]| \le |\beta_k|d + |\beta_{k-1}|\,|\mathsf{Tr}\left[\boldsymbol{Q}_{t-\tau+k-1}\right]| + \cdots + |\beta_0|\,|\mathsf{Tr}\left[\boldsymbol{Q}_{t-\tau+k-1}\cdots\boldsymbol{Q}_{t-\tau}\right]| \le d\sum_{l=0}^{k}|\beta_l|\,,$$

where we bound each of the traces by d using Lemma F.19 (since we have $\boldsymbol{Q}_k \preceq \boldsymbol{I}$ under $\mathcal{E}$).

(c) By a similar logic, we get

$$\begin{aligned}&\left|\mathsf{Tr}\left[\boldsymbol{Q}_{t-\tau+k}\boldsymbol{A}_k + \boldsymbol{A}_k^\top\boldsymbol{Q}_{t-\tau+k}\right]\right| \\ &\le 2|\beta_k|\mathsf{Tr}\left[\boldsymbol{Q}_{t-\tau+k}\right] + 2|\beta_1|\,|\mathsf{Tr}\left[\boldsymbol{Q}_{t-\tau+k}\boldsymbol{Q}_{t-\tau+k-1}\right]| + \cdots + 2|\beta_0|\,|\mathsf{Tr}\left[\boldsymbol{Q}_{t-\tau+k}\cdots\boldsymbol{Q}_{t-\tau}\right]| \\ &\le 2d\sum_{l=0}^{k}|\beta_l|\,.\end{aligned}$$

(d) We prove by induction that $\mathsf{Tr}\left[\boldsymbol{A}_k\boldsymbol{A}_k^\top\right] \le d\left(\sum_{l=0}^{k}|\beta_l|\right)^2$.

The base case holds since $\mathsf{Tr}\left[\boldsymbol{A}_0\boldsymbol{A}_0^\top\right] = d\beta_0^2$. Supposing the induction hypothesis holds for some integer $1 \le k < t-1$, we use the recursion of $\boldsymbol{A}_{k+1}$ to calculate

$$\begin{aligned}\mathsf{Tr}\left[\boldsymbol{A}_{k+1}\boldsymbol{A}_{k+1}^\top\right] &= d\beta_{k+1}^2 + \beta_{k+1}\mathsf{Tr}\left[\boldsymbol{Q}_{t-\tau+k}\boldsymbol{A}_k + \boldsymbol{A}_k^\top Q_{t-\tau+k}\right] + \mathsf{Tr}\left[\boldsymbol{Q}_{t-\tau+k}\boldsymbol{A}_k\boldsymbol{A}_k^\top\boldsymbol{Q}_{t-\tau+k}\right] \\ &\le d\beta_{k+1}^2 + 2d|\beta_{k+1}|\sum_{l=0}^{k}|\beta_l| + \mathsf{Tr}\left[\boldsymbol{A}_k\boldsymbol{A}_k^\top\right] \le d\left(\sum_{l=0}^{k+1}|\beta_l|\right)^2.\end{aligned}$$

Finally, item (d) together with $\boldsymbol{A}_\tau = \boldsymbol{V}_{t,\tau}$ completes the proof. □

D.3 Expected Bounds on Noisy-FTRL

Our goal of this section is to prove the following finite-time convergence guarantee of Noisy-FTRL in terms of the asymptotic suboptimality.

Theorem D.11. *Consider problem* (15) *and suppose Assumption C.2 holds. For a given a starting iterate $\boldsymbol{\theta}_0 \in \mathbb{R}^d$, weights $\boldsymbol{\beta} \in \ell^2$, learning rate $\eta < 1/R^2$, consider the sequence $(\boldsymbol{\theta}_t)_{t=0}^\infty$ produced by the iteration* (16) *where $\boldsymbol{w}_t \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ with $\sigma^2 = G^2\gamma_\infty(\boldsymbol{\beta})^2/(2\rho)$. Then, for any $t \ge 0$, we have,*

$$\mathbb{E}\left[F(\boldsymbol{\theta}_t) - F(\boldsymbol{\theta}_\star)\right] \le \left(\sqrt{\tfrac{L}{\mu}\exp(-\eta\mu t)\left(F(\boldsymbol{\theta}_0) - F(\boldsymbol{\theta}_\star)\right)} + \sqrt{F_\infty(\boldsymbol{\beta})}\right)^2.$$

We start with some preliminary lemmas. The first lemma is about the covariance of the noise process and is a generalization of [29, Lemma 3] to linearly correlated additive noise.

Lemma D.12. *Consider the sequence $(\widetilde{\boldsymbol{\theta}}_t)_{t=0}^{\infty}$ generated by Noisy-FTRL starting from $\widetilde{\boldsymbol{\theta}}_t = \boldsymbol{\theta}_\star$ with noise correlations $\boldsymbol{\beta} \in \ell^2$ and learning rate $\eta \le 1/R^2$. Under Assumption C.2, we have that its covariance*

$$\boldsymbol{S}_t := \mathbb{E}\left[\left(\widetilde{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_\star\right) \otimes \left(\widetilde{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_\star\right)\right]$$

satisfies: (a) $\boldsymbol{S}_t \preceq \boldsymbol{S}_{t+1}$ for all $t \ge 0$, and (b) the sequence $(\boldsymbol{S}_t)_{t=0}^{\infty}$ converges element-wise as $t \to \infty$.

Proof. Recall the notation $\boldsymbol{Q}_t = \boldsymbol{I} - \eta \boldsymbol{x} \otimes \boldsymbol{x}_t$ and $\mathcal{P}\boldsymbol{M} = \mathbb{E}[\boldsymbol{Q}_t \boldsymbol{M} \boldsymbol{Q}_t]$. We use the shorthand $\widetilde{\boldsymbol{\theta}}'_t := \widetilde{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_\star$. We first prove that the covariance is increasing in a PSD sense and argue that its limit exists.

Part 1: Non-decreasing noise: By unrolling the update equation and using $\widetilde{\boldsymbol{\theta}}'_t = \boldsymbol{0}$, we get

$$\begin{aligned}\widetilde{\boldsymbol{\theta}}'_t = \eta\left(\boldsymbol{x}_{t-1}\xi_{t-1} + \boldsymbol{Q}_{t-1}\boldsymbol{x}_{t-2}\xi_{t-2} + \cdots + \boldsymbol{Q}_{t-1}\cdots\boldsymbol{Q}_1\boldsymbol{x}_0\xi_0\right) \\ - \eta\left(\beta_0\boldsymbol{w}_{t-1} + (\beta_1\boldsymbol{I} + \beta_0\boldsymbol{Q}_{t-1})\boldsymbol{w}_{t-2} + \cdots + (\beta_{t-1}\boldsymbol{I} + \beta_{t-2}\boldsymbol{Q}_{t-1} + \cdots + \beta_0\boldsymbol{Q}_{t-1}\cdots\boldsymbol{Q}_1)\boldsymbol{w}_0\right).\end{aligned} \tag{71}$$

Next, we calculate $\mathbb{E}\left[\widetilde{\boldsymbol{\theta}}'_t \otimes \widetilde{\boldsymbol{\theta}}'_t\right]$. By independence, all the cross terms cancel out, so it suffices to write out the second moment of each of the terms above. For the SGD noise terms that contain $\boldsymbol{x}_\tau\xi_\tau$, we get for $\tau = 0, \ldots, t-1$ that

$$\mathbb{E}\left[(\boldsymbol{Q}_{t-1}\cdots\boldsymbol{Q}_{t-\tau+1}\boldsymbol{x}_{t-\tau}\xi_{t-\tau}) \otimes (\boldsymbol{Q}_{t-1}\cdots\boldsymbol{Q}_{t-\tau+1}\boldsymbol{x}_{t-\tau}\xi_{t-\tau})\right] = \mathcal{P}^\tau\left(\mathbb{E}[\xi^2\boldsymbol{x}\otimes\boldsymbol{x}]\right) =: \mathcal{T}_\tau\,. \tag{72}$$

Since it is a second moment term, we have $\mathcal{T}_\tau \succeq \boldsymbol{0}$. For the DP noise terms, denote $\boldsymbol{x}^{\otimes 2} = \boldsymbol{x}\otimes\boldsymbol{x} = \boldsymbol{x}\boldsymbol{x}^\top$. Then, we have for $\tau = 0$ to $t-1$ that

$$\begin{aligned}&\frac{1}{\sigma^2}\mathbb{E}\left((\beta_\tau\boldsymbol{I} + \beta_{\tau-1}\boldsymbol{Q}_{t-1} + \beta_{\tau-2}\boldsymbol{Q}_{t-1}\boldsymbol{Q}_{t-2} + \cdots + \beta_0\boldsymbol{Q}_{t-1}\cdots\boldsymbol{Q}_{t-\tau})\boldsymbol{w}_{t-\tau-1}\right)^{\otimes 2} \\ &\quad= \mathbb{E}\left(\beta_\tau\boldsymbol{I} + \beta_{\tau-1}\boldsymbol{Q}_{t-1} + \beta_{\tau-2}\boldsymbol{Q}_{t-1}\boldsymbol{Q}_{t-2} + \cdots + \beta_0\boldsymbol{Q}_{t-1}\cdots\boldsymbol{Q}_{t-\tau}\right)^{\otimes 2} \\ &\quad= \beta_\tau^2\boldsymbol{I} + 2\beta_\tau\sum_{k=0}^{\tau-1}\beta_k(\boldsymbol{I}-\eta\boldsymbol{H})^{\tau-k} + \sum_{k=0}^{\tau-1}\sum_{l=0}^{\tau-1}\beta_k\beta_l\,\mathbb{E}\left[\boldsymbol{Q}_{t-1}\cdots\boldsymbol{Q}_{t-\tau+k}\boldsymbol{Q}_{t-\tau+l}\cdots\boldsymbol{Q}_{t-1}\right] \\ &\quad= \beta_\tau^2\boldsymbol{I} + 2\beta_\tau\sum_{k=0}^{\tau-1}\beta_k(\boldsymbol{I}-\eta\boldsymbol{H})^{\tau-k} + 2\sum_{k=0}^{\tau-1}\sum_{l=0}^{k}\beta_k\beta_l\,\mathbb{E}\left[\boldsymbol{Q}_{t-1}\cdots\boldsymbol{Q}_{t-\tau+l}(\boldsymbol{I}-\eta\boldsymbol{H})^{k-l}\boldsymbol{Q}_{t-\tau+l}\cdots\boldsymbol{Q}_{t-1}\right] \\ &\quad= \beta_\tau^2\boldsymbol{I} + 2\beta_\tau\sum_{k=0}^{\tau-1}\beta_k(\boldsymbol{I}-\eta\boldsymbol{H})^{\tau-k} + 2\sum_{k=0}^{\tau-1}\sum_{l=0}^{k}\beta_k\beta_l\,\mathcal{P}^{\tau-k}\left((\boldsymbol{I}-\eta\boldsymbol{H})^{k-l}\right) =: \mathcal{T}'_\tau\,.\end{aligned} \tag{73}$$

By virtue of this being a second moment, we have that $\mathcal{T}'_\tau \succeq \boldsymbol{0}$. Plugging in (72) and (73) into the second moment of (71), we get,

$$\begin{aligned}\mathbb{E}\left[\widetilde{\boldsymbol{\theta}}'_{t+1}\otimes\widetilde{\boldsymbol{\theta}}'_{t+1}\right] &= \eta^2\sum_{\tau=0}^{t}(\mathcal{T}_\tau + \sigma^2\mathcal{T}'_\tau) \\ &= \mathbb{E}\left[\widetilde{\boldsymbol{\theta}}'_t\otimes\widetilde{\boldsymbol{\theta}}'_t\right] + \eta^2(\mathcal{T}_t + \sigma^2\mathcal{T}'_t) \succeq \mathbb{E}\left[\widetilde{\boldsymbol{\theta}}'_t\otimes\widetilde{\boldsymbol{\theta}}'_t\right].\end{aligned}$$

This shows that the noise is non-decreasing in a PSD sense.

