OPTIMIZATION RISK BOUNDS FOR KOLMOGOROV–ARNOLD NETWORKS TRAINED BY DP-SGD WITH CORRELATED NOISE

Anonymous authors
Paper under double-blind review

ABSTRACT

The theoretical understanding of differentially private stochastic gradient descent (DP-SGD) with temporally correlated noise remains limited, particularly for non-convex neural network training. As a first step, we study two-layer Kolmogorov–Arnold Networks (KANs), a recently introduced architecture with learnable spline-based edge functions. We establish the first optimization risk bounds for clipped mini-batch DP-SGD with correlated noise in this setting, with explicit dependence on temporal correlation, clipping, mini-batch sampling, and network width. Existing arguments fail for three reasons: temporal dependence breaks the conditional-centering step; projection obstructs the cross-iteration cancellation of correlated perturbations; and active clipping breaks the empirical-gradient structure. We address these issues through shifted and auxiliary dynamics, a weighted empirical loss, and a high-probability localization argument. Our bound shows that temporal correlation reduces the leading noise terms, the clipping threshold enters essentially through an effective step size, and private training admits an explicit network width range. Experiments on synthetic data and MNIST support the predicted optimization effect of temporal correlation. As an application, we derive population risk guarantees via algorithmic stability. Our framework recovers non-private mini-batch SGD, independent-noise DP-SGD, and their full-batch counterparts as special cases.

1 INTRODUCTION

Differentially private stochastic gradient descent (DP-SGD) (Song et al., 2013) is the standard approach to training machine learning models with formal privacy guarantees (Dwork, 2006). Standard DP-SGD adds fresh Gaussian noise at every iteration, whose accumulation can substantially degrade optimization and predictive performance. Recent correlated noise mechanisms instead introduce temporal correlations across perturbations, allowing consecutive noise terms to partially cancel and improve privacy–utility tradeoffs. Recent studies have demonstrated substantial empirical gains over independent noise training (Choquette-Choo et al., 2024a; 2023; Kalinin et al., 2026).

Despite this empirical progress, optimization theory for correlated noise private training remains limited, particularly for neural networks. Koloskova et al. (2023) provide stationarity guarantees for smooth non-convex optimization with linearly correlated noise, but do not address optimization risk for clipped mini-batch neural network training with local network geometry. Optimization risk guarantees in this regime therefore remain largely unexplored.

We investigate this problem through Kolmogorov-Arnold Networks (KANs) (Liu et al., 2025b), a structured alternative to multilayer perceptrons that uses learnable univariate functions on edges. KANs have shown promising performance in scientific applications (Cherednichenko & Poptsova, 2025; Li et al., 2025a; Patra et al., 2025; Shukla et al., 2024; Wang et al., 2025d; Vaca-Rubio et al., 2024). The closest KAN theory, Wang et al. (2026), analyzes full-batch GD and its independent noise DP counterpart. Extending this setting to clipped mini-batch training with temporally correlated noise introduces new analytical difficulties from temporal dependence, localization, and clipping. Our analysis provides a first step toward understanding correlated noise DP training for non-convex neural networks more broadly.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

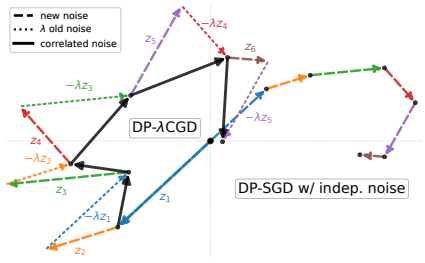


Figure 1: DP-λCGD subtracts a λ-fraction of the previous noise, inducing partial cancellation across consecutive iterations. The figure is reproduced from Kalinin et al. (2026) with the authors' permission.

Concretely, for two-layer KANs, we consider clipped mini-batch DP-λCGD (Kalinin et al., 2026) with correlated perturbations $\xi_t = \kappa(Z_t - \lambda Z_{t-1})$, where Z_t is standard Gaussian noise, $\kappa \geq 0$ is the noise multiplier, and $\lambda = 0$ recovers standard DP-SGD. Our main focus is optimization risk: we characterize how temporal correlation, clipping, mini-batch sampling, and network width enter the optimization bound, and use this result as the input to a stability-based population risk analysis.

This analysis poses three main technical challenges: *(i)* temporal dependence breaks the conditional-centering arguments underlying standard stochastic optimization analyses; *(ii)* the projection used for localization obstructs the cross-iteration cancellation structure of correlated noise (Figure 1); and *(iii)* active clipping makes the averaged update no longer the gradient of the empirical loss, preventing direct application of the local comparator argument. Consequently, the analysis does not follow from a direct combination of the full-batch KAN argument of Wang et al. (2026) and existing correlated noise optimization analyses such as Koloskova et al. (2023).

Our main contributions are summarized as follows.

- We provide the *first* optimization risk analysis of clipped mini-batch DP-SGD with temporally correlated noise for non-convex neural networks, taking two-layer KANs as a concrete setting. The analysis combines shifted and auxiliary dynamics with a high-probability localization bootstrap to exploit cross-iteration cancellation while certifying projection inactivity. Active clipping is handled through a weighted empirical loss with frozen clipping coefficients, which restores the gradient structure needed for the local comparator argument.
- Our optimization bound explicitly characterizes how the *clipping threshold*, *temporal correlation*, *mini-batch sampling*, and *network width* affect optimization risk. In the active clipping regime, the clipping threshold enters essentially through an effective step size, while temporal correlation reduces the leading noise terms through explicit correlation dependent factors. The result also identifies an admissible width range for private KAN training. Experiments on synthetic data and MNIST at fixed noise scales support the predicted optimization effect of temporal correlation.
- As an application, we combine the optimization bound with algorithmic stability to obtain population risk guarantees. We further study how privacy calibration interacts with the optimization effect of temporal correlation under representative privacy regimes.
- Our framework recovers non-private mini-batch SGD, independent noise DP-SGD, and the corresponding full-batch GD and DP-GD regimes as special cases. In the full-batch setting, our bounds match the known dependence on n, d, and ϵ in Wang et al. (2026).

2 Related Work

KAN theory has focused on approximation, expressiveness, and optimization (Eshtehardian et al., 2026; Gao & Tan, 2025; Li et al., 2025b; Liu et al., 2025a; Wang et al., 2025c). The closest prior work Wang et al. (2026) studies optimization and population risk for two-layer KANs under full-batch GD and independent-noise DP-GD. Our setting adds mini-batch stochasticity, clipping, and temporally correlated noise, while recovering these regimes as special cases.

Theory of private neural network (NN) training. Recent work studies private NN training with independent noise Ding et al. (2025); Shi et al. (2026); Wang et al. (2025a); Xu & Chen (2026); Zhang et al. (2026). Wang et al. (2025a) analyzes DP-GD for three-layer MLPs in regression, while Shi et al. (2026) studies DP-GD for two-layer CNNs. Ding et al. (2025); Xu & Chen (2026); Zhang et al. (2026) study DP-SGD for NNs from a feature-learning perspective, including noisy feature dynamics, fairness and robustness degradation, and long-tailed memorization. These works do not cover KANs or temporally correlated noise mechanisms.

Theory of correlated noise DP. Correlated noise mechanisms have been analyzed from several perspectives Denisov et al. (2022); Koloskova et al. (2023); Choquette-Choo et al. (2024a). The work

Koloskova et al. (2023) studies GD with linearly correlated noise and gives stationarity guarantees for smooth non-convex objectives, while Choquette-Choo et al. (2024a) establishes utility separations between correlated and independent noise for private convex learning. Related work develops privacy accounting and amplification for subsampled or structured matrix mechanisms Choquette-Choo et al. (2023; 2024b;c); Schuchardt & Kalinin (2026). These works establish privacy guarantees or optimization results in different settings, but do not address optimization risk for clipped mini-batch DP-SGD in KANs, where the analysis must additionally control local curvature and clipping.

A broader overview of related work, including neural network theory and privacy amplification by subsampling, is provided in Appendix A.

3 Problem Setting

We now introduce the learning problem, the two-layer KAN architecture, the mini-batch DP-SGD algorithm with correlated noise, and the assumptions underlying our analysis.

Notation and learning problem. Let $\mathcal{P}$ be a probability distribution on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subseteq \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 \leq 1\}$ and $\mathcal{Y} = \{-1, +1\}$. For a positive integer q, let $[q] = \{1, \ldots, q\}$. We use $\|\cdot\|_2$ for the Euclidean norm, and $\langle \cdot, \cdot \rangle$ for the inner product. Given a training dataset $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ drawn i.i.d. from $\mathcal{P}$, we measure the quality of a classifier $f : \mathcal{X} \to \mathbb{R}$ by the population and empirical risks $\mathcal{L}(f) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}}[\ell(yf(\mathbf{x}))]$ and $\mathcal{L}_S(f) = \frac{1}{n}\sum_{i=1}^n \ell(y_i f(\mathbf{x}_i))$, respectively, where $\ell(z) = \log(1 + \exp(-z))$ is the logistic loss.