Part 2: Convergence of the covariance: Next, we show that the noise sequence converges. From the update equation $\widetilde{\boldsymbol{\theta}}'_{t+1} = \boldsymbol{Q}_t\widetilde{\boldsymbol{\theta}}'_t + \eta\boldsymbol{x}_t\xi_t - \eta\sum_{\tau=0}^{t}\beta_\tau\boldsymbol{w}_{t-\tau}$, we get

$$\begin{aligned}\boldsymbol{S}_{t+1} = \mathcal{P}\boldsymbol{S}_t + \eta^2\mathbb{E}[\xi^2\boldsymbol{x}\otimes\boldsymbol{x}] + \eta^2\sigma^2\sum_{\tau=0}^{t}\beta_\tau^2\boldsymbol{I} \\ - \eta(\boldsymbol{I}-\eta\boldsymbol{H})\sum_{\tau=0}^{t}\beta_\tau\mathbb{E}\left[\widetilde{\boldsymbol{\theta}}'_t\otimes\boldsymbol{w}_{t-\tau}\right] - \eta\sum_{\tau=0}^{t}\beta_\tau\mathbb{E}\left[\boldsymbol{w}_{t-\tau}\otimes\widetilde{\boldsymbol{\theta}}'_t\right](\boldsymbol{I}-\eta\boldsymbol{H})\,.\end{aligned}$$

For $\tau = 0$, the term $\mathbb{E}[\widetilde{\boldsymbol{\theta}}_t' \otimes \boldsymbol{w}_{t-\tau}]$ and its transpose are both $\mathbf{0}$. For $\tau > 0$, we have from (71) that

$$\begin{aligned}-\mathbb{E}\left[\widetilde{\boldsymbol{\theta}}_t' \otimes \boldsymbol{w}_{t-\tau}\right] &= \eta\mathbb{E}\left[\beta_{\tau-1}\boldsymbol{I} + \beta_{\tau-2}\boldsymbol{Q}_{t-1} + \cdots + \beta_0\boldsymbol{Q}_{t-1}\cdots\boldsymbol{Q}_{t-\tau+1}\right]\ \mathbb{E}[\boldsymbol{w}_{t-\tau}\otimes\boldsymbol{w}_{t-\tau}] \\ &= \eta\sigma^2\left(\beta_{\tau-1}\boldsymbol{I} + \beta_{\tau-2}(\boldsymbol{I}-\eta\boldsymbol{H}) + \cdots + \beta_0(\boldsymbol{I}-\eta\boldsymbol{H})^{\tau-1}\right).\end{aligned}$$

Plugging this back in gives

$$\begin{aligned}\boldsymbol{S}_{t+1} &= \mathcal{P}\boldsymbol{S}_t + \eta^2\mathbb{E}[\xi^2\boldsymbol{x}\otimes\boldsymbol{x}] + \eta^2\sigma^2\sum_{\tau=0}^{t}\beta_\tau^2\boldsymbol{I} + 2\eta^2\sigma^2\sum_{\tau=1}^{t}\sum_{k=0}^{\tau-1}\beta_\tau\beta_k(\boldsymbol{I}-\eta\boldsymbol{H})^{\tau-k} \\ &= \mathcal{P}\boldsymbol{S}_t + \eta^2\mathbb{E}[\xi^2\boldsymbol{x}\otimes\boldsymbol{x}] + \eta^2\sigma^2\sum_{\tau=0}^{t}\sum_{k=0}^{t}\beta_\tau\beta_k(\boldsymbol{I}-\eta\boldsymbol{H})^{|\tau-k|}. \end{aligned} \tag{74}$$

Next, we take a trace of (74). For the first term, we get

$$\begin{aligned}\mathsf{Tr}\left[\mathcal{P}\boldsymbol{S}_t\right] &= \mathsf{Tr}\left[\boldsymbol{S}_t\right] - 2\eta\mathsf{Tr}\left[\boldsymbol{H}\boldsymbol{S}_t\right] + \eta^2\mathsf{Tr}\left[\boldsymbol{S}_t\mathbb{E}[\|\boldsymbol{x}_t\|_2^2\boldsymbol{x}_t\otimes\boldsymbol{x}_t]\right] \\ &\le \mathsf{Tr}\left[\boldsymbol{S}_t\right] - \eta\mathsf{Tr}\left[\boldsymbol{H}\boldsymbol{S}_t\right](2-\eta R^2) \\ &\le (1-\eta\mu)\mathsf{Tr}\left[\boldsymbol{S}_t\right],\end{aligned}$$

where we use (a) $\mathbb{E}[\|\boldsymbol{x}_t\|_2^2\boldsymbol{x}_t\otimes\boldsymbol{x}_t] \preceq R^2\boldsymbol{H}$, (b) $\eta \le 1/R^2$, and (c) $\boldsymbol{H} \succeq \mu\boldsymbol{I}$. By assumption, we also get that $\mathsf{Tr}\left[\mathbb{E}[\xi^2\boldsymbol{x}\otimes\boldsymbol{x}]\right] \le \sigma_{\mathsf{sgd}}^2\mathsf{Tr}\left[\boldsymbol{H}\right] \le \sigma_{\mathsf{sgd}}^2R^2$. Finally, we have using Lemma F.17 that

$$\sum_{\tau=0}^{t}\sum_{k=0}^{t}\beta_\tau\beta_k\sum_{j=1}^{d}(1-\eta\lambda_j)^{|\tau-k|} \le \|\boldsymbol{\beta}\|_2^2\sum_{j=1}^{d}\left(\frac{2-\eta\lambda_j}{\eta\lambda_j}\right) \le \frac{2\|\boldsymbol{\beta}\|_2^2\mathsf{Tr}\left[\boldsymbol{H}^{-1}\right]}{\eta}.$$

Thus, we get

$$\mathsf{Tr}\left[\boldsymbol{S}_{t+1}\right] \le (1-\eta\mu)\mathsf{Tr}\left[\boldsymbol{S}_t\right] + 2\eta\sigma^2\|\boldsymbol{\beta}\|_2^2\,\mathsf{Tr}\left[\boldsymbol{H}^{-1}\right] + \eta^2R^2\sigma_{\mathsf{sgd}}^2.$$

By unrolling this out, we get a uniform bound for all t:

$$\mathsf{Tr}\left[\boldsymbol{S}_t\right] \le \frac{1}{\mu}\left(2\sigma^2\|\boldsymbol{\beta}\|_2^2\,\mathsf{Tr}\left[\boldsymbol{H}^{-1}\right] + \eta R^2\sigma_{\mathsf{sgd}}^2\right) < \infty$$

since $\boldsymbol{\beta} \in \ell^2$. For any fixed vector $\boldsymbol{v}$, $\langle \boldsymbol{v}, \boldsymbol{S}_t\boldsymbol{v}\rangle$ thus has a limit from the monotone convergence theorem. From this, it follows that every diagonal entry of $\boldsymbol{S}_t$ converges (take $\boldsymbol{v}$ as a standard basis vector) and then every off-diagonal entry of $\boldsymbol{S}_t$ also converges (take $\boldsymbol{v}$ as the sum of two standard basis vectors). This shows that $\boldsymbol{S}_t$ converges element-wise. □

We are now ready to prove Theorem D.11.

Proof of Theorem D.11. Define $F_\infty^\star(\boldsymbol{\beta})$ as the asymptotic suboptimality of a process that starts from $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_\star$. We will prove the desired result with $F_\infty^\star(\boldsymbol{\beta})$ in the place of $F_\infty(\boldsymbol{\beta})$. Finally, we will show that $F_\infty(\boldsymbol{\beta})$ is independent of its starting iterate so $F_\infty(\boldsymbol{\beta}) = F_\infty^\star(\boldsymbol{\beta})$.

We first separate out the effects of the noise and the initial iterate using Property D.2. We invoke Lemma D.12 for the former and directly bound the latter. Lastly, we combine them both with a triangle inequality. Recall that use the shorthand $\boldsymbol{\theta}_t' := \boldsymbol{\theta}_t - \boldsymbol{\theta}_\star$ and $\boldsymbol{Q}_t := \boldsymbol{I} - \eta\boldsymbol{x}_t\otimes\boldsymbol{x}_t$.

Effect of the initialization: We first calculate

$$\mathbb{E}[\boldsymbol{Q}_t^2] = \boldsymbol{I} - 2\eta\boldsymbol{H} + \eta^2\mathbb{E}\left[\|\boldsymbol{x}_t\|_2^2\boldsymbol{x}_t\otimes\boldsymbol{x}_t\right] \preceq \boldsymbol{I} - 2\eta\boldsymbol{H} + \eta^2R^2\boldsymbol{H} \preceq \boldsymbol{I} - \eta\boldsymbol{H} \preceq (1-\eta\mu)\boldsymbol{I},$$

where the first inequality follows from (47), the second since $\eta \le 1/R^2$, and the third since $\boldsymbol{H} \succeq \mu\boldsymbol{I}$. Letting $\mathcal{F}_t$ denote the sigma algebra generated by $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{t-1}$, we get

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{\theta}}_{t+1}\right\|_2^2\middle|\mathcal{F}_t\right] = \left\langle\hat{\boldsymbol{\theta}}_t, \mathbb{E}[\boldsymbol{Q}_t^2]\hat{\boldsymbol{\theta}}_t\right\rangle \le (1-\eta\mu)\left\|\hat{\boldsymbol{\theta}}_t\right\|_2^2 \le \exp(-\eta\mu)\left\|\hat{\boldsymbol{\theta}}_t\right\|_2^2.$$

Taking an unconditional expectation and unrolling this and using $\mu \boldsymbol{I} \preceq \boldsymbol{H} \preceq L\boldsymbol{I}$ (Assumption (**A1**)) gives

$$\mathbb{E}\left\|\hat{\boldsymbol{\theta}}_t\right\|_{\boldsymbol{H}}^2 \leq L\,\mathbb{E}\left\|\hat{\boldsymbol{\theta}}_t\right\|_2^2 \leq L\,\exp(-\eta\mu t)\,\|\boldsymbol{\theta}_0'\|_2^2 \leq \frac{L}{\mu}\,\exp(-\eta\mu t)\,\|\boldsymbol{\theta}_0'\|_{\boldsymbol{H}}^2\,. \tag{75}$$

Effect of the noise: Define $\widetilde{\boldsymbol{\theta}}_t' := \widetilde{\boldsymbol{\theta}}_t^{\mathsf{sgd}} + \widetilde{\boldsymbol{\theta}}_t^{\mathsf{dp}}$. We get from Lemma D.12 that there exists a PSD matrix $\boldsymbol{S}_\infty$ such that

$$\mathbf{0} = \mathbb{E}\left[\widetilde{\boldsymbol{\theta}}_0' \otimes \widetilde{\boldsymbol{\theta}}_0'\right] \preceq \mathbb{E}\left[\widetilde{\boldsymbol{\theta}}_1' \otimes \widetilde{\boldsymbol{\theta}}_1'\right] \preceq \cdots \preceq \lim_{t\to\infty} \mathbb{E}\left[\widetilde{\boldsymbol{\theta}}_t' \otimes \widetilde{\boldsymbol{\theta}}_t'\right] =: \boldsymbol{S}_\infty\,.$$

Multiplying by $\boldsymbol{H}$ and taking a trace, we get,

$$0 \leq \mathbb{E}\left\|\widetilde{\boldsymbol{\theta}}_0'\right\|_{\boldsymbol{H}}^2 \leq \mathbb{E}\left\|\widetilde{\boldsymbol{\theta}}_1'\right\|_{\boldsymbol{H}}^2 \leq \cdots \leq \lim_{t\to\infty} \mathbb{E}\left\|\widetilde{\boldsymbol{\theta}}_t'\right\|_{\boldsymbol{H}}^2 = \mathsf{Tr}\left[\boldsymbol{H}\boldsymbol{S}_\infty\right]\,. \tag{76}$$

Thus, $\widetilde{\boldsymbol{\theta}}_t = \widetilde{\boldsymbol{\theta}}_t' + \boldsymbol{\theta}_\star$ is a process that starts from $\widetilde{\boldsymbol{\theta}}_0 = \boldsymbol{\theta}_\star$ and satisfies the conditions of Lemma D.12. This in turn gives

$$0 \leq \mathbb{E}\left[F(\widetilde{\boldsymbol{\theta}}_0) - F(\boldsymbol{\theta}_\star)\right] \leq \mathbb{E}\left[F(\widetilde{\boldsymbol{\theta}}_1) - F(\boldsymbol{\theta}_\star)\right] \leq \cdots \leq \lim_{t\to\infty} \mathbb{E}\left[F(\widetilde{\boldsymbol{\theta}}_t) - F(\boldsymbol{\theta}_\star)\right] = \frac{1}{2}\mathsf{Tr}\left[\boldsymbol{H}\boldsymbol{S}_\infty\right]\,, \tag{77}$$

which equals $F_\infty^\star(\boldsymbol{\beta})$ by definition.

Combining both processes: From the triangle inequality of the norm $\boldsymbol{u} \mapsto \sqrt{\mathbb{E}\|\boldsymbol{u}\|_{\boldsymbol{H}}^2}$, we get

$$\sqrt{\mathbb{E}\|\boldsymbol{\theta}_t'\|_{\boldsymbol{H}}^2} \leq \sqrt{\mathbb{E}\left\|\hat{\boldsymbol{\theta}}_t\right\|_{\boldsymbol{H}}^2} + \sqrt{\mathbb{E}\left\|\widetilde{\boldsymbol{\theta}}_t'\right\|_{\boldsymbol{H}}^2}\,.$$

Plugging in (75) and (76) gives

$$\begin{aligned}\sqrt{\mathbb{E}\left[F(\boldsymbol{\theta}_t) - F(\boldsymbol{\theta}_\star)\right]} &\leq \sqrt{\frac{L}{2\mu}\exp(-\eta\mu t)\left\|\hat{\boldsymbol{\theta}}_0'\right\|_{\boldsymbol{H}}^2} + \sqrt{\frac{1}{2}\mathsf{Tr}\left[\boldsymbol{H}\boldsymbol{S}_\infty\right]} \\ &= \sqrt{\frac{L}{\mu}\exp(-\eta\mu t)\left(F(\boldsymbol{\theta}_0) - F(\boldsymbol{\theta}_\star)\right)} + \sqrt{F_\infty^\star(\boldsymbol{\beta})}\,,\end{aligned}$$

where the last equality followed from (77). This establishes the required statement with $F_\infty^\star$ in place of F^∞. Taking $t \to \infty$, we see that

$$\sqrt{F_\infty(\boldsymbol{\beta})} = \lim_{t\to\infty} \sqrt{\mathbb{E}\left[F(\boldsymbol{\theta}_t) - F(\boldsymbol{\theta}_\star)\right]} = \sqrt{F_\infty^\star(\boldsymbol{\beta})}\,,$$

for any fixed η or that $F_\infty = F_\infty^\star$ irrespective of $\boldsymbol{\theta}_0$. □

D.4 Privacy-Utility Guarantees of DP-FTRL

We now state a general privacy-utility bound for DP-FTRL in terms of the asymptotics of Noisy-FTRL run with the same parameters.

Theorem D.13. *Fix a constant $0 < p < 1$ and suppose the Assumption D.1 holds. Fix some noise coefficients $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{T-1})$ that satisfy Half-Expo Decay with parameter $\eta\widetilde{\nu}$ for some $\widetilde{\nu} \leq \mu$. Consider the sequence $(\boldsymbol{\theta}_t)_{t=0}^{T-1}$ of iterates and the sequence $(\boldsymbol{g}_t)_{t=0}^{T-1}$ of gradients when running DP-FTRL for T iterations with noise coefficients $\boldsymbol{\beta}$, gradient clip norm $G = cR^2 \max\left\{\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_\star\|_2, \sqrt{\eta R^2 \sigma_{\mathsf{sgd}}^2/\mu}, \sigma_{\mathsf{sgd}}/R\right\} \log^{5/2}\left(\frac{T}{p}\right)$, and a learning rate*

$$\eta \leq \min\left\{\frac{1}{CR^2\log(T/p)}, \frac{\widetilde{\nu}\rho}{8C^2R^4 d\gamma_\infty(\boldsymbol{\beta})^2\|\boldsymbol{\beta}\|_1^2\log^5(T/p)}\right\}\,,$$

and DP noise $\boldsymbol{w}_t \sim \mathcal{N}(\mathbf{0}, \sigma_{\mathsf{dp}}^2 G^2 \boldsymbol{I})$ with squared noise multiplier $\sigma_{\mathsf{dp}}^2 = \gamma(\boldsymbol{\beta})^2/(2\rho)$. Then, we have the following:

(a) $(\boldsymbol{\theta}_t)_{t=0}^T$ *is* ρ*-zCDP.*

(b) Let $\mathcal{E}$ *denote the event where no gradients are clipped, i.e,* $\mathcal{E} = \cap_{t=0}^{T-1}\{\|\boldsymbol{g}_t\|_2 \leq G\}$*. We have,* $\mathbb{P}(\mathcal{E}) \geq 1 - p$*.*

(c) We have,

$$\mathbb{E}\left[(F(\boldsymbol{\theta}_t) - F(\boldsymbol{\theta}_\star)) \cdot \mathbb{1}\left(\mathcal{E}\right)\right] \leq \frac{2L}{\mu}\exp(-\eta\mu t)\left(F(\boldsymbol{\theta}_0) - F(\boldsymbol{\theta}_\star)\right) + 2\,\hat{F}_\infty(\boldsymbol{\beta})\,,$$

where $\hat{F}_\infty(\boldsymbol{\beta})$ *is the asymptotic suboptimality of Noisy-FTRL run with the same parameters.*

Proof. Part (a) follows from Theorem 2.1. For part (b), we bound the gradient norms from Theorem D.4 as

$$\begin{aligned}
\|\boldsymbol{g}_t\|_2 &\leq CR^2\left(\|\boldsymbol{\theta}_0'\|_2 + \sqrt{\frac{\eta R^2\sigma_{\mathsf{sgd}}^2}{\mu}} + \frac{\sigma_{\mathsf{sgd}}}{R} + G\sqrt{\frac{\eta\sigma^2 d\|\boldsymbol{\beta}\|_1^2}{\widetilde{\nu}}}\right)\log^{5/2}\left(\frac{T}{p}\right) \\
&\leq CR^2\left(\|\boldsymbol{\theta}_0'\|_2 + \sqrt{\frac{\eta R^2\sigma_{\mathsf{sgd}}^2}{\mu}} + \frac{\sigma_{\mathsf{sgd}}}{R}\right)\log^{5/2}\left(\frac{T}{p}\right) + \frac{G}{4} \\
&\leq 4\max\left\{CR^2\max\left\{\|\boldsymbol{\theta}_0'\|_2, \sqrt{\frac{\eta R^2\sigma_{\mathsf{sgd}}^2}{\mu}}, \frac{\sigma_{\mathsf{sgd}}}{R}\right\}\log^{5/2}\left(\frac{T}{p}\right), \frac{G}{4}\right\} \leq G
\end{aligned}$$

where the second inequality followed from the condition on the learning rate and we take $c = 4C$ in the definition of G for the last inequality. Thus, $\mathcal{E}$ holds whenever the bound of Theorem D.4 holds, so we have $\mathbb{P}(\mathcal{E}) \geq 1 - p$.

For part (c), consider the sequence $(\boldsymbol{\phi}_t)_{t=0}^T$ produced by running Noisy-FTRL with $\boldsymbol{\phi}_0 = \boldsymbol{\theta}_0$ and the same realizations $(\boldsymbol{x}_t, \xi_t, \boldsymbol{w}_t)$ of random inputs, linear model noise, and DP noise. On $\mathcal{E}$, we have that $\boldsymbol{\phi}_t = \boldsymbol{\theta}_t$ for all t. Thus, we have,

$$\mathbb{E}\left[(F(\boldsymbol{\theta}_t) - F(\boldsymbol{\theta}_\star)) \cdot \mathbb{1}\left(\mathcal{E}\right)\right] = \mathbb{E}\left[(F(\boldsymbol{\phi}_t) - F(\boldsymbol{\theta}_\star)) \cdot \mathbb{1}\left(\mathcal{E}\right)\right] \leq \mathbb{E}\left[F(\boldsymbol{\phi}_t) - F(\boldsymbol{\theta}_\star)\right]\,,$$

since $\mathbb{1}\left(\mathcal{E}\right) \leq 1$. This can now be bounded using Theorem D.11 to complete the proof. □

We can instantiate these rates for DP-SGD and DP-FTRL. Recall that we have $\kappa = L/\mu$, $d_{\mathsf{eff}} = \mathsf{Tr}\left[\boldsymbol{H}\right]/L$, and $R^2 = \Theta(\mathsf{Tr}\left[\boldsymbol{H}\right])$.