3.1 Architecture: Two-layer KANs with B-spline Bases

Let m be the hidden width. Following the spline-based two-layer KAN formulation studied in Gao & Tan (2025); Wang et al. (2026), we consider a model with B-spline basis $\{b_k\}_{k=1}^p$ and a fixed activation function $\sigma : \mathbb{R} \to \mathbb{R}$. For an input $\mathbf{x} = (x_1, \ldots, x_d)$, the network is defined as

$$f_{\mathbf{W}}(\mathbf{x}) = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} \sum_{k=1}^{p} c_{j,k}\, b_k(x_{1,j}) \quad \text{with} \quad x_{1,j} = \sigma\Big(\frac{1}{\sqrt{d}} \sum_{i=1}^{d} \sum_{k=1}^{p} w_{i,j,k}\, b_k(x_i)\Big),$$

where $\mathbf{W} = \{w_{i,j,k}\}_{i\in[d],\, j\in[m],\, k\in[p]} \in \mathbb{R}^{mdp}$ denotes the trainable first-layer spline coefficients, $\mathbf{c} = \{c_{j,k}\}_{j\in[m],\, k\in[p]} \in \mathbb{R}^{mp}$ denotes the second-layer spline coefficients, and $x_{1,j}$ is the output of the j-th hidden unit. Following the first-layer training formulation of Eshtehardian et al. (2026), we draw $\mathbf{c}$ once at initialization from $\mathcal{N}(\mathbf{0}, \mathbf{I}_{mp})$ and keep it fixed throughout training, while optimizing only $\mathbf{W}$. We further discuss this modeling choice in Section 6. For notational simplification, we write $\mathcal{L}_S(\mathbf{W}) = \mathcal{L}_S(f_{\mathbf{W}})$ and $\mathcal{L}(\mathbf{W}) = \mathcal{L}(f_{\mathbf{W}})$.

3.2 Algorithm: Mini-batch DP-SGD with Correlated Noise

To enable training KANs on sensitive datasets, we study a differentially private variant of mini-batch SGD. We call two datasets neighboring if they differ in the contribution of one data record.

Definition 1 (Differential privacy (Dwork et al., 2006))**.** A randomized algorithm $\mathcal{A}$ satisfies (ϵ, δ)-DP if, for any two neighboring datasets S and S' and any measurable set E in the output space of $\mathcal{A}$, it holds that $\mathbb{P}(\mathcal{A}(S) \in E) \leq e^{\epsilon}\mathbb{P}(\mathcal{A}(S') \in E) + \delta$. We say $\mathcal{A}$ satisfies ϵ-DP if $\delta = 0$.

The algorithm we analyze is mini-batch DP-SGD with correlated noise. It also subsumes independent noise DP-SGD and non-private mini-batch SGD as special cases. We use DP-λCGD (Kalinin et al., 2026), a simple correlated noise mechanism, as the noise model. After initializing $\mathbf{W}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{mdp})$, at each iteration t we draw a mini-batch $\mathcal{B}_t \subseteq [n]$ of size B uniformly without replacement and form $g_{t,i} = \nabla\ell(y_i f_{\mathbf{W}_{t-1}}(\mathbf{x}_i))$, which are clipped at threshold C_{clip} and averaged: $\tilde{g}_{t,i} = g_{t,i} \min\big\{1, \frac{C_{\text{clip}}}{\|g_{t,i}\|_2}\big\}$ and $v_t = \frac{1}{B}\sum_{i\in\mathcal{B}_t} \tilde{g}_{t,i}$. The noise takes the correlated form $\xi_t = \kappa(Z_t - \lambda Z_{t-1})$ with $Z_0 = \mathbf{0}$ and $\{Z_t\}_{t\geq 1} \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}_{mdp})$, where $\kappa \geq 0$ is the noise multiplier and $\lambda \in [0, 1)$ controls the strength of temporal correlation. The iterate is then updated by a projected step onto $\mathcal{K} = \mathcal{B}(\mathbf{W}_0, R_*)$: $\mathbf{W}_t = \Pi_{\mathcal{K}}(\mathbf{W}_{t-1} - \eta\,\hat{v}_t)$ with $\hat{v}_t = v_t + \frac{C_{\text{clip}}}{B}\xi_t$. The full procedure is given in Algorithm 1. We empirically evaluate it on two-layer KANs in Section 4, with additional fixed-privacy predictive results provided in Appendix H.

Algorithm 1 Mini-batch DP-SGD with Correlated Gaussian Noise

Require: Dataset $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$; total iterations T; step size $\eta > 0$; batch size B; clip threshold $C_{\text{clip}} > 0$; correlation $\lambda \in [0, 1)$; noise multiplier $\kappa \geq 0$; localization radius $R_* > 0$.

Initialize $\mathbf{W}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{mdp})$ and $\mathbf{c} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{mp})$; keep $\mathbf{c}$ fixed; set $Z_0 = \mathbf{0} \in \mathbb{R}^{mdp}$ and $\mathcal{K} = \mathcal{B}(\mathbf{W}_0, R_*)$.
for $t = 1, 2, \ldots, T$ **do**
  Sample a mini-batch $\mathcal{B}_t \subseteq [n]$ of size B uniformly without replacement
  for each $i \in \mathcal{B}_t$ **do**
    $\tilde{g}_{t,i} \leftarrow g_{t,i} \cdot \min\left\{1, \frac{C_{\text{clip}}}{\|g_{t,i}\|_2}\right\}$ with $g_{t,i} = \nabla \ell(y_i f_{\mathbf{W}_{t-1}}(\mathbf{x}_i))$ ▷ per-example clipping
  end for
  $v_t \leftarrow \frac{1}{B} \sum_{i \in \mathcal{B}_t} \tilde{g}_{t,i}$ ▷ clipped mini-batch gradient
  Sample fresh noise $Z_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{mdp})$
  $\xi_t \leftarrow \kappa\,(Z_t - \lambda Z_{t-1})$ ▷ correlated Gaussian noise
  $\hat{v}_t \leftarrow v_t + \frac{C_{\text{clip}}}{B}\,\xi_t$
  $\mathbf{W}_t \leftarrow \Pi_{\mathcal{K}}(\mathbf{W}_{t-1} - \eta\,\hat{v}_t)$
end for
return $\{\mathbf{W}_t\}_{t=0}^T$.

Special cases. This formulation recovers the standard baselines as special cases. Setting $\lambda = 0$ gives $\xi_t = \kappa Z_t$, recovering *independent-noise DP-SGD*. Setting $\kappa = 0$ removes the noise and yields *non-private mini-batch SGD*; taking in addition $C_{\text{clip}} \to \infty$ and $R_* \to \infty$ removes the clipping and projection, recovering plain *mini-batch SGD*.

We impose two standard assumptions used in prior KAN analyses (Gao & Tan, 2025; Taheri et al., 2025; Wang et al., 2026). The first bounds the activation and B-spline basis. It is satisfied, e.g., by cubic (or higher-degree) B-splines together with sigmoid or hyperbolic tangent activations. Let $\mathcal{I}_b$ be an interval containing $[-1, 1] \cup \text{range}(\sigma)$.

Assumption 1. Assume σ satisfies $|\sigma(u)| \leq B_\sigma$, $|\sigma'(u)| \leq B'_\sigma$, and $|\sigma''(u)| \leq B''_\sigma$ for all $u \in \mathbb{R}$. Further, assume $\{b_k\}_{k=1}^p$ satisfy $|b_k(v)| \leq B_b$, $|b'_k(v)| \leq B'_b$, and $|b''_k(v)| \leq B''_b$ for all $v \in \mathcal{I}_b$.

The second is a margin-style separability condition on the NTK features at initialization (Lei et al., 2026; Taheri et al., 2025; Wang et al., 2026). It is weaker than the NTK Gram-matrix positive-definiteness commonly used in the literature (Arora et al., 2019; Gao & Tan, 2025; Nitanda et al., 2019), and was empirically validated for KANs in Wang et al. (2026).

Assumption 2 (NTK separability)**.** There exists $\gamma \in (0, 1]$ and $\mathbf{u} \in \mathbb{R}^{mdp}$ with $\|\mathbf{u}\|_2 = 1$ such that $y_i \langle \nabla f_{\mathbf{W}_0}(\mathbf{x}_i), \mathbf{u} \rangle \geq \gamma$ for all $i \in [n]$.

4 Optimization Risk Bound for DP-SGD with Correlated Noise

This section develops our optimization risk bound for DP-SGD with correlated noise, which forms the technical core of the paper. We first identify the main analytical challenges, then present and interpret the optimization theorem, empirically illustrate its correlation-dependent behavior, and finally give a proof sketch showing how the main difficulties are resolved.

4.1 Technical Challenges

To control the optimization risk $\frac{1}{T}\sum_{t=0}^{T-1} \mathcal{L}_S(\mathbf{W}_t)$, a standard approach is to use a comparator argument based on the local curvature of the empirical loss. In particular, for a suitable comparator $\mathbf{W}^* \in \mathcal{K}$, one expects an inequality of the form

$$\mathcal{L}_S(\mathbf{W}_{t-1}) - \mathcal{L}_S(\mathbf{W}^*) \lesssim \left\langle \bar{v}_t, \mathbf{W}_{t-1} - \mathbf{W}^* \right\rangle + \text{Err}_t,$$

where $\bar{v}_t = \frac{1}{n}\sum_{i=1}^n \tilde{g}_{t,i}$ is the full-batch clipped gradient used only in the analysis, and Err_t contains the local curvature error terms. Thus, bounding the optimization risk reduces to controlling the accumulated first-order term $\sum_t \langle \bar{v}_t, \mathbf{W}_{t-1} - \mathbf{W}^* \rangle$. This reduction leads to three distinct difficulties: the correlated perturbations break the conditional-centering argument used in the independent noise

analysis; the curvature-based comparator inequality needed above is only valid locally for KANs; and under active clipping, $\bar{v}_t$ is no longer the gradient of the empirical loss.