Corollary D.14. *Consider the setting of Theorem D.13 with* T *large enough that* $T/\log^5(T/p) \geq c\kappa^2 d_{\mathsf{eff}}^2 d/\rho$*. The the final suboptimality of DP-SGD at an appropriate choice of the learning rate is (ignoring absolute constants),*

$$\begin{aligned}
\mathbb{E}\left[(F(\boldsymbol{\theta}_T) - F(\boldsymbol{\theta}_\star)) \cdot \mathbb{1}\left(\mathcal{E}\right)\right] \leq \frac{L}{\mu}\exp\left(-\frac{\rho T}{c\kappa^2 d_{\mathsf{eff}}^2 d\log^5(T/p)}\right) \\
+ \kappa\, d_{\mathsf{eff}}\left(\frac{d\mathsf{Tr}\left[\boldsymbol{H}\right]\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_\star\|_2^2}{\rho T} + \frac{d\sigma_{\mathsf{sgd}}^2}{\rho T} + \frac{\sigma_{\mathsf{sgd}}^2}{T}\right)\mathrm{polylog}\left(T\right)\,.
\end{aligned}$$

Proof. We plug in the asymptotic suboptimality bound of Noisy-SGD into the bound of Theorem D.13. We get two terms depending on the learning rate η: the first $\exp(-\mu\eta T)$ term and the second $O(\eta)$ term coming from the asymptotic suboptimality. We balance both the terms subject to the maximum bound on η using Lemma F.21 to get

$$\begin{aligned}
\mathbb{E}\left[(F(\boldsymbol{\theta}_T) - F(\boldsymbol{\theta}_\star)) \cdot \mathbb{1}\left(\mathcal{E}\right)\right] \leq \frac{L}{\mu}\exp\left(-\frac{\rho\mu^2 T}{cR^4 d\log^5(T/p)}\right) \\
+ \frac{\mathrm{polylog}\left(T\right)}{\mu T}\left(\frac{dR^4\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_\star\|_2^2}{\rho} + \frac{d\sigma_{\mathsf{sgd}}^2 R^2}{\rho} + \sigma_{\mathsf{sgd}}^2 R^2\right)\,.
\end{aligned}$$

Rearranging the constants completes the proof. □

Corollary D.15. *Consider the setting of Theorem D.13 with T large enough that $T/\log^7(T/p) \geq \frac{c\kappa^2 d_{\mathsf{eff}}^2 d}{\rho}\log\left(\frac{c\kappa^2 d_{\mathsf{eff}}^2 d}{\rho}\right)$. For ν-DP-FTRL with an appropriate choice of the parameter ν and learning rate η, we have (ignoring absolute constants),*

$$\mathbb{E}\left[(F(\boldsymbol{\theta}_T) - F(\boldsymbol{\theta}_\star)) \cdot \mathbb{1}\left(\mathcal{E}\right)\right] \leq \frac{L}{\mu}\exp\left(-\frac{\rho T}{c\kappa^2 d_{\mathsf{eff}}^2\, d\log^7(T/p)\log(\kappa^2 d_{\mathsf{eff}}^2\, d/\rho)}\right) + \kappa d_{\mathsf{eff}}\left(\frac{\kappa d_{\mathsf{eff}}\mathsf{Tr}\left[\boldsymbol{H}\right]\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_\star\|_2^2}{\rho T^2} + \frac{\kappa d_{\mathsf{eff}}\sigma_{\mathsf{sgd}}^2}{\rho T^2} + \frac{\sigma_{\mathsf{sgd}}^2}{T}\right)\mathrm{polylog}\left(T\right).$$

Proof. We plug in the asymptotic error for ν-Noisy-FTRL from Proposition C.20 into Theorem D.13 to get that

$$\mathbb{E}\left[(F(\boldsymbol{\theta}_T) - F(\boldsymbol{\theta}_\star)) \cdot \mathbb{1}\left(\mathcal{E}\right)\right] \leq \frac{L}{\mu}\exp(-\mu\eta T) + \eta\sigma_{\mathsf{sgd}}^2 R^2 + \eta^2\frac{R^2G^2}{\rho}\log^2\frac{1}{\eta\mu}, \tag{78}$$

where G^2 is as given in the statement of Theorem D.13. For our choice of $\boldsymbol{\beta}$, we have $\|\boldsymbol{\beta}\|_1^2 \leq 4$ always and $\gamma(\boldsymbol{\beta})^2 \leq 5\log(1/\eta\mu)$ from Equation (42) (from the proof of Proposition C.20). Thus, the largest learning rate permitted must satisfy

$$\eta\log^2\frac{1}{\eta\mu} \leq \frac{\eta\rho}{cR^2 d\log^5(T/p)}.$$

From Lemma F.22, we can ensure with a more stringent condition

$$\eta \leq \frac{\mu\rho}{cR^4 d\log^5(T/p)\log^2(cR^4 d\log^(T/p)/(\mu^2\rho))}.$$

Finally, this is implied by imposing the requirement

$$\eta \leq \frac{\mu\rho}{cR^4 d\log^7(T/p)\log\left(\frac{R^4 d}{\mu^2\rho}\right)} =: \eta_{\max}.$$

We now tune η to minimize the bound (78) subject to $\eta \leq \eta_{\max}$ using Lemma F.21. Thus gives,

$$\mathbb{E}\left[(F(\boldsymbol{\theta}_T) - F(\boldsymbol{\theta}_\star)) \cdot \mathbb{1}\left(\mathcal{E}\right)\right] \leq \frac{L}{\mu}\exp\left(-\frac{\rho\mu^2 T}{cR^4 d\log^7(T/p)\log\frac{R^4 d}{\rho\mu^2}}\right) + \frac{\mathrm{polylog}\left(T\right)}{\mu T}\left(\frac{R^6\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_\star\|_2^2}{\rho\mu T} + \frac{R^4\sigma_{\mathsf{sgd}}^2}{\rho\mu^2 T^2} + \sigma_{\mathsf{sgd}}^2 R^2\right).$$

Rewriting the constants completes the proof. □

E Proofs for General Strongly Convex Functions

We now generalize §4 to general strongly convex problems. Here, we bound the asymptotic suboptimality of DP-FTRL and DP-SGD by the value of a convex program.

Theorem E.1. *Suppose $f(\,\cdot\,;\boldsymbol{z})$ is G-Lipschitz, and the stochastic gradients are uniformly bounded as $\|\nabla_\theta f\left(\boldsymbol{\theta};\boldsymbol{z}\right) - \mathbb{E}_{\boldsymbol{z}'\sim\mathbb{P}_{\mathsf{data}}}\left[\nabla_\theta f\left(\boldsymbol{\theta};\boldsymbol{z}'\right)\right]\|_2 \leq \sigma_{\mathsf{sgd}}$. Then, if F is μ-strongly convex and L-smooth, we can bound F_∞ for any noise correlation $B\left(\omega\right)$ in the frequency domain as:*

$$\min_{\substack{\psi:[-\pi,\pi]\to\mathbb{R}_+ \\ \psi\in\mathcal{C}(\eta,L,\mu)}} \frac{Ld}{2\pi}\int_{-\pi}^{\pi}\left(G^2\rho^{-1}|B\left(\omega\right)|^2\gamma_\infty(B)^2 + \sigma_{\mathsf{sgd}}^2\right)\psi(\omega)\,\mathrm{d}\omega, \tag{79}$$

where $\gamma_\infty(B)$ is the limiting sensitivity in the Fourier domain, and $\mathcal{C}\left(\eta,\mu,L\right)$ is a convex set (details in Appendix E).

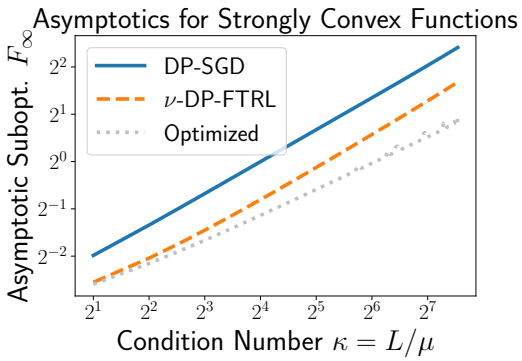

Figure 4: **DP-FTRL attains a tighter bound on** F_∞ with the growing condition number. Here, "Optimized" approximately minimizes (79). The plots holds for smooth and strongly convex functions ($L = 1 = G, \sigma_{\text{sgd}} = 0$).

While technically an infinite dimensional optimization problem over the function ψ, we can approximate the solution by discretizing ψ into k points uniformly over $[-\pi, \pi]$. Further, if we discretize B similarly, we can obtain a **second-order cone program** with k conic constraints and $O(k)$ decision variables. As $k \to \infty$, the solution approaches the solution to (79). Empirically, we observe that the values stabilize quickly as k increases. We stop the computation when the change in bound as a function of k drops below a threshold. We use $k = 1000$ for.

Further, given the optimal $\psi = \psi^\star$, we can run an alternating minimization where we minimize the objective of (79) with respect to ψ for fixed B and with respect to B for fixed ψ. This leads to an iteratively improving choice of B. We find empirically that this iterative procedure converges quickly and leads to a provable theoretical gap between the upper bounds on F_∞ achievable by DP-SGD and DP-FTRL.

We numerically compare the bound (79) for DP-SGD and ν-DP-FTRL with weights from (7). Figure 4 shows that the gap between DP-SGD and ν-DP-FTRL is multiplicative, i.e., the absolute gap grows with the increasing condition number $\kappa = L/\mu$ (which reflects practical scenarios). The suboptimality of "Optimized" DP-FTRL (optimized as described above) grows even more slowly with κ.

Overall, ν-DP-FTRL significantly improves upon DP-SGD and has only a single tunable parameter ν and no expensive computation to generate the noise correlations. We focus on ν-DP-FTRL for experiments in this paper, but leave the possibility of improving results further based on Optimized DP-FTRL for future work.

E.1 Proofs

We prove the results from Theorem E.1. Under the assumptions of the theorem, clipping does not occur in DP-FTRL so the updates can be written as

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \left((\boldsymbol{B}\boldsymbol{w})_t + (\boldsymbol{g}_t + \hat{\boldsymbol{w}}_t) \right) \tag{80}$$

where

$$\boldsymbol{g}_t = \nabla F\left(\boldsymbol{\theta}_t\right), \hat{\boldsymbol{w}}_t = \nabla f\left(\boldsymbol{\theta}_t; \boldsymbol{z}_t\right) - \mathbb{E}_{\boldsymbol{z} \sim \mathbb{P}_{\text{data}}}\left[\nabla f\left(\boldsymbol{\theta}_t; \boldsymbol{z}\right)\right]$$

where $\hat{\boldsymbol{w}}_t$ is a random variable that, conditioned on $\boldsymbol{\theta}_t$, is bounded by σ_{sgd} with probability 1.

Let $\boldsymbol{I}_d$ denote the $d \times d$ identity matrix.

Theorem E.2. *Let* $\boldsymbol{\lambda} = \{\lambda_t\}_{t=-\infty}^{\infty}$ *be such that* $\lambda_t \geq 0 \quad \forall t \in \mathbb{Z}$,

$$\sum_{t=-\infty}^{\infty} \lambda_t \leq 2\lambda_0$$

and let Λ denote the Discrete-time Fourier transform (DTFT) of $\boldsymbol{\lambda}$. Let

$$M_\lambda(\omega) = A(\omega)^{*\top} \widetilde{M}_\lambda(\omega) A(\omega) \tag{81a}$$

$$A(\omega) = \begin{pmatrix} \eta \boldsymbol{I}_d & 0 \\ (1 - \exp(i\omega)) \boldsymbol{I}_d & -\eta \boldsymbol{I}_d \end{pmatrix} \tag{81b}$$

$$\widetilde{M}_\lambda(\omega) = \begin{pmatrix} -\mu L \left(\Lambda(\omega) + \Lambda(\omega)^*\right) \boldsymbol{I}_d & \mu \Lambda(\omega) \boldsymbol{I}_d + L \Lambda(\omega)^* \boldsymbol{I}_d \\ \mu \Lambda^*(\omega) \boldsymbol{I}_d + L \Lambda(\omega) \boldsymbol{I}_d & -\left(\Lambda(\omega) + \Lambda(\omega)^*\right) \boldsymbol{I}_d \end{pmatrix} \tag{81c}$$

Then, for any non-negative valued function $\psi : [-\pi, \pi] \mapsto \mathbb{R}_+$ such that

$$M_\lambda(\omega) \preceq \begin{pmatrix} -\eta^2 \boldsymbol{I}_d & 0 \\ 0 & \psi(\omega) \boldsymbol{I}_d \end{pmatrix} \quad \forall \omega \in [-\pi, \pi] \tag{82}$$

We have that

$$\lim_{t \to \infty} \mathbb{E}\left[\frac{\sum_{t=-T}^{T} \|\boldsymbol{\theta}_t - \boldsymbol{\theta}^\star\|_2^2}{2T+1}\right] \leq \frac{2d}{2\pi\eta^2} \int_{-\pi}^{\pi} \left(|B(\omega)|^2 G^2 \rho^{-1} \gamma_\infty(B)^2 + \sigma_{\mathsf{sgd}}^2\right) \psi(\omega) \, \mathrm{d}\omega$$

where S_{sgd} is the power spectral density of $\widetilde{\boldsymbol{w}}$. In particular, if the density of $\boldsymbol{\theta}_t$ converges to a stationary distribution, the expected value of

$$\lim_{t \to \infty} \mathbb{E}\left[\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^\star\|_2^2\right]$$

under the stationary distribution is bounded as above.

Proof. We assume without loss of generality that $\nabla F(0) = 0$ so that the origin is the global optimum of F (else we can translate the origin to achieve this).

Since $\boldsymbol{g} = \nabla F(\boldsymbol{\theta})$ satisfies

$$\langle \boldsymbol{g} - L\boldsymbol{\theta}, \mu\boldsymbol{\theta} - \boldsymbol{g} \rangle \geq 0 \quad \forall \boldsymbol{\theta}, \boldsymbol{g}$$

Then, we can write down the following family of integral quadratic constraints relating $\boldsymbol{g} = (\ldots, \boldsymbol{g}_0, \boldsymbol{g}_1, \boldsymbol{g}_2, \ldots)$ and $\boldsymbol{\theta} = (\ldots, \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots)$ in terms of their Fourier transforms $\Theta(\omega), G(\omega)$ ([23] equations 27-29):

$$\int_{-\pi}^{\pi} \begin{pmatrix} \Theta(\omega) \\ G(\omega) \end{pmatrix}^* \begin{pmatrix} -\mu L \left(\Lambda(\omega) + \Lambda(\omega)^*\right) \boldsymbol{I}_d & \mu\left(\Lambda(\omega)\right) \boldsymbol{I}_d + L\left(\Lambda(\omega)^*\right) \boldsymbol{I}_d \\ \mu\left(\Lambda^*(\omega)\right) \boldsymbol{I}_d + L\left(\Lambda(\omega)\right) \boldsymbol{I}_d & -\left(\Lambda(\omega) + \Lambda(\omega)^*\right) \boldsymbol{I}_d \end{pmatrix} \begin{pmatrix} \Theta(\omega) \\ G(\omega) \end{pmatrix} \mathrm{d}\omega \geq 0 \tag{83}$$

Noting that from (80), we have that

$$\Theta(\omega)\left(\exp(i\omega) - 1\right) = -\eta\left(G(\omega) + Z(\omega)\right) \implies G(\omega) = \left(\frac{1 - \exp(i\omega)}{\eta}\right) \Theta(\omega) - Z(\omega)$$

where Z denotes the DTFT of $\boldsymbol{\zeta} = \boldsymbol{B}\boldsymbol{w} + \hat{\boldsymbol{w}}$. Plugging this into the above quadratic constraint and multiplying by η^2, we obtain

$$\int_{-\pi}^{\pi} \begin{pmatrix} \Theta(\omega) \\ Z(\omega) \end{pmatrix}^* M_\lambda(\omega) \begin{pmatrix} \Theta(\omega) \\ Z(\omega) \end{pmatrix} \mathrm{d}\omega \geq 0 \tag{84}$$

Since $M_\lambda(\omega) \preceq \begin{pmatrix} -\eta^2 \boldsymbol{I}_d & 0 \\ 0 & \psi(\omega) \boldsymbol{I}_d \end{pmatrix}$ we obtain that

$$\int_{-\pi}^{\pi} \begin{pmatrix} \Theta(\omega) \\ Z(\omega) \end{pmatrix}^* \begin{pmatrix} -\eta^2 \boldsymbol{I}_d & 0 \\ 0 & \psi(\omega) \end{pmatrix} \begin{pmatrix} \Theta(\omega) \\ Z(\omega) \end{pmatrix} \mathrm{d}\omega \geq 0 \implies \frac{\mathbb{E}\left[\int_{-\pi}^{\pi} \|\Theta(\omega)\|^2\right]}{\mathbb{E}\left[\int_{-\pi}^{\pi} \left\|\sqrt{\psi(\omega)} Z(\omega)\right\|^2\right]} \leq 1$$

$$\implies \quad \frac{\lim_{T \to \infty} \mathbb{E}\left[\frac{\sum_{t=-T}^{T} \|\boldsymbol{\theta}_t\|^2}{2T+1}\right]}{\lim_{T \to \infty} \mathbb{E}\left[\frac{\sum_{t=-T}^{T} \left\|\sqrt{\psi}[\boldsymbol{\zeta}](t)\right\|^2}{2T+1}\right]} \leq \frac{1}{\eta^2}$$

where $\sqrt{\boldsymbol{\zeta}}[z]$ denotes the LTI operator with transfer function $\sqrt{\boldsymbol{\zeta}(\omega)}$ applied to the signal $\boldsymbol{\zeta}$.

The denominator of the final line above is the power spectral density of $\sqrt{\kappa}[\boldsymbol{\zeta}]$ (since $\sqrt{\kappa}[\boldsymbol{\zeta}]$ is a wide-sense stationary stochastic process). By the Cauchy-Schwartz inequality for random variables, this is bounded above by

$$2d\left(|B(\omega)|^2\rho^{-1}\gamma_\infty(B)^2+\sigma_{\mathsf{sgd}}^2\right)\psi(\omega)$$

where the first term in brackets is the power spectral density of the Gaussian random process $\boldsymbol{Bw}$ and the second term is an upper bound on the power spectral density of $\hat{\boldsymbol{w}}$. Hence, by Theorem F.2, we have the result. □

Full Proof of Theorem E.1: We are now ready to prove the theorem.

Proof of Theorem E.1. Given the above theorem and smooth convexity parameter L, we know that the asymptotic suboptimality F_∞ is bounded above by

$$\frac{2Ld}{2\pi\eta^2}\int_{-\pi}^{\pi}\left(|B(\omega)|^2\rho^{-1}\gamma_\infty(B)^2G^2+\sigma_{\mathsf{sgd}}^2\right)\psi(\omega)\,\mathrm{d}\omega$$

Now, the constraint (82) can be rewritten as

$$\begin{pmatrix}-\eta^2 & 0\\ 0 & \psi(\omega)\end{pmatrix}- \begin{pmatrix}\eta & 0\\ 1-\exp(i\omega) & -\eta\end{pmatrix}^{*\top}\begin{pmatrix}-\mu L\left(\Lambda(\omega)+\Lambda(\omega)^*\right) & \mu\Lambda(\omega)+L\Lambda(\omega)^*\\ \mu\Lambda^*(\omega)+L\Lambda(\omega) & -\left(\Lambda(\omega)+\Lambda(\omega)^*\right)\end{pmatrix}\begin{pmatrix}\eta & 0\\ 1-\exp(i\omega) & -\eta\end{pmatrix}\succeq 0 \tag{85}$$

since all the matrices involved are Hadamard products of the 2×2 matrices above and the identity matrix.

Thus, for each ω, $\psi(\omega)$ must satisfy a 2×2 PSD constraint which can be rewritten as a Second Order Cone Program (SOCP) constraint. Furthermore, the constraint on λ from theorem E.2 is a linear constraint. Since the projection of a convex set in ψ,λ to ψ is convex, ψ belongs to a convex set. Furthermore, if we take λ such that $\lambda_\tau=0$ for $|\tau|>T_{\max}$ for some $T_{\max}>0$, the constraint on λ can be written as

$$2\lambda_0\geq\sum_{\tau=-T_{\max}}^{T_{\max}}\lambda_t$$

Further, if we discretize ω to a uniform grid on $[-\pi,\pi]$, the constraints (85) can be written as a finite collection of SOCP constraints linking $\psi(\omega)$ and λ.

□

F Technical Definitions and Lemmas

We review a number of relevant technical definitions and lemmas here:

- **Appendix F.1**: Fourier Analysis of Linear Time-Invariant Systems.
- **Appendix F.2**: Stationary covariance of SGD.
- **Appendix F.3**: Concentration of Measure.
- **Appendix F.4**: Review of definitions and useful properties of elliptic integrals.

F.1 Linear Time-Invariant (LTI) Systems

We first review the definition and some useful properties of discrete-time Linear Time-Invariant (LTI) systems. We refer to the textbook [43] for a more detailed description.