A natural first attempt is to follow the independent noise proof, which applies the projected one-step recursion and uses conditional centering to control the perturbation terms. To see the key step, substitute the decomposition of the noisy averaged clipped gradient $\hat{v}_t$ into the first order term:

$$\langle \hat{v}_t, \mathbf{W}_{t-1} - \mathbf{W}^* \rangle = \langle \bar{v}_t, \mathbf{W}_{t-1} - \mathbf{W}^* \rangle + \langle v_t - \bar{v}_t, \mathbf{W}_{t-1} - \mathbf{W}^* \rangle + \frac{C_{\text{clip}}}{B} \langle \xi_t, \mathbf{W}_{t-1} - \mathbf{W}^* \rangle.$$

In the independent noise case, the mini-batch fluctuation $v_t - \bar{v}_t$ and the Gaussian perturbation $\xi_t = \kappa Z_t$ are both conditionally centered given the past. Hence, after taking conditional expectations in the one-step recursion, the two perturbation terms vanish and only the deterministic descent term remains, while the perturbations contribute only through higher order variance terms. For correlated noise, however, this cancellation fails. Since $\xi_t = \kappa(Z_t - \lambda Z_{t-1})$, we have $\mathbb{E}[\xi_t \mid \mathcal{F}_{t-1}] = -\kappa\lambda Z_{t-1}$ with $\mathcal{F}_t$ the sigma-algebra containing all randomness revealed up to the end of iteration t . Hence, ignoring the scaling factor $\frac{C_{\text{clip}}}{B}$, the noise component in the first-order term satisfies

$$\mathbb{E}\big[\langle \xi_t, \mathbf{W}_{t-1} - \mathbf{W}^* \rangle \,|\, \mathcal{F}_{t-1}\big] = -\kappa\lambda \langle Z_{t-1}, \mathbf{W}_{t-1} - \mathbf{W}^* \rangle.$$

Thus, instead of vanishing after conditioning, the perturbation leaves a first order drift term. This term cannot be removed by conditional centering, because $\mathbf{W}_{t-1}$ already depends on Z_{t-1}. Moreover, the projection prevents us from directly exploiting the cross-iteration cancellation structure of the correlated perturbations. Hence the standard projected one-step argument no longer applies.

The second difficulty comes from the local KAN geometry. The empirical loss is non-convex in the first-layer spline coefficients, and the comparator inequality that converts the first order term into empirical loss is valid only when the iterate and the comparator stay in a localized region around the initialization. Therefore, even after the correlated noise drift is handled through the shifted dynamics, the proof must still certify that the auxiliary trajectory remains in the region where the local KAN curvature and self-boundedness estimates can be applied.

The third difficulty comes from active gradient clipping. When clipping is inactive, $\bar{v}_t$ coincides with the gradient of the empirical loss, so the local comparator argument can be applied directly. Under active clipping, however, the coefficients multiplying the per-example gradients depend on the current iterate, and $\bar{v}_t$ is no longer the gradient of $\mathcal{L}_S$. Consequently, the comparator inequality above cannot be applied directly to the empirical loss, and a suitable gradient structure must be recovered while explicitly accounting for the clipping threshold.

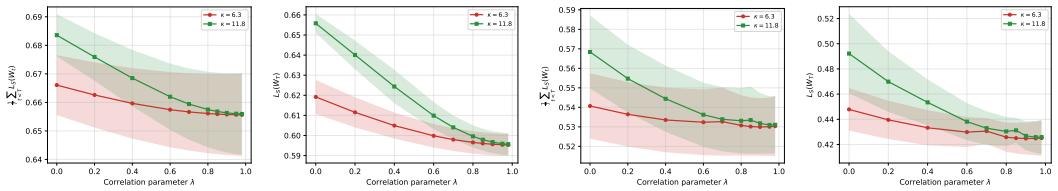
4.2 Main Theorem

We now state our main optimization theorem, which addresses the difficulties described above. The theorem characterizes how mini-batch sampling, clipping, temporal correlation, and network width affect the optimization risk, and identifies an admissible width range. Let $C_{\sigma,b} > 0$ depend only on σ, b, and let $G_\delta = B'_\sigma B_b B'_b p\big(4\sqrt{p} + 2\sqrt{\log(2/\delta)/m}\big)$ be a uniform per-example gradient bound holding with probability at least $1-\delta$. Define $\rho_{\text{clip}} = \min\big\{1, \frac{C_{\text{clip}}}{G_\delta}\big\}$ and $c_{\text{eff}} = \frac{C_{\text{clip}}\kappa}{B\rho_{\text{clip}}}$. Moreover, $\widetilde{\mathcal{O}}$ and $\widetilde{\Omega}$ suppress polylogarithmic factors in $m, n, T, \delta^{-1}, \rho_{\text{clip}}^{-1}$.

Theorem 1. *Let $\delta \in (0,1)$. Suppose Assumptions 1 and 2 hold. Let $\{\mathbf{W}_t\}_{t=0}^{T-1}$ be generated by Algorithm 1 with $\eta_{\text{eff}} \le \frac{1}{12 C_{\sigma,b} p^3}$ with $\eta_{\text{eff}} = \rho_{\text{clip}}\eta$ and $C_{\text{clip}} > 0$. Assume $\eta_{\text{eff}}\gamma\sqrt{T}\big(\frac{1}{\sqrt{B}} + (1-\lambda)c_{\text{eff}}\big) + \eta_{\text{eff}}^2\gamma^2\big(\frac{T}{B}+1\big) \lesssim 1$ and $\widetilde{\Omega}(\gamma^{-4}) \le m \le \widetilde{\mathcal{O}}\big((\eta_{\text{eff}}^2 c_{\text{eff}}^2 \gamma^2 d)^{-1} \min\big\{1, \big(T((1-\lambda)^2 + \lambda^2\eta_{\text{eff}})\big)^{-1}\big\}\big)$. With probability at least $1-\delta$ over the initialization and the algorithmic randomness,*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathcal{L}_S(\mathbf{W}_t) \lesssim \text{polylog}\big(\frac{nT}{\delta\rho_{\text{clip}}}\big) \cdot \Big[\frac{1}{\gamma^2\eta_{\text{eff}}T} + \frac{1}{\gamma\sqrt{BT}} + \eta_{\text{eff}}\big(\frac{1}{B} + \frac{1}{T}\big) + \frac{(1-\lambda)c_{\text{eff}}}{\gamma\sqrt{T}} + \big((1-\lambda)^2 + \lambda^2\eta_{\text{eff}}\big)\eta_{\text{eff}} c_{\text{eff}}^2 md \Big] =: A_{\text{corr}}.$$

Interpretation. The structure of A_{corr} separates the different sources of error. The first three terms arise from optimization and mini-batch sampling, whereas the last two capture the effect of the



(a) Synthetic logistic data. (b) MNIST.

Figure 2: Optimization effect of temporal correlation at fixed noise multipliers. For each dataset, the left panel reports the trajectory-averaged empirical loss $\frac{1}{T}\sum_{t=0}^{T-1}\mathcal{L}_S(\mathbf{W}_t)$, while the right panel reports the final iterate empirical loss $\mathcal{L}_S(\mathbf{W}_T)$. We hold the noise multiplier κ fixed and vary only the correlation parameter λ, thereby isolating the optimization-side effect of temporal correlation. Curves show the mean over five random seeds, with shaded regions indicating one standard deviation.

correlated Gaussian perturbations. In particular, temporal correlation enters these noise-dependent terms through the factors $1-\lambda$ and $(1-\lambda)^2+\lambda^2\eta_{\text{eff}}$, respectively, making the effect of temporal correlation explicit in the bound.

The admissible width range also reflects a tradeoff: m must be large enough for the local KAN curvature argument to apply, but small enough that the private noise does not break localization.

Remark 1 (Effect of the clipping threshold C_{clip})**.** The dependence of the optimization bound on C_{clip} is captured by the effective step size $\eta_{\text{eff}}=\rho_{\text{clip}}\eta$ and the effective noise scale $c_{\text{eff}}=C_{\text{clip}}\kappa/(B\rho_{\text{clip}})$. When $C_{\text{clip}}\le G_\delta$, we have $\rho_{\text{clip}}=C_{\text{clip}}/G_\delta$ and $c_{\text{eff}}=G_\delta\kappa/B$, so c_{eff} is independent of the clipping threshold. Consequently, apart from logarithmic factors in ρ_{clip}^{-1}, the clipping threshold enters Theorem 1 only through η_{eff}. In particular, stronger clipping can be compensated by increasing the nominal step size η to keep η_{eff} fixed, in which case the optimization guarantee is unchanged up to logarithmic factors. When $C_{\text{clip}}\ge G_\delta$, we have $\rho_{\text{clip}}=1$ and $\eta_{\text{eff}}=\eta$, while $c_{\text{eff}}=C_{\text{clip}}\kappa/B$. The clipping operator is then inactive on the high-probability event underlying G_δ, and increasing C_{clip} further provides no optimization benefit while enlarging the noise-dependent terms and tightening the admissible upper bound on the width. Thus, G_δ naturally separates the active clipping regime from the regime in which increasing the clipping threshold only increases the injected noise.