Definition F.1. *An input-output system $\boldsymbol{y}_t = \mathcal{A}_t(\boldsymbol{x})$ with an input sequence $\boldsymbol{x} = (\boldsymbol{x}_t)_{t=-\infty}^{\infty}$ in some input space $\mathcal{X}$ and an output sequence $(\boldsymbol{y}_t)_{t=-\infty}^{\infty}$ in an output space $\mathcal{Y}$ is said to be LTI if satisfies two properties:*

- ***Linearity**: For any $\mathcal{X}$-valued sequences $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots$ and scalars $\alpha_1, \alpha_2, \ldots$, we have*

$$\mathcal{A}_t\left(\sum_{j=1}^{\infty} \alpha_j \boldsymbol{x}^{(j)}\right) = \sum_{j=1}^{\infty} \alpha_j \mathcal{A}_t(\boldsymbol{x}^{(j)})\,.$$

- ***Time-Invariance**: For any $t_0 \in \mathbb{Z}$, the sequence $\boldsymbol{x}'$ defined as $\boldsymbol{x}'_t := \boldsymbol{x}_{t-t_0}$ satisfies $\mathcal{A}_t(\boldsymbol{x}') = \mathcal{A}_{t-t_0}(\boldsymbol{x})$.*

LTI systems can be viewed as linear operators defined on the Hilbert space of *signals*

$$\ell_{2e}^d = \left\{(\boldsymbol{x}_t)_{t=-\infty}^{\infty} : \sum_{\tau=-t}^{t} \|\boldsymbol{x}\|^2 < \infty \qquad \forall t \in \mathbb{Z}\right\}$$

with the inner product $\langle \boldsymbol{x}, \boldsymbol{y}\rangle = \sum_{t=-\infty}^{\infty} \boldsymbol{x}_t{}^\top \boldsymbol{y}_t$ with the convention $\ell_{2e} = \ell_{2e}^1$. LTI systems can be described in linear algebraic notation by the action of an infinite Toeplitz matrix on an element of ℓ_{2e}:

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} \implies \boldsymbol{y}_t = \sum_{\tau=-\infty}^{\infty} \boldsymbol{H}_{t,\tau}\boldsymbol{x}_\tau = \sum_{\tau=-\infty}^{\infty} \boldsymbol{h}_{t-\tau}\boldsymbol{x}_\tau = (\boldsymbol{h} \star \boldsymbol{x})_t \quad \forall t \in \mathbb{Z}\,.$$

This property is represented more elegantly in the Fourier domain. Consider the discrete-time Fourier transform (DTFT)

$$\boldsymbol{X}(\omega) = \sum_{t=-\infty}^{\infty} \boldsymbol{x}_t \, \exp(-i\omega t)\,,$$

of $\boldsymbol{x}_t$ and the corresponding DTFTs $\boldsymbol{Y}(\omega), \boldsymbol{G}(\omega)$ of $\boldsymbol{y}, \boldsymbol{h}$ respectively. Then, we have $\boldsymbol{Y}(\omega) = \boldsymbol{G}(\omega)\boldsymbol{X}(\omega)$. Here, $\boldsymbol{h}$ is known as the **impulse response** and $\boldsymbol{G}(\omega)$ is known as the **transfer function**.

An LTI system is said to be asymptotically stable if its output decays to zero for any input sequence that is bounded, ie, for which there exists $T > 0$ such that $x_t = 0 \quad \forall t > T$.

Variance of LTI systems driven by white noise: The Fourier-domain analysis of an LTI system (particularly its transfer function) helps us characterize the covariance of the output $\boldsymbol{y}_t$ as a function of the covariance of the input $\boldsymbol{x}_t$. For simplicity, we assume that $\boldsymbol{x}_t \in \mathbb{R}^d$ and $\boldsymbol{y}_t \in \mathbb{R}^p$ so that $\boldsymbol{H}_t$ and $\boldsymbol{G}(\omega)$ can be identified with matrices in $\mathbb{R}^{p\times d}$.

Theorem F.2. *Consider an asymptotically-stable LTI system with $\mathbb{R}^d$-valued inputs $(\boldsymbol{x}_t)_{t=-\infty}^{\infty}$ and $\mathbb{R}^p$-valued outputs $(\boldsymbol{y}_t)_{-\infty}^{\infty}$ and a transfer function $\boldsymbol{G}(\omega) \in \mathbb{R}^{p\times d}$. Suppose that $\boldsymbol{x}_t$ is a stationary white noise sequence with covariance $\boldsymbol{\Sigma}$, i.e., $\mathbb{E}[\boldsymbol{x}_t] = \boldsymbol{0}$ and $\mathbb{E}[\boldsymbol{x}_t \otimes \boldsymbol{x}_\tau] = \boldsymbol{\Sigma}$ if $t = \tau$ and $\boldsymbol{0}$ otherwise for all t, τ. Then, we have for all $t > -\infty$ that*

$$\mathbb{E}[\boldsymbol{y}_t \otimes \boldsymbol{y}_t] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \boldsymbol{G}(\omega)\,\boldsymbol{\Sigma}\,\boldsymbol{G}(\omega)^* \,\mathrm{d}\omega\,.$$

F.2 Stationary Covariance of Stochastic Gradient Descent for Linear Regression

We now give a result characterizing the stationary covariance of SGD for linear regression [7, 18, 29, 30].

Theorem F.3 (Lemma 5 of [29])**.** *Given a fixed non-random vector $\boldsymbol{\delta}_0 \in \mathbb{R}^d$, consider the recursion*

$$\boldsymbol{\delta}_{t+1} = (\boldsymbol{I} - \eta \boldsymbol{x}_t \otimes \boldsymbol{x}_t)\,\boldsymbol{\delta}_t + \eta\boldsymbol{\zeta}_t\,,$$

for all $t \geq 0$ where

- *$\boldsymbol{x}_t$ are i.i.d. with mean $\boldsymbol{0}$, covariance $\boldsymbol{H}$, and*
- *$\boldsymbol{\zeta}_t$ are i.i.d. with mean $\boldsymbol{0}$, covariance $\mathbb{E}[\boldsymbol{\zeta}_t \otimes \boldsymbol{\zeta}_t] \preceq \nu^2\boldsymbol{H}$.*

Further, if $\mathbb{E}\left[\|\boldsymbol{x}_t\|_2^2\,(\boldsymbol{x}_t \otimes \boldsymbol{x}_t)\right] \preceq R^2\boldsymbol{H}$ and $\eta < 1/R^2$, then we have for all $t \geq 0$.

$$\mathbb{E}[\boldsymbol{\delta}_t \otimes \boldsymbol{\delta}_t] \preceq \frac{\eta\nu^2}{1-\eta R^2}\,\boldsymbol{I}\,.$$

F.3 Concentration of Measure

We recall the definition of sub-Gaussian random variables and list some useful concentration inequalities.

Definition F.4. *A real-valued random variable X is said to be sub-Gaussian with variance proxy σ^2 if for all $\lambda \in \mathbb{R}$, we have*

$$\mathbb{E}[\exp(\lambda(X-\mu))] \leq \exp(\lambda^2\sigma^2/2)\,,$$

where $\mu = \mathbb{E}[X]$. If in addition, the variance of X exactly equals σ^2, it is said to be strictly sub-Gaussian*.*

The cumulants of strict sub-Gaussian random variables are closely related to those of a Gaussian [6, Prop. 3.2].

Property F.5. *If X is strictly sub-Gaussian with mean zero and variance σ^2, we have $\mathbb{E}[X^3] = 0$ and $\mathbb{E}[X^4] \leq 3\sigma^4 = \mathbb{E}[Y^4]$ for $Y \sim \mathcal{N}(0,\sigma^2)$.*

Next, we state the **Hanson-Wright inequality** for the concentration of quadratic forms; see e.g. [46].

Lemma F.6. *Let $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_d)$ be such that each ξ_j is independent and sub-Gaussian with mean zero and variance proxy σ^2. Then, we have for any matrix $\boldsymbol{A} \in \mathbb{R}^{d\times d}$,*

$$\mathbb{P}(\langle \boldsymbol{\xi}, \boldsymbol{A}\boldsymbol{\xi}\rangle - \mathbb{E}[\langle \boldsymbol{\xi}, \boldsymbol{A}\boldsymbol{\xi}\rangle] > t) \leq \exp\left(-c\min\left\{\frac{t^2}{\sigma^4\|\boldsymbol{A}\|_F^2}, \frac{t}{\sigma^2\|\boldsymbol{A}\|_2}\right\}\right)\,,$$

for a universal constant c. Consequently, for any $\rho < 1/3$ and symmetric PSD matrix $\boldsymbol{A}$, we have with probability $1-\rho$ that

$$\langle \boldsymbol{\xi}, \boldsymbol{A}\boldsymbol{\xi}\rangle \leq C\sigma^2\left(\mathsf{Tr}\left[\boldsymbol{A}\right]\sqrt{\log\frac{1}{\rho}} + \|\boldsymbol{A}\|_2\log\frac{1}{\rho}\right) \leq C'\sigma^2\mathsf{Tr}\left[\boldsymbol{A}\right]\log\frac{1}{\rho}\,,$$

for universal constants C, C'.

The second part follows from the first one under the simplifications $\|\boldsymbol{A}\|_2 \leq \|\boldsymbol{A}\|_F \leq \mathsf{Tr}\left[\boldsymbol{A}\right]$ and $\mathbb{E}[\langle \boldsymbol{\xi}, \boldsymbol{A}\boldsymbol{\xi}\rangle] \leq \sigma^2\mathsf{Tr}\left[\boldsymbol{A}\right]$ for $\boldsymbol{A}$ PSD.

Remark F.7. *Explicit values for the constant c in Lemma F.6 (and thus for C, C') are known for the case when $\xi_1, \ldots, \xi_d \sim \mathcal{N}(0,\sigma^2)$: $c \approx 0.1457 \geq 1/8$, $C \leq 8$, $C' \leq 16$ [42].*

F.4 Review of Elliptic Integrals

We recall some definitions and useful properties of elliptic integrals. We refer to [1, §19] and [11] for details.

The three canonical elliptic integral forms are:

(i) The complete elliptic integral of the first kind $K : (0,1) \to [0,\infty)$ is

$$K(k) := \int_0^{\pi/2} \frac{\mathrm{d}\omega}{\sqrt{1-k^2\sin^2(\omega)}}\,. \tag{86}$$

(ii) The complete elliptic integral of the second kind $E : (0,1) \to [0,\infty)$ is

$$E(k) := \int_0^{\pi/2} \sqrt{1-k^2\sin^2(\omega)}\;\mathrm{d}\omega\,. \tag{87}$$

(iii) The complete elliptic integral of the third kind $\Pi : (\mathbb{R}\setminus\{\pm 1\})\times(0,1) \to \mathbb{R}$ is denoted conventionally as $\Pi(\alpha^2, k)$ where α^2 is allowed to take negative values. It is defined as

$$\Pi(\alpha^2, k) := \int_0^{\pi/2} \frac{\mathrm{d}\omega}{(1-\alpha^2\sin^2(\omega))\sqrt{1-k^2\sin^2(\omega)}}\,. \tag{88}$$

The corresponding integrals where $1 - k^2 \sin^2(\omega)$ is replaced with $1 + k^2 \sin^2(\omega)$ can also be expressed using the elliptic integrals [1, Eq. (19.7.2), (19.7.5)].

Property F.8. *For any $m \in (0, 1)$, we have*

$$\int_0^{\pi/2} \frac{\mathrm{d}\omega}{\sqrt{1 + m \sin^2(\omega)}} = \frac{1}{\sqrt{1+m}} \, K\left(\sqrt{\frac{m}{1+m}}\right) . \tag{89}$$

Property F.9. *For any $m \in (0, 1)$ and any $\alpha^2 \in \mathbb{R} \setminus \{\pm 1\}$ such that $\alpha^2 + m \neq 0$, we have*

$$\begin{aligned} &\int_0^{\pi/2} \frac{\mathrm{d}\omega}{(1 - \alpha^2 \sin^2(\omega))\sqrt{1 + m \sin^2(\omega)}} \\ &\quad = \frac{m}{(m+\alpha^2)\sqrt{1+m}} \, K\left(\sqrt{\frac{m}{1+m}}\right) + \frac{\alpha^2}{(m+\alpha^2)\sqrt{1+m}} \Pi\left(\frac{m+\alpha^2}{1+m}, \sqrt{\frac{m}{1+m}}\right) . \end{aligned} \tag{90}$$

The next few properties are about the asymptotics of the elliptic integrals; see e.g. [1, Eq. (19.9.1)] for $K(\cdot)$ and [1, Eq. (19.12.4)] for Π.

Property F.10. *For all $k \in (0, 1)$, we have*

$$\log\left(\frac{4}{\sqrt{1-k^2}}\right) \leq K(k) \leq \left(1 + \frac{1-k^2}{4}\right) \log\left(\frac{4}{\sqrt{1-k^2}}\right) \leq \frac{5}{4} \log\left(\frac{4}{\sqrt{1-k^2}}\right) .$$

Property F.11. *For all $k, \alpha^2 \in (0, 1)$, we have*

$$\Pi(\alpha^2, k) \leq \frac{1}{1-\alpha^2} \log\left(\frac{4}{\sqrt{1-k^2}}\right) \left(1 + O\left(\sqrt{1-k^2}\right)\right) .$$

F.5 Useful Integrals

We list a number of useful definite integrals in this section.

Direct Evaluation: The first one is a cosine integral divided by a quadratic form.[5]

Lemma F.12. *For reals $0 < |b| < a$ and an integer l, we have*

$$\int_{-\pi}^{\pi} \frac{\cos(l\omega)\mathrm{d}\omega}{a^2 + b^2 - 2ab\cos\omega} = \frac{2\pi}{a^2 - b^2} \left(\frac{b}{a}\right)^{|l|} .$$

The next lemma is also about rational cosine functions.[6]

Lemma F.13. *For scalar a, we have*

$$\int_{-\pi}^{\pi} \frac{\mathrm{d}\omega}{1 + a\cos(\omega)} = \begin{cases} \frac{2\pi}{1-a^2}, & \text{if } |a| < 1, \\ +\infty, & \text{if } |a| = 1 . \end{cases}$$

The next one is similar to the previous one.

Lemma F.14. *We have that*

$$\int_{-\pi}^{\pi} \frac{\mathrm{d}\omega}{\sqrt{1 - \cos(\omega)}} = +\infty .$$

Proof. We successively deduce

$$\int_{-\pi}^{\pi} \frac{\mathrm{d}\omega}{\sqrt{1 - \cos(\omega)}} = \frac{1}{\sqrt{2}} \int_{-\pi}^{\pi} \frac{\mathrm{d}\omega}{|\sin(\omega/2)|} = 2\sqrt{2} \int_0^{\pi/2} \frac{\mathrm{d}\omega}{\sin(\omega)} = +\infty ,$$

where we used that $\int \mathrm{d}\omega / \sin(\omega) = -\log|\csc(\omega) + \cot(\omega)| + C$. □

[5] See https://math.stackexchange.com/a/816253.

[6] See https://math.stackexchange.com/a/1235309.

Reductions to Elliptic Integrals: We now list several cosine integrals that can be reduced to elliptic integrals (see Appendix F.4 for their definitions).

Lemma F.15. *For any $a \in (0,1)$, we have*

$$\int_{-\pi}^{\pi} \frac{\mathrm{d}\omega}{|1-a-\exp(i\omega)|} = \frac{4}{2-a} K\left(\frac{\sqrt{1-a}}{1-a/2}\right), \tag{91}$$

where $K(\cdot)$ is the complete elliptic integral of the first kind, cf. (86)*.*

Proof. Using $\cos(\omega) = 1 - 2\sin^2(\omega/2)$ and the substitution $\omega' = \omega/2$, we successively deduce

$$\begin{aligned}
\int_{-\pi}^{\pi} \frac{\mathrm{d}\omega}{|1-a-\exp(i\omega)|} &= 2\int_0^{\pi} \frac{\mathrm{d}\omega}{\sqrt{1+(1-a)^2-2(1-a)\cos(\omega)}} \\
&= 2\int_0^{\pi} \frac{\mathrm{d}\omega}{\sqrt{a^2+4(1-a)\sin^2(\omega/2)}} \\
&= \frac{4}{a}\int_0^{\pi/2} \frac{\mathrm{d}\omega'}{\sqrt{1+4\left(\frac{1-a}{a^2}\right)\sin^2(\omega')}}.
\end{aligned}$$

Applying Property F.8 to reduce this to the standard elliptic integral completes the proof. □

The next lemma handles a more general case. Note that it recovers Lemma F.15 when $a = b$ since $\Pi(0,k) = K(k)$ by definition.

Lemma F.16. *For any $a, b \in (0,1)$, we have*

$$\int_{-\pi}^{\pi} \frac{|1-a-\exp(i\omega)|}{|1-b-\exp(i\omega)|^2}\,\mathrm{d}\omega = \frac{2a^2}{b^2(1-a/2)}\,\Pi\left(\frac{b^2(1-a)-a^2(1-b)}{b^2(1-a/2)^2},\ \frac{\sqrt{1-a}}{1-a/2}\right), \tag{92}$$

where Π is the complete elliptic integral of the third kind, cf. (88)*.*

Proof. We assume that $a \neq b$ to begin and handle the case of $a = b$ by continuity. Denote $h(a,\omega) = \sqrt{1+(1-a)^2-2(1-a)\cos(\omega)}$

$$\begin{aligned}
\int_{-\pi}^{\pi} \frac{|1-a-\exp(i\omega)|}{|1-b-\exp(i\omega)|^2}\,\mathrm{d}\omega &= \int_{-\pi}^{\pi} \frac{|1-a-\exp(i\omega)|^2}{|1-a-\exp(i\omega)|\,|1-b-\exp(i\omega)|^2}\,\mathrm{d}\omega \\
&= \frac{1+(1-a)^2}{h(a,\omega)\,h(b,\omega)^2} - 2(1-a)\frac{\cos(\omega)}{h(a,\omega)\,h(b,\omega)^2}.
\end{aligned}$$

We next add and subtract terms to make the numerator of the second term read $h(b,\omega)^2$ to give

$$\int_{-\pi}^{\pi} \frac{|1-a-\exp(i\omega)|}{|1-b-\exp(i\omega)|^2}\,\mathrm{d}\omega = \int_{-\pi}^{\pi} \frac{1+(1-a)^2-\frac{1-a}{1-b}\left(1+(1-b)^2\right)}{h(a,\omega)\,h(b,\omega)^2}\,\mathrm{d}\omega + \frac{1-a}{1-b}\int_{-\pi}^{\pi} \frac{\mathrm{d}\omega}{h(a,\omega)}. \tag{93}$$

From Lemma F.15, the second term above can be written as

$$\frac{1-a}{1-b}\int_{-\pi}^{\pi} \frac{\mathrm{d}\omega}{h(a,\omega)} = \frac{4(1-a)}{(1-b)(2-a)} K\left(\frac{\sqrt{1-a}}{1-a/2}\right). \tag{94}$$

The first term of (93) can similarly be reduced to the elliptic integral form with $\cos(\omega) = 1 - 2\sin^2(\omega/2)$ and the substitution $\omega' = \omega/2$ as

$$\begin{aligned}
\int_{-\pi}^{\pi} \frac{\mathrm{d}\omega}{h(a,\omega)\,h(b,\omega)^2} &= \frac{2}{ab^2}\int_0^{\pi} \frac{\mathrm{d}\omega}{\sqrt{1+\frac{4(1-a)}{a^2}\sin^2(\omega/2)}\left(1+\frac{4(1-b)}{b^2}\sin^2(\omega/2)\right)} \\
&= \frac{4}{ab^2}\int_0^{\pi/2} \frac{\mathrm{d}\omega'}{\sqrt{1+\frac{4(1-a)}{a^2}\sin^2(\omega')}\left(1+\frac{4(1-b)}{b^2}\sin^2(\omega')\right)}.
\end{aligned}$$

This can be written in terms of elliptic integrals using Property F.9 as

$$\int_0^{\pi/2} \frac{\mathrm{d}\omega'}{\sqrt{1+\frac{4(1-a)}{a^2}\sin^2(\omega')}\left(1+\frac{4(1-b)}{b^2}\sin^2(\omega')\right)} = \frac{a}{2-a}\left(\frac{b^2(1-a)}{b^2(1-a)-a^2(1-b)}\right)K(k) - \frac{a^3(1-b)}{(2-a)(b^2(1-a)-a^2(1-b))}\,\Pi(\alpha^2,k)\,, \tag{95}$$

with $k = \sqrt{1-a}/(1-a/2)$ and

$$\alpha^2 = \frac{b^2(1-a)-a^2(1-b)}{b^2(1-a/2)^2}\,.$$

Plugging in (94) and (95) into (93), we find that the $K(\cdot)$ term cancels out, completing the proof. □

F.6 Other Helper Results

We list a number of other miscellaneous useful results.