Remark 2 (Role of λ)**.** Theorem 1 shows that, for fixed κ and C_{clip}, the linear noise-fluctuation term scales proportionally to $1-\lambda$, while the quadratic Gaussian term is governed by $(1-\lambda)^2+\lambda^2\eta_{\text{eff}}$. Thus, for small η_{eff}, increasing λ substantially reduces the effect of the correlated perturbations, with the quadratic factor minimized at $\lambda=1/(1+\eta_{\text{eff}})$, which is close to 1. This describes the optimization effect of temporal correlation for a fixed noise multiplier κ. Under DP, however, the required value of κ itself depends on λ, so the resulting privacy–optimization tradeoff depends on the privacy calibration. We illustrate this interaction in Section 5 under two representative settings.

Empirical Illustration. To isolate the optimization effect of temporal correlation from privacy calibration, we fix the noise multiplier κ and vary only the correlation parameter λ. We consider synthetic logistic data generated following the setup of Wang et al. (2026) and MNIST 3 vs. 8, using a two-layer KAN with $m=32$ and $p=8$ cubic B-splines. We set $n=5{,}000$, $B=125$, $T=400$, $C_{\text{clip}}=0.5$, and $\eta=1$, and consider two fixed noise multipliers, $\kappa=6.3$ and $\kappa=11.8$. Thus, these experiments focus on the optimization-side effect of temporal correlation rather than comparisons under a common privacy budget.

Figure 2 reports both the trajectory-averaged empirical loss $\frac{1}{T}\sum_{t=0}^{T-1}\mathcal{L}_S(\mathbf{W}_t)$, which is the quantity controlled by Theorem 1, and the empirical loss of the final iterate. On both datasets and for both noise multipliers, the averaged training loss decreases consistently as λ increases and levels off under strong correlation. The final-iterate loss exhibits the same qualitative behavior. These results are consistent with the reduction of the correlation-dependent noise terms identified in Theorem 1. Additional experimental details and further results are provided in Appendix H.

4.3 Proof Sketch of Theorem 1

We outline the main steps and defer the full proof to Appendix C. The proof uses an auxiliary unprojected trajectory and a shifted iterate to expose the cross-iteration cancellation of correlated

noise, and then combines a localization bootstrap with a weighted empirical loss to make this cancellation compatible with the local KAN geometry and active clipping. The proof proceeds in six steps.

Proof Sketch. STEP 1: REDUCE TO AN AUXILIARY UNPROJECTED TRAJECTORY. The correlated perturbations break conditional centering, while the projection obstructs their cross-iteration cancellation. We therefore introduce an auxiliary unprojected trajectory $\{\widetilde{\mathbf{W}}_t\}_{t=0}^T$ driven by the same mini-batches and Gaussian variables:

$$\widetilde{\mathbf{W}}_t = \widetilde{\mathbf{W}}_{t-1} - \eta\big(\bar{g}_t + \widetilde{\Delta}_t + c_{\mathrm{priv}}(Z_t - \lambda Z_{t-1})\big), \qquad c_{\mathrm{priv}} = C_{\mathrm{clip}}\kappa/B,$$

where $\bar{g}_t$ is the full-batch average of the clipped gradients and $\widetilde{\Delta}_t$ is the centered mini-batch fluctuation. We first analyze this auxiliary process and later show that it remains inside the projection region, so that it coincides with the original iterates.

STEP 2: INTRODUCE THE SHIFTED ITERATE. Applying a comparator recursion directly to the auxiliary trajectory still fails because the correlated perturbation is not conditionally centered, leaving a first order drift term.

To handle this obstruction, we define the shifted iterate $\mathbf{U}_t = \widetilde{\mathbf{W}}_t + \eta c_{\mathrm{priv}} Z_t$. Then the auxiliary update admits the exact reformulation $\mathbf{U}_t = \mathbf{U}_{t-1} - \eta\big(\bar{g}_t + \widetilde{\Delta}_t + (1-\lambda)c_{\mathrm{priv}} Z_{t-1}\big)$.

This identity is the key cancellation step. The leading correlated perturbation is reduced to a residual term proportional to $1-\lambda$, which is the source of the improved noise dependence in the final bound.

STEP 3: RECOVER A LOCAL DESCENT INEQUALITY UNDER CLIPPING. Applying the squared-distance recursion to $\|\mathbf{U}_t - \mathbf{W}^*\|_2^2$ produces a descent term involving $\langle \bar{g}_t, \widetilde{\mathbf{W}}_{t-1} - \mathbf{W}^*\rangle$ together with mini-batch and Gaussian fluctuation terms. Under active clipping, $\bar{g}_t$ is not the gradient of the empirical loss. On the initialization event, the per-example gradient norm is uniformly bounded by G_δ, and we freeze the clipping coefficients at $\widetilde{\mathbf{W}}_{t-1}$ to define

$$\mathcal{L}_{S,t}^{\alpha}(\mathbf{W}) = \frac{1}{n}\sum_{i=1}^{n} \alpha_{t,i}\ell(y_i f_{\mathbf{W}}(\mathbf{x}_i)), \qquad \alpha_{t,i} \in [\rho_{\mathrm{clip}}, 1].$$

Then $\bar{g}_t = \nabla\mathcal{L}_{S,t}^{\alpha}(\widetilde{\mathbf{W}}_{t-1})$ and $\rho_{\mathrm{clip}}\mathcal{L}_S \le \mathcal{L}_{S,t}^{\alpha} \le \mathcal{L}_S$. We stop the recursion when the shifted trajectory leaves the localization region; before this stopping time, the local KAN comparator inequality and self-bounding estimate yield descent in $\mathcal{L}_{S,t}^{\alpha}$ while controlling $\|\bar{g}_t\|_2^2$.

STEP 4: CONTROL THE STOPPED STOCHASTIC TERMS. The Gaussian terms, the mini-batch quadratic variation, and the martingale fluctuations in the stopped recursion are controlled with high probability using Gaussian and chi-square concentration, a without-replacement variance bound, and martingale concentration. Summing the resulting stopped recursion gives

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathcal{L}_S(\widetilde{\mathbf{W}}_t) \lesssim \frac{\mathcal{L}_S(\mathbf{W}^*)}{\rho_{\mathrm{clip}}} + \frac{\|\mathbf{W}_0 - \mathbf{W}^*\|_2^2}{\eta_{\mathrm{eff}} T} + \eta_{\mathrm{eff}}\Big(\frac{1}{B} + \frac{1}{T}\Big) + \mathrm{Noise}_{\lambda,\kappa},$$

as long as the trajectory remains localized, where $\eta_{\mathrm{eff}} = \rho_{\mathrm{clip}}\eta$ and $\mathrm{Noise}_{\lambda,\kappa}$ collects the λ-dependent Gaussian fluctuation terms.

STEP 5: CLOSE THE LOCALIZATION BOOTSTRAP AND ESTABLISH PROJECTION INACTIVITY. For a comparator satisfying the properties verified in Step 6, the width and step-size conditions in Theorem 1, together with the stopped potential bound from Step 4, close the localization bootstrap and yield $\sup_{0\le t\le T}\|\mathbf{U}_t - \mathbf{W}^*\|_2 \lesssim \tau_\gamma$. Since $\widetilde{\mathbf{W}}_t = \mathbf{U}_t - \eta c_{\mathrm{priv}} Z_t$, the Gaussian maximum-norm bound and the upper bound on m imply $\eta c_{\mathrm{priv}}\|Z_t\|_2 \lesssim \tau_\gamma$ uniformly in t. Combining these bounds with $\|\mathbf{W}^* - \mathbf{W}_0\|_2 \lesssim \tau_\gamma$ gives $\sup_{0\le t\le T}\|\widetilde{\mathbf{W}}_t - \mathbf{W}_0\|_2 \lesssim \tau_\gamma$. Choosing $R_* = C_R\tau_\gamma$ with C_R sufficiently large therefore shows that projection is inactive, and hence $\mathbf{W}_t = \widetilde{\mathbf{W}}_t$ for all $t \le T$.

STEP 6: CONSTRUCT THE COMPARATOR FROM NTK SEPARABILITY. It remains to verify that a comparator satisfying the required localization and loss properties exists. Assumption 2 yields a comparator $\mathbf{W}^* = \mathbf{W}_0 + \tau_{\mathrm{ntk}}\mathbf{u}$ such that, with high probability,

$$\|\mathbf{W}^* - \mathbf{W}_0\|_2^2 \lesssim \tau_\gamma^2, \qquad \mathcal{L}_S(\mathbf{W}^*) \le \frac{\rho_{\mathrm{clip}}}{T},$$

where $\tau_\gamma^2 \asymp \big(\log^2(T/\rho_{\text{clip}}) + \log(n/\delta)\big)/\gamma^2$. Substituting this comparator into the bound from Step 4 and using the projection-inactivity result from Step 5 gives $\frac{1}{T}\sum_{t=0}^{T-1} \mathcal{L}_S(\mathbf{W}_t) \lesssim A_{\text{corr}}$. A union bound over the required high-probability events completes the proof. □

5 APPLICATIONS

This section applies the optimization result from Section 4 to derive population risk guarantees via algorithmic stability. We then illustrate, through a finite-horizon example, how privacy calibration can preserve the optimization-side benefit of temporal correlation. Finally, we derive the corresponding guarantees for independent noise DP-SGD, non-private mini-batch SGD, and full-batch training.