Lemma F.17. *For a sequence $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots) \in \ell^2$ and a constant $0 \le c < 1$, we have*

$$\sum_{t=0}^{\infty}\sum_{\tau=0}^{\infty} \beta_t \beta_\tau c^{|t-\tau|} \le \left(\frac{1+c}{1-c}\right)\|\boldsymbol{\beta}\|_2^2\,.$$

Proof. We break the sum into powers of c and use the Cauchy-Schwarz inequality $(*)$ to get

$$\begin{aligned}\sum_{t=0}^{\infty}\sum_{\tau=0}^{\infty} \beta_t \beta_\tau c^{|t-\tau|} &= \|\boldsymbol{\beta}\|_2^2 + 2\sum_{k=1}^{\infty} c^k \left(\sum_{t=0}^{\infty} \beta_t \beta_{t+k}\right) \\ &\overset{(*)}{\le} \|\boldsymbol{\beta}\|_2^2 + 2\sum_{k=1}^{\infty} c^k \|\boldsymbol{\beta}\|_2^2\,.\end{aligned}$$

Summing up the geometric series with a multiplier $0 \le c < 1$ completes the proof. □

Lemma F.18. *Consider a random vector $\boldsymbol{x}$ that satisfies $\mathbb{E}[\boldsymbol{x}] = 0$, $\mathbb{E}[\boldsymbol{x}\otimes\boldsymbol{x}] = \boldsymbol{H} \succeq \mu I$ for some $\mu > 0$ and $\mathbb{E}\left[\|\boldsymbol{x}\|_2^2 \boldsymbol{x}\otimes\boldsymbol{x}\right] \preceq R^2\boldsymbol{H}$. Then, we have for all $\eta \le 1/R^2$ and all PSD matrices $\boldsymbol{M}$ that*

$$\mathsf{Tr}\left[(\boldsymbol{I}-\eta\boldsymbol{x}\otimes\boldsymbol{x})\boldsymbol{M}(\boldsymbol{I}-\eta\boldsymbol{x}\otimes\boldsymbol{x})\right] \le (1-\eta\mu)\mathsf{Tr}\left[\boldsymbol{M}\right]\,.$$

Proof. The left side above (call it "LHS") is bounded by

$$\begin{aligned}\text{LHS} &= \mathsf{Tr}\left[\boldsymbol{M}\right] - 2\eta\mathsf{Tr}\left[\boldsymbol{M}\boldsymbol{M}\right] + \eta^2\mathsf{Tr}\left[\mathbb{E}\left[\|\boldsymbol{x}\|_2^2\boldsymbol{x}\otimes\boldsymbol{x}\right]\boldsymbol{M}\right] \\ &\le \mathsf{Tr}\left[\boldsymbol{M}\right] - 2\eta\mathsf{Tr}\left[\boldsymbol{H}\boldsymbol{M}\right] + \eta^2R^2\mathsf{Tr}\left[\boldsymbol{H}\boldsymbol{M}\right] \\ &\le \mathsf{Tr}\left[\boldsymbol{M}\right] - \eta\mathsf{Tr}\left[\boldsymbol{H}\boldsymbol{M}\right] \\ &\le (1-\eta\mu)\mathsf{Tr}\left[\boldsymbol{M}\right]\,,\end{aligned}$$

where we used (a) $\mathbb{E}\left[\|\boldsymbol{x}\|_2^2\boldsymbol{x}\otimes\boldsymbol{x}\right] \preceq R^2\boldsymbol{H}$, (b) $\eta \le 1/R^2$, and (c) $\boldsymbol{H} \succeq \mu\boldsymbol{I}$. □

Lemma F.19. *For PSD matrices $\boldsymbol{0} \preceq \boldsymbol{A}_1, \ldots, \boldsymbol{A}_k \preceq \boldsymbol{I}$ of shape $d \times d$, we have $|\mathsf{Tr}\left[\boldsymbol{A}_1\cdots\boldsymbol{A}_k\right]| \le d$.*

Proof. Recall the inner product $\langle \boldsymbol{A}, \boldsymbol{B}\rangle = \mathsf{Tr}\left[\boldsymbol{A}\boldsymbol{B}^\top\right]$ on the space of real $d \times d$ matrices. Using Hölder's inequality on the Schatten p-norms, we get

$$|\mathsf{Tr}\left[\boldsymbol{A}_1\ldots\boldsymbol{A}_k\right]| = |\langle \boldsymbol{A}_1, \boldsymbol{A}_k\cdots\boldsymbol{A}_2\rangle| \le \|\boldsymbol{A}_1\|_{S_1}\|\boldsymbol{A}_k\cdots,\boldsymbol{A}_2\|_{S_\infty}\,.$$

DP-FTRL Variant	Citation	Corr. matrix B	Anytime?	Computation Cost	
				Generation	Training (per step)
DP-SGD	[2]	Identity	✓	$O(1)$	$O(1)$
Honaker/TreeAgg	[33]	Lower-Triangular (LT)	✓	$O(1)$	$O(\log T)$
Optimal CC	[21]	Toeplitz & LT	✓	$O(1)$	$O(T)$
ν-DP-FTRL	Ours	Toeplitz & LT	✓	$O(1)$	$O(T)$
FFT	[17]	Toeplitz	-	$O(1)$	$O\left(T\log^2 T\right)$
Full Honaker	[25]	Arbitrary	-	$O(T^2)$	$O(T^2)$
Multi-Epoch (ME)	[17]	Arbitrary	-	$O\left(T^3\right)$	$O\left(T^2\right)$

Table 4: **Variants of DP-FTRL**: the noise correlation matrix $\boldsymbol{B}$ and whether the correlation matrix $\boldsymbol{B}$ can be created/optimized agnostic to the time horizon T (denoted as "**Anytime**"), and the computation cost.

Here, the Schatten 1-norm $\|\cdot\|_{S_1}$ is the ℓ_1 norm of the singular values (i.e. the nuclear norm); this is just the trace for a PSD matrix. Thus,

$$\|\boldsymbol{A}_1\|_{S_1} = \mathsf{Tr}\left[\boldsymbol{A}_1\right] \le \mathsf{Tr}\left[\boldsymbol{I}\right] = 1\,.$$

The $\|\cdot\|_{S_\infty}$ is the ℓ_∞ norm of the singular values, i.e. the operator norm $\|\cdot\|_2$. We get,

$$\|\boldsymbol{A}_k \cdots \boldsymbol{A}_2\|_2 \le \|\boldsymbol{A}_k\|_2 \cdots \|\boldsymbol{A}_2\|_2 \le 1\,.$$

□

Lemma F.20. *For some fixed integer $t \ge 1$ and constants $a > 0$, $\rho \in (0,1)$, define the function*

$$f(\tau) = \tau + \frac{1}{\rho a}\exp(-a\tau)\,\mathbb{1}\left(\tau < t-1\right)\,.$$

For $\hat{\tau} = \min\{t-1, a^{-1}\log(1/\rho)\}$, we have,

$$f(\hat{\tau}) = \min\left\{t-1, \frac{1}{a}(1+\log(1/\rho))\right\} \le \frac{1}{a}(1+\log(1/\rho))\,.$$

Proof. The convex function $\tau \mapsto \tau + \frac{1}{\rho a}\exp(-a\tau)$ is minimized at $\tau_\star = a^{-1}\log(1/\rho) > 0$ with a minimum value of $a^{-1}(1+\log(1/\rho))$. If $t-1 \le \hat{\tau}_\star$, we take $\hat{\tau} = t-1$ and $f(\hat{\tau}) = t-1 \le \hat{\tau} \le a^{-1}(1+\log(1/\rho))$. □

The next lemma is from [45, Lemma 13].

Lemma F.21. *Consider a function $\varphi : [0, \eta_{\max}] \to \mathbb{R}_+$ given by*

$$\varphi(\eta) = A\exp(-\mu\eta T) + B\eta + C\eta^2\log^2\left(\frac{1}{\eta\mu}\right)\,,$$

given some constants $\eta_{\max}, \mu, A, B, C > 0$. If $T \ge (\mu\eta_{\max})^{-1}$, then we have

$$\varphi(\eta_\star) \le A\exp(-\mu\eta_{\max}T) + \frac{3B}{\mu T}\left(1 \vee \log\frac{A\mu T}{B}\right) + \frac{3C}{\mu^2T^2}\left(1 \vee \log\frac{A\mu^2T^2}{C}\right)^2\log^2(T)\,,$$

for some $\eta_\star \le \eta_{\max}$ depending on A, B, C, μ, T.

Lemma F.22. *For $0 < c < 1/4$, we have,*

$$0 < x \le \frac{c}{9\log^2(9/c)} \quad \Longrightarrow \quad x\log^2(1/x) \le c\,.$$

G Empirical Details

We train image-classification models using the CIFAR-10 dataset and language models using the Stack Overflow Next Word Prediction (SONWP) dataset available on `tensorflow-datasets`.

G.1 Image classification

Image classification has long been studied in DP ML. For example, the original DP-SGD work of Abadi et al. [2] focused on this task. We use CIFAR10 which has 50,000 training and 10,000 test examples. We evaluate and compute test accuracies on the entire test set, following the open-sourced code of Kairouz et al. [33]. We reuse the network architecture, dataset processing and initialization strategies presented in Kairouz et al. [33]; in particular, the architecture we use can be found in their Table 2 (b).

Setup and Tuning: We train all mechanisms for 2000 steps using a batch size of 500 and a clip norm of 1. This leads to a ML training dynamics of 20 epochs and 100 steps per epoch. We performed some initial small grid searches which showed nearly ubiquitously that momentum of 0.95 (searched over the grid $0, 0.85, 0.9, 0.95$) and a linear learning rate cooldown $0.05\times$ the initial learning rate over the last 500 steps of training improved model utility for all privacy levels. Thus, we fix these settings for all mechanisms except DP-SGD, for which no momentum performed best. For each mechanism, we then run a tuning grid search for the learning rate on coefficients in $\{1, 2, 5\}$ on powers in [-2, 3], selecting the best mechanism for each privacy level from this interval. Final experiments are repeated 12 times in each setting and show 95% bootstrapped confidence intervals.

Some mechanisms include additional hyper parameters that specify the exact mechanism's structure. For example, ME is specified by both the number of steps n and the max number of participations k. We include such parameters in the mechanism name. For all mechanisms, $n = 2000$.

G.2 Language modeling

Language modeling has been prominently studied in user-level DP contexts, starting with McMahan et al. [41]. DP training is important for real-world applications of language models trained on user data as these models can memorize their training data if appropriate mitigations are not applied [5, 13–15, 27, 37]. Indeed, DP already plays an important role in this application, as evidenced by Google's use of DP for training on-device language models McMahan & Thakurta [40], Xu et al. [50]. StackOverflow Next Word Prediction (SONWP) contains over 10^8 examples non-iid over 342,477 users. The goal of this task is to predict the next word given a sequence of words. We use the same setup as Choquette-Choo et al. [17].

Setup and Tuning: All mechanisms use a clip norm of 1, a server momentum of 0.95, and a client learning rate of 1.0. They also use a server learning rate cooldown over the last 25% rounds. Initial tuning showed these to favorable settings. We train all mechanisms for 2052 steps and report the final evaluation accuracy of the model as reported on a held-out set of $10,000$ examples. We zero out large updates of $\ell_\infty \geq 100$. We use the tuned server learning rates from Choquette-Choo et al. [17] for all existing mechanisms. For our new ν-DP-FTRL mechanisms, we do not perform extensive tuning due to computational costs and instead choose a near-optimal value of the server learning rate based on its correlation with normalized loss of the mechanism (i.e., normalized noise added due to $\boldsymbol{B}$) and the results of Choquette-Choo et al. [16, Figure 11].