5.1 POPULATION RISK OF DP-SGD WITH CORRELATED NOISE

The following theorem converts the optimization quantity A_{corr} from Theorem 1 into a population risk guarantee through algorithmic stability, without requiring a particular privacy calibration of the noise scale. The detailed proof is given in Appendix D. Denote the algorithmic randomness by $\mathcal{A} = \{\mathcal{B}_0, \dots, \mathcal{B}_{T-1}, Z_0, \dots, Z_{T-1}\}$. We suppress logarithmic factors in the bound below.

Theorem 2 (Population risk bound)**.** *Suppose Assumption 1 holds and the parameter conditions of Theorem 1 are satisfied. Assume further that, conditional on the initialization, an i.i.d. sample $S \sim \mathcal{P}^n$ satisfies Assumption 2 with margin γ with probability at least $1 - \delta_{\text{sep}}$, where $\delta_{\text{sep}} \le \delta/8$. Let $\{\mathbf{W}_t\}_{t=0}^T$ be generated by Algorithm 1 with $C_{\text{clip}} > 0$ and $\eta_{\text{eff}} \le \frac{1}{12C_{\sigma,b}p^3}$. Assume $m \gtrsim R_*^2 \log(m/\delta)\eta^2\big(TA_{\text{corr}} + (\log(n/\delta) + R_*)\big(\sqrt{T/B} + \log(1/\delta)\big)\big)^2$. Then, with probability at least $1-\delta$ over the initialization,*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}_{S,\mathcal{A}}[\mathcal{L}(\mathbf{W}_t)] \lesssim \Big(1 + \frac{\eta T}{n}\Big)\Big(A_{\text{corr}} + (\sqrt{m} + R_*)\delta\Big).$$

Theorem 2 shows that the dependence on temporal correlation and clipping obtained in the optimization analysis is preserved at the population level. In particular, when $\eta T/n = \mathcal{O}(1)$ and δ is sufficiently small, the averaged population risk inherits the leading order of A_{corr} up to constant and logarithmic factors.

A finite-horizon example where privacy calibration preserves the optimization gain. Consider full-batch training $B = n$ with any fixed finite horizon T, $C_{\text{clip}} = G_\delta$, and the same privacy budget (ϵ, δ) for the correlated and independent mechanisms. Under a finite-horizon matched calibration, the linear and quadratic privacy-dependent terms relative to independent noise are multiplied by

$$r_1(\lambda) = \Big(\frac{1}{T}\sum_{t=1}^{T}(1-\lambda^t)^2\Big)^{1/2}, \qquad r_2(\lambda) = \big((1-\lambda)^2 + \eta\lambda^2\big)\frac{1}{T}\sum_{t=1}^{T}\Big(\frac{1-\lambda^t}{1-\lambda}\Big)^2.$$

For every $\lambda \in (0,1)$, $r_1(\lambda) < 1$, while $r_2(\lambda) = 1 - \frac{2\lambda}{T} + \mathcal{O}(\lambda^2)$ as $\lambda \to 0^+$. Hence, for every fixed finite T, there exists $\lambda_T \in (0,1)$ such that $r_1(\lambda_T) < 1$ and $r_2(\lambda_T) < 1$. Under the common admissibility conditions of Theorem 2, temporal correlation therefore yields a strictly smaller population risk bound than independent noise DP-SGD under the same privacy budget. The detailed privacy calculation and resulting risk bound are given in Appendix E.1.

Conservative worst-case benchmark. Appendix E.2 analyzes a conservative worst-case mini-batch privacy calibration. Even under this calibration, the resulting population risk bound still retains, up to logarithmic and model-dependent factors, the optimal dependence on n, d, and ϵ known for Lipschitz convex DP stochastic convex optimization (Bassily et al., 2019).

5.2 RECOVERY OF STANDARD REGIMES

Our general framework recovers independent noise DP-SGD, non-private mini-batch SGD, and the corresponding full-batch regimes as special cases. For the representative rates below, we take

$C_{\text{clip}} = G_\delta$ for simplicity; other clipping thresholds can be handled through the general clipping analysis in Theorem 1.

Independent noise DP-SGD. Setting $\lambda = 0$ recovers standard DP-SGD with independent Gaussian noise. Under the standard privacy calibration $\kappa^2 \asymp B^2 T \log(1/\delta)/(n^2\epsilon^2)$, Algorithm 1 satisfies (ϵ, δ)-DP. Under a representative parameter regime, the resulting averaged optimization and population risks are both of order $\mathcal{O}\big(\frac{1}{\sqrt{n}} + \frac{\sqrt{d}}{n\epsilon}\big)$. The complete optimization and generalization analyses, together with the precise parameter choices, are given in Appendix F.

Non-private mini-batch SGD. Setting $\kappa = 0$ removes the Gaussian perturbation and recovers non-private mini-batch SGD. Under a representative parameter regime, both the averaged optimization and population risks are of order $\mathcal{O}(1/n)$. The corresponding general result and its proof are given in Appendix G.

Full-batch GD and DP-GD. Taking $B = n$ recovers the corresponding full-batch regimes. In the fixed-second-layer setting, the independent noise DP-GD specialization yields a population risk rate $\mathcal{O}\big(\frac{\sqrt{d}}{n\epsilon}\big)$ under $m \asymp \text{polylog}(n/\delta)/\gamma^4$, matching the established sample and privacy dependence in Wang et al. (2026). Similarly, the non-private full-batch specialization yields a population risk rate $\mathcal{O}(1/n)$, matching the rate in Wang et al. (2026).

6 Discussion and Limitations

On fixing the second layer. In this work, we draw the second-layer coefficients $\mathbf{c}$ at initialization and keep them fixed throughout training, while optimizing only the first-layer parameters $\mathbf{W}$. This is a common theoretical formulation for shallow neural networks. In the KAN literature, Eshtehardian et al. (2026) explicitly study two-layer KANs with first-layer training. The same modeling choice is also common in the theory of fully connected neural networks Lei et al. (2026); Taheri & Thrampoulidis (2024); Lei et al. (2022) fix the output-layer weights and optimize only the hidden-layer parameters. We adopt this formulation for analytical simplicity and to avoid introducing an additional trainable parameter block into the main analysis.

More importantly, fixing the second layer is not responsible for the leading-order rates obtained in this paper. Wang et al. (2026) analyze the corresponding two-layer KAN with both layers trained and obtain the same leading optimization and population risk rates. Their joint gradient and Hessian bounds, including the mixed first–second-layer blocks, provide the local estimates needed in our analysis. Thus, replacing $\mathbf{W}$ by the full parameter vector and incorporating these estimates would mainly require additional bookkeeping for the second-layer trajectory and is not expected to change the leading-order bounds. We therefore fix the second layer for analytical simplicity.

Limitation. Our analysis focuses on binary classification with the logistic loss, while corresponding optimization and population risk guarantees for regression remain open. A second limitation is the restriction to two-layer KANs. Extending the analysis to deep KANs is tied to the broader open problem of understanding the optimization and generalization of deep KANs, and is therefore beyond a routine extension of the present arguments. Finally, our results are developed specifically for KANs. Extending the correlated noise optimization and population risk analysis to other neural network architectures, such as MLPs and CNNs, would require architecture-specific control of their local training dynamics and remains an interesting direction for future work.

7 Conclusion

We established optimization risk bounds for two-layer KANs trained by clipped mini-batch DP-SGD with correlated noise, explicitly characterizing the effects of temporal correlation, clipping, mini-batch sampling, and network width. Combining this analysis with algorithmic stability yields corresponding population risk guarantees. We further showed that the optimization benefit of temporal correlation can survive privacy calibration in a finite-horizon full-batch regime, while even a conservative worst-case mini-batch calibration retains, up to logarithmic and model-dependent factors, the optimal dependence on n, d, and ϵ known for Lipschitz convex DP-SCO. Our framework also recovers standard private and non-private training regimes as special cases.

AI USE STATEMENT

We used generative AI tools to improve the readability of the manuscript, identify relevant literature, suggest the manuscript structure, and propose titles and keywords. We also used these tools as a research assistant for interpreting results and assisting with checking and revision of theoretical arguments and proofs. All AI-assisted material was independently verified, revised, and integrated by the authors. Generative AI was not used to generate experimental data or results. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pp. 242–252. PMLR, 2019.

Meenatchi Sundaram Muthu Selva Annamalai, Borja Balle, Jamie Hayes, Georgios Kaissis, and Emiliano De Cristofaro. The hitchhiker's guide to efficient, end-to-end, and tight dp auditing. *arXiv preprint arXiv:2506.16666*, 2025.

Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pp. 322–332. PMLR, 2019.

Gilles Barthe and Federico Olmedo. Beyond differential privacy: Composition theorems and relational logic for f-divergences between probabilistic programs. In *International Colloquium on Automata, Languages, and Programming*, pp. 49–60. Springer, 2013.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Guha Thakurta. Private stochastic convex optimization with optimal rates. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yuan Cao and Quanquan Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Zixiang Chen, Yuan Cao, Difan Zou, and Quanquan Gu. How much over-parameterization is sufficient to learn deep ReLU networks? In *International Conference on Learning Representation*, 2021.

Oleksandr Cherednichenko and Maria Poptsova. Kolmogorov–arnold networks for genomic tasks. *Briefings in Bioinformatics*, 26(2):bbaf129, 2025.

Christopher A Choquette-Choo, Arun Ganesh, Ryan McKenna, H Brendan McMahan, John Rush, Abhradeep Guha Thakurta, and Zheng Xu. (amplified) banded matrix factorization: A unified approach to private training. In *Advances in Neural Information Processing Systems*, volume 36, pp. 74856–74889, 2023. URL https://openreview.net/forum?id=zEm6hF97Pz.

Christopher A Choquette-Choo, Krishnamurthy Dj Dvijotham, Krishna Pillutla, Arun Ganesh, Thomas Steinke, and Abhradeep Guha Thakurta. Correlated noise provably beats independent noise for differentially private learning. In *International Conference on Learning Representations*, 2024a.

Christopher A Choquette-Choo, Arun Ganesh, Saminul Haque, Thomas Steinke, and Abhradeep Thakurta. Near exact privacy amplification for matrix mechanisms. *arXiv preprint arXiv:2410.06266*, 2024b.

Christopher A. Choquette-Choo, Arun Ganesh, Thomas Steinke, and Abhradeep Guha Thakurta. Privacy amplification for matrix mechanisms. In *International Conference on Learning Representations*, 2024c. URL https://openreview.net/forum?id=xUzWmFdglP.

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. In *Advances in Neural Information Processing Systems*, volume 35, pp. 5910–5924, 2022.

Meng Ding, Mingxi Lei, Shaopeng Fu, Shaowei Wang, Di Wang, and Jinhui Xu. Understanding private learning from feature perspective. *arXiv preprint arXiv:2511.18006*, 2025.

Cynthia Dwork. Differential privacy. In *International colloquium on automata, languages, and programming*, 2006.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In Shai Halevi and Tal Rabin (eds.), *Theory of Cryptography*, pp. 265–284, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-32732-5.

Seyed Mohammad Eshtehardian, Mohammad Hossein Yassaee, and Babak Khalaj. On the convergence of two-layer kolmogorov-arnold networks with first-layer training. In *International Conference on Learning Representations*, 2026.

Spencer Frei, Niladri S Chatterji, and Peter L Bartlett. Random feature amplification: Feature learning and generalization in neural networks. *Journal of Machine Learning Research*, 24(303):1–49, 2023.

Yihang Gao and Vincent YF Tan. On the convergence of (stochastic) gradient descent for Kolmogorov–Arnold networks. *IEEE Transactions on Information Theory*, 2025.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020. doi: 10.1214/18-PS321.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in Neural Information Processing Systems*, 31, 2018.

Ziwei Ji and Matus Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow ReLU networks. In *International Conference on Learning Representations*, 2020.

Nikita P Kalinin, Ryan McKenna, Rasmus Pagh, and Christoph H Lampert. DP-λCGD: efficient noise correlation for differentially private model training, 2026. arXiv preprint arXiv:2601.22334.

Anastasiia Koloskova, Ryan McKenna, Zachary Charles, John Rush, and H Brendan McMahan. Gradient descent with linearly correlated noise: Theory and applications to differential privacy. In *Advances in Neural Information Processing Systems*, volume 36, pp. 35761–35773, 2023.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pp. 1302–1338, 2000.

Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *International Conference on Machine Learning*, pp. 5809–5819. PMLR, 2020.

Yunwen Lei, Rong Jin, and Yiming Ying. Stability and generalization analysis of gradient methods for shallow neural networks. In *Advances in Neural Information Processing Systems*, volume 35, pp. 38557–38570, 2022.

Yunwen Lei, Puyu Wang, Yiming Ying, and Ding-Xuan Zhou. Optimization and generalization of gradient descent for shallow ReLU networks with minimal width. *Journal of Machine Learning Research*, 27(34):1–35, 2026.

Longlong Li, Yipeng Zhang, Guanghui Wang, and Kelin Xia. Kolmogorov–arnold graph neural networks for molecular property prediction. *Nature Machine Intelligence*, 7(8):1346–1354, 2025a.

Pengqi Li, Lizhong Ding, Jiarun Fu, Guoren Wang, Ye Yuan, et al. Generalization bounds for kolmogorov-arnold networks (KANs) and enhanced KANs with lower lipschitz complexity. In *Advances in Neural Information Processing Systems*, 2025b.

Yuanfan Li, Yunwen Lei, Zheng-Chu Guo, and Yiming Ying. Optimal rates for generalization of gradient descent for deep ReLU classification. In *Advances in Neural Information Processing Systems*, 2026.

Wei Liu, Eleni Chatzi, and Zhilu Lai. On the rate of convergence of kolmogorov-arnold network regression estimators. *arXiv preprint arXiv:2509.19830*, 2025a.

Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y Hou, and Max Tegmark. KAN: Kolmogorov-Arnold networks. In *International Conference on Learning Representations*, 2025b.

Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pp. 263–275. IEEE, 2017.

Mike Nguyen and Nicole Muecke. How many neurons do we need? a refined analysis for shallow networks trained with gradient descent. *Journal of Statistical Planning and Inference*, 233:106169, 2024.

Atsushi Nitanda and Taiji Suzuki. Optimal rates for averaged stochastic gradient descent under neural tangent kernel regime. In *International Conference on Learning Representations*, 2021.

Atsushi Nitanda, Geoffrey Chinot, and Taiji Suzuki. Gradient descent can learn less over-parameterized two-layer neural networks on classification problems. *arXiv preprint arXiv:1905.09870*, 2019.

Subhajit Patra, Sonali Panda, Bikram Keshari Parida, Mahima Arya, Kurt Jacobs, Denys I Bondar, and Abhijit Sen. Physics informed kolmogorov-arnold neural networks for dynamical analysis via efficient-kan and wav-kan. *Journal of Machine Learning Research*, 26(233):1–39, 2025.

Dominic Richards and Ilja Kuzborskij. Stability & generalisation of gradient descent for shallow neural networks without the neural tangent kernel. In *Advances in Neural Information Processing Systems*, volume 34. PMLR, 2021.

Jan Schuchardt and Nikita Kalinin. Sampling-free privacy accounting for matrix mechanisms under random allocation, 2026.

Zhongjie Shi, Puyu Wang, Chenyang Zhang, and Yuan Cao. Towards understanding generalization in DP-GD: A case study in training two-layer CNNs. In *AAAI Conference on Artificial Intelligence*, 2026.

Khemraj Shukla, Juan Diego Toscano, Zhicheng Wang, Zongren Zou, and George Em Karniadakis. A comprehensive and fair comparison between mlp and kan representations for differential equations and operator networks. *Computer Methods in Applied Mechanics and Engineering*, 431:117290, 2024.

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.

Hossein Taheri and Christos Thrampoulidis. Generalization and stability of interpolating neural networks with minimal width. *Journal of Machine Learning Research*, 25(156):1–41, 2024.

Hossein Taheri, Christos Thrampoulidis, and Arya Mazumdar. Sharper guarantees for learning neural network classifiers with gradient methods. In *International Conference on Learning Representations*, 2025.

Cristian J Vaca-Rubio, Luis Blanco, Roberto Pereira, and Màrius Caus. Kolmogorov-arnold networks (kans) for time series analysis. In *2024 IEEE Globecom Workshops (GC Wkshps)*, pp. 1–6. IEEE, 2024.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Puyu Wang, Yunwen Lei, Marius Kloft, and Yiming Ying. Optimal utility bounds for differentially private gradient descent in three-layer neural networks. In *2025 IEEE 12th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 1–8. IEEE, 2025a.

Puyu Wang, Yunwen Lei, Di Wang, Yiming Ying, and Ding-Xuan Zhou. Generalization guarantees of gradient descent for shallow neural networks. *Neural Computation*, 37(2):344–402, 2025b.

Puyu Wang, Junyu Zhou, Philipp Liznerski, and Marius Kloft. Optimization, generalization and differential privacy bounds for gradient descent on Kolmogorov-Arnold networks. In *International Conference on Machine Learning*, 2026.

Yixuan Wang, Jonathan W Siegel, Ziming Liu, and Thomas Y Hou. On the expressiveness and spectral bias of kans. In *International Conference on Learning Representations*, 2025c.

Yizheng Wang, Jia Sun, Jinshuai Bai, Cosmin Anitescu, Mohammad Sadegh Eshaghi, Xiaoying Zhuang, Timon Rabczuk, and Yinghua Liu. Kolmogorov–arnold-informed neural network: A physics-informed deep learning framework for solving forward and inverse problems based on kolmogorov–arnold networks. *Computer Methods in Applied Mechanics and Engineering*, 433: 117518, 2025d.

Yu-Xiang Wang, Borja Balle, and Shiva Prasad Kasiviswanathan. Subsampled rényi differential privacy and analytical moments accountant. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

Ruichen Xu and Kexin Chen. Differential privacy in two-layer networks: How dp-sgd harms fairness and robustness. *arXiv preprint arXiv:2603.04881*, 2026.

Jiaming Zhang, Huanyi Xie, Meng Ding, Shaopeng Fu, Jinyan Liu, and Di Wang. Understanding the impact of differentially private training on memorization of long-tailed data. *arXiv preprint arXiv:2602.03872*, 2026.

Junyu Zhou, Puyu Wang, and Ding-Xuan Zhou. Generalization analysis with deep relu networks for metric and similarity learning. *arXiv preprint arXiv:2405.06415*, 2024.

Yuqing Zhu, Jinshuo Dong, and Yu-Xiang Wang. Optimal accounting of differential privacy via characteristic function. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep relu networks. *Machine learning*, 109:467–492, 2020.

Appendix

A Further Related Work

This appendix expands on the related work referenced in Section 2, covering generalization theory for neural networks (Appendix A.1) and privacy amplification by subsampling for correlated-noise mechanisms (Appendix A.2).

A.1 Population risk bounds for neural networks

Most existing optimization and generalization results for neural networks have been developed for *fully connected* multilayer perceptrons (MLPs), rather than architectures with edge-wise functional parametrizations such as KANs. Broadly speaking, the theoretical literature on MLPs can be organized into three main directions.

A first line of work is based on the neural tangent kernel (NTK) viewpoint Jacot et al. (2018). In this regime, a sufficiently wide network behaves approximately like its linearization around initialization, which makes it possible to analyze gradient-based training through an associated kernel. This perspective has led to influential convergence guarantees for overparameterized MLPs and, in the lazy-training regime, to sharp statistical and excess risk bounds in a variety of settings Allen-Zhu et al. (2019); Arora et al. (2019); Cao & Gu (2019); Chen et al. (2021); Nguyen & Muecke (2024); Nitanda & Suzuki (2021); Zou et al. (2020).

A second line of work studies generalization through uniform convergence approach. Typical tools include Rademacher complexity, covering numbers, norm-based complexity measures, and margin-based arguments Bartlett et al. (2017); Frei et al. (2023); Ji & Telgarsky (2020); Lei et al. (2026); Li et al. (2026); Nitanda et al. (2019); Shi et al. (2026); Zhou et al. (2024). These approaches yield broad generalization guarantees for MLPs, often under relatively mild assumptions on the training algorithm itself.

More recently, there has been growing interest in algorithm-dependent generalization analyses based on algorithmic stability Lei et al. (2022); Richards & Kuzborskij (2021); Taheri & Thrampoulidis (2024); Taheri et al. (2025); Wang et al. (2025b). This line of work studies how perturbations in the training data propagate through the optimization dynamics, and consequently yields generalization and excess risk bounds that are directly tied to the learning algorithm. This viewpoint is especially relevant to the present paper, since our goal is likewise to derive algorithm-dependent risk guarantees, but now in the more structured setting of KANs and under minibatch, clipped, and differentially private training.

A.2 Privacy amplification by subsampling for correlated noise

Prior analyses of correlated-noise mechanisms with privacy amplification cover banded matrix mechanisms, Poisson-subsampled matrix mechanisms, and Balls-in-Bins schemes Choquette-Choo et al. (2024b; 2023; 2024c); Schuchardt & Kalinin (2026). We complement this line by deriving an analytical noise-multiplier bound for DP-λCGD under mini-batches drawn uniformly without replacement, which underlies the privacy guarantee for Algorithm 1.

B Useful Lemmas

In this section, we introduce some useful lemmas that will be used in the proofs.

Lemma 3 (Wang et al. (2026)). *Let $s \in \mathbb{N}$ and $\mathbf{A}_i \in \mathbb{R}^{m_i \times n_i}$ for $i \in [s]$. Define the diagonal block matrix*

$$\mathbf{B} = \begin{bmatrix} \mathbf{A}_1 & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{A}_s \end{bmatrix} \in \mathbb{R}^{\sum_{i=1}^{s} m_i \times \sum_{i=1}^{s} n_i}$$

Then, it holds that

$$\|\mathbf{B}\|_2 = \max_{i \in [s]} \|\mathbf{A}_i\|_2.$$

Lemma 4. *For any $n \geq 2$, let $x_1, \ldots, x_n \in \mathbb{R}^D$ be deterministic vectors, and let $\bar{x} := \frac{1}{n}\sum_{i=1}^n x_i$. Suppose that $\mathcal{B} \subseteq [n]$ is sampled uniformly without replacement among all subsets of cardinality B. Define $\bar{x}_\mathcal{B} = \frac{1}{B}\sum_{i\in\mathcal{B}} x_i$. Then, $\mathbb{E}_\mathcal{B}[\bar{x}_\mathcal{B}] = \bar{x}$ and*

$$\mathbb{E}_\mathcal{B}\big[\|\bar{x}_\mathcal{B} - \bar{x}\|_2^2\big] = \frac{n-B}{B(n-1)} \cdot \frac{1}{n}\sum_{i=1}^n \|x_i - \bar{x}\|_2^2.$$

Furthermore, if $\|x_i\|_2 \leq G$ for all $i \in [n]$, then

$$\mathbb{E}_\mathcal{B}\big[\|\bar{x}_\mathcal{B} - \bar{x}\|_2^2\big] \leq \frac{G^2}{B}.$$

Proof. The proof is standard, we include it for completeness. The unbiasedness follows from

$$\mathbb{E}_\mathcal{B}[\bar{x}_\mathcal{B}] = \frac{1}{B}\sum_{i=1}^n \mathbb{P}(i \in \mathcal{B})\, x_i = \frac{1}{B}\sum_{i=1}^n \frac{B}{n} x_i = \bar{x}.$$

To prove the second identity, let $x_i' = x_i - \bar{x}$ for any $i \in [n]$. Then it holds $\sum_{i=1}^n x_i' = 0$ and $\bar{x}_\mathcal{B} - \bar{x} = \frac{1}{B}\sum_{i\in\mathcal{B}} x_i'$. Hence,

$$\mathbb{E}_\mathcal{B}\big[\|\bar{x}_\mathcal{B} - \bar{x}\|_2^2\big] = \frac{1}{B^2}\,\mathbb{E}_\mathcal{B}\Big[\Big\|\sum_{i\in\mathcal{B}} x_i'\Big\|_2^2\Big] = \frac{1}{B^2}\Big(\sum_{i=1}^n \mathbb{P}(i\in\mathcal{B})\|x_i'\|_2^2 + \sum_{i\neq j}\mathbb{P}(i,j\in\mathcal{B})\langle x_i', x_j'\rangle\Big).$$

Since $\mathcal{B}$ is sampled uniformly without replacement with $|\mathcal{B}| = B$, we know

$$\mathbb{P}(i\in\mathcal{B}) = \frac{B}{n}, \qquad \mathbb{P}(i,j\in\mathcal{B}) = \frac{B(B-1)}{n(n-1)} \quad (i\neq j).$$

Moreover, from $\sum_{i=1}^n x_i' = 0$ we get

$$\sum_{i\neq j}\langle x_i', x_j'\rangle = \Big\|\sum_{i=1}^n x_i'\Big\|_2^2 - \sum_{i=1}^n \|x_i'\|_2^2 = -\sum_{i=1}^n \|x_i'\|_2^2.$$

Therefore,

$$\mathbb{E}_\mathcal{B}\big[\|\bar{x}_\mathcal{B} - \bar{x}\|_2^2\big] = \frac{1}{B^2}\Big(\frac{B}{n} - \frac{B(B-1)}{n(n-1)}\Big)\sum_{i=1}^n \|x_i'\|_2^2 = \frac{n-B}{B(n-1)}\cdot\frac{1}{n}\sum_{i=1}^n \|x_i - \bar{x}\|_2^2.$$

Further, if $\|x_i\|_2 \leq G$ for all $i \in [n]$, then

$$\frac{1}{n}\sum_{i=1}^n \|x_i - \bar{x}\|_2^2 = \frac{1}{n}\sum_{i=1}^n \|x_i\|_2^2 - \|\bar{x}\|_2^2 \leq \frac{1}{n}\sum_{i=1}^n \|x_i\|_2^2 \leq G^2.$$

Hence,

$$\mathbb{E}_\mathcal{B}\big[\|\bar{x}_\mathcal{B} - \bar{x}\|_2^2\big] \leq \frac{n-B}{B(n-1)}\,G^2 \leq \frac{G^2}{B}.$$

This completes the proof. □

Lemma 5 (Chernoff bound for adapted bounded variables)**.** *Let $\{X_t\}_{t=1}^T$ be a nonnegative adapted process with respect to $\{\mathscr{F}_t\}_{t=0}^T$. Suppose $0 \leq X_t \leq L$ a.s. and $\sum_{t=1}^T \mathbb{E}[X_t \mid \mathscr{F}_{t-1}] \leq V$ a.s. Then, for every $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\sum_{t=1}^T X_t \leq 2V + 2L\log\Big(\frac{1}{\delta}\Big).$$

Lemma 6 (Sub-Gaussianity of sampling without replacement)**.** *Let $a_1, \ldots, a_n \in \mathbb{R}$ be deterministic and let $\bar{a} = \frac{1}{n}\sum_{i=1}^n a_i$. Let $\mathcal{B} \subset [n]$ be sampled uniformly without replacement with $|\mathcal{B}| = B$. If $|a_i| \leq R$ for all $i \in [n]$, then, for every $\theta \in \mathbb{R}$,*

$$\mathbb{E}\Big[\exp\Big(\theta\Big(\frac{1}{B}\sum_{i\in\mathcal{B}} a_i - \bar{a}\Big)\Big)\Big] \leq \exp\Big(\frac{\theta^2 R^2}{2B}\Big).$$

Proof. Let $I_1, \dots, I_B$ be sampled uniformly without replacement from $[n]$, and let $J_1, \dots, J_B$ be i.i.d. uniform on $[n]$. By Hoeffding's comparison theorem for sampling without replacement (Hoeffding, 1963, Theorem 4),

$$\mathbb{E}\Big[\exp\Big(\theta\Big(\frac{1}{B}\sum_{s=1}^{B} a_{I_s} - \bar{a}\Big)\Big)\Big] \le \mathbb{E}\Big[\exp\Big(\frac{\theta}{B}\sum_{s=1}^{B}(a_{J_s} - \bar{a})\Big)\Big].$$

Since $a_{J_s} - \bar{a}$ has mean zero and lies in an interval of length at most $2R$, Hoeffding's lemma yields

$$\mathbb{E}\Big[\exp\Big(\frac{\theta}{B}(a_{J_s} - \bar{a})\Big)\Big] \le \exp\left(\frac{\theta^2 R^2}{2B^2}\right).$$

Independence of $J_1, \dots, J_B$ completes the proof. □

Our convergence analysis relies on the following standard mirror descent inequality (see, e.g., Wang et al. (2026)).

Lemma 7 (Mirror descent inequality)**.** *Let $\mathcal{K} \subseteq \mathbb{R}^{mdp}$ be a nonempty closed convex set, and let $Proj_{\mathcal{K}}(\cdot)$ denote the Euclidean projection onto $\mathcal{K}$. Fix any $\eta > 0$ and any vector $g \in \mathbb{R}^{mdp}$. Define*

$$\mathbf{W}^+ = Proj_{\mathcal{K}}\big(\mathbf{W} - \eta g\big).$$

Then, for any comparator $\mathbf{W}^* \in \mathcal{K}$, the following inequality holds

$$\big\langle g,\ \mathbf{W} - \mathbf{W}^* \big\rangle \le \frac{1}{2\eta}\Big(\|\mathbf{W} - \mathbf{W}^*\|_2^2 - \|\mathbf{W}^+ - \mathbf{W}^*\|_2^2\Big) + \frac{\eta}{2}\|g\|_2^2.$$

Denote $[\mathbf{W}_1, \mathbf{W}_2] = \{\alpha\mathbf{W}_1 + (1-\alpha)\mathbf{W}_2 : \alpha \in [0,1]\}$ as the line segment between $\mathbf{W}_1$ and $\mathbf{W}_2$.

Lemma 8 (Local quasi-convexity property Taheri & Thrampoulidis (2024))**.** *Suppose $s \in \mathbb{N}$ and $G : \mathbb{R}^s \to \mathbb{R}$ be a second-order differentiable function satisfying $\lambda_{\min}(\nabla^2 G(\mathbf{W})) \ge -\kappa G(\mathbf{W})$. Let $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^s$ be two arbitrary points with distance $\|\mathbf{W}_1 - \mathbf{W}_2\|_2 \le D \le \sqrt{2/\kappa}$. Let $\tau := (1 - D^2\kappa/2)^{-1}$. Then,*

$$\max_{\mathcal{V} \in [\mathbf{W}_1, \mathbf{W}_2]} G(\mathcal{V}) \le \tau \max\big\{G(\mathbf{W}_1), G(\mathbf{W}_2)\big\}.$$

For any scalar $v \in \mathbb{R}$, we denote $\mathbf{h}(v) = [b_1(v), \dots, b_p(v)]^\top \in \mathbb{R}^p$. For $s \in \mathbb{N}$ and a vector $\mathbf{u} = [u_1, \dots, u_s]^\top \in \mathbb{R}^s$, we denote $\mathbf{h}(\mathbf{u}) = [\mathbf{h}(u_1)^\top, \dots, \mathbf{h}(u_s)^\top]^\top \in \mathbb{R}^{sp}$. For each hidden unit $j \in [m]$, the first-layer spline coefficients $\{w_{i,j,k}\}_{i\in[d],\, k\in[p]}$ are arranged into a vector $\mathbf{w}_j \in \mathbb{R}^{dp}$ according to a fixed ordering of (i,k). We write $\mathbf{W} = (\mathbf{w}_1, \dots, \mathbf{w}_m)^\top \in \mathbb{R}^{m\times dp}$. The second-layer spline coefficients $\{c_{j,k}\}_{j\in[m],\, k\in[p]}$ are collected as $\mathbf{c} = (\mathbf{c}_1, \dots, \mathbf{c}_m) \in \mathbb{R}^{mp}$ with $\mathbf{c}_j = (c_{j,1}, \dots, c_{j,p}) \in \mathbb{R}^p$. Then $f_{\mathbf{W}}$ can be rewritten as

$$f_{\mathbf{W}}(\mathbf{x}) = \frac{1}{\sqrt{m}}\mathbf{c}^\top \mathbf{h}\Big(\sigma\big(\frac{1}{\sqrt{d}}\mathbf{W}\mathbf{h}(\mathbf{x})\big)\Big).$$

To control the gradients and Hessians, we use the following lemma Wainwright (2019), which provides estimates for $\mathbf{c} \in \mathbb{R}^{mp}$.

Lemma 9. *For any $\delta \in (0,1)$, define the event*

$$\mathcal{E}_\delta := \Big\{\|\mathbf{c}\|_2 \le 4\sqrt{pm} + 2\sqrt{\log(2/\delta)}\ \text{ and }\ \max_{j\in[m]}\|\mathbf{c}_j\|_2 \le 4\sqrt{p} + 2\sqrt{\log(2m/\delta)}\Big\}. \tag{1}$$

It holds that $\mathbb{P}(\mathcal{E}_\delta) \ge 1 - \delta$.

To simplify the analysis, we work on the above high-probability event associated with the initialization of $\mathbf{c}$.

Lemma 10 (Gradient and Hessian)**.** *Let $\delta \in (0,1)$ and $C_{\sigma,b} > 0$ be a constant that depends solely on σ, b. Suppose Assumption 1 and the event $\mathcal{E}_\delta$ occurs. It holds for any $\mathbf{W}$ and any $\mathbf{x} \in \mathcal{X}$ that*

$$\big\|\nabla f_{\mathbf{W}}(\mathbf{x})\big\|_2 \le C_{\sigma,b}\, p\left(\sqrt{p} + \sqrt{\frac{\log(1/\delta)}{m}}\right)$$

and

$$\big\|\nabla^2 f_{\mathbf{W}}(\mathbf{x})\big\|_2 \le \frac{C_{\sigma,b}\, p^{\frac{3}{2}}\big(\sqrt{p} + \sqrt{\log(m/\delta)}\big)}{\sqrt{m}}.$$

Proof. The proof is based on arguments from Wang et al. (2026), where both $\mathbf{c}$ and $\mathbf{W}$ are trained. For completeness, we provide a detailed proof. We first estimate $\|\nabla f_{\mathbf{W}}(\mathbf{x})\|_2$. For all $v \in \mathbb{R}$, we denote $\mathbf{h}'(v) = [b_1'(v), \ldots, b_p'(v)]^\top \in \mathbb{R}^p$ and $\mathbf{h}''(v) = [b_1''(v), \ldots, b_p''(v)]^\top \in \mathbb{R}^p$. Define

$$\mathbf{u}(\mathbf{x}) = \sigma\big(\frac{1}{\sqrt{d}}\mathbf{W}'\mathbf{h}(\mathbf{x})\big) \text{ and } \mathbf{D}(\mathbf{x}) = \mathrm{diag}\big(\sigma'(\frac{1}{\sqrt{d}}\mathbf{w}_i^\top \mathbf{h}(\mathbf{x}))\big)_{i=1}^m \in \mathbb{R}^{m\times m}. \tag{2}$$

From the form $f_{\mathbf{W}}(\mathbf{x}) = \frac{1}{\sqrt{m}}\mathbf{c}^\top \mathbf{h}(\mathbf{u}(\mathbf{x}))$, we have

$$\frac{\partial \mathbf{h}(\mathbf{u}(\mathbf{x}))}{\partial \mathbf{u}(\mathbf{x})} = \begin{bmatrix} \mathbf{h}'(u_1(\mathbf{x})) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{h}'(u_2(\mathbf{x})) & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{h}'(u_m(\mathbf{x})) \end{bmatrix} \in \mathbb{R}^{mp\times m}$$

and

$$\frac{\partial \mathbf{u}(\mathbf{x})}{\partial \mathbf{w}_i} = \begin{bmatrix} \mathbf{0} \\ \vdots \\ \frac{1}{\sqrt{d}}\sigma'(\frac{1}{\sqrt{d}}\mathbf{w}_i^\top \mathbf{h}(\mathbf{x}))\mathbf{h}(\mathbf{x})^\top \\ \vdots \\ \mathbf{0} \end{bmatrix} \in \mathbb{R}^{m\times pd}.$$

According to the chain rule, for any $i \in [m]$, it holds that

$$\partial_{\mathbf{w}_i} f_{\mathbf{W}}(\mathbf{x}) = \frac{\partial f_{\mathbf{W}}(\mathbf{x})}{\partial \mathbf{h}(\mathbf{u}(\mathbf{x}))}\frac{\partial \mathbf{h}(\mathbf{u}(\mathbf{x}))}{\partial \mathbf{u}(\mathbf{x})}\frac{\partial \mathbf{u}(\mathbf{x})}{\partial \mathbf{w}_i} = \frac{1}{\sqrt{md}}\langle \mathbf{c}_i, \mathbf{h}'(u_i(\mathbf{x}))\rangle \sigma'\big(\frac{1}{\sqrt{d}}\mathbf{w}_i^\top \mathbf{h}(\mathbf{x})\big)\mathbf{h}(\mathbf{x})^\top.$$

Since $\mathbf{W} = \mathrm{Vec}\big(\{\mathbf{w}_i\}_{i=1}^m\big) \in \mathbb{R}^{mpd}$ is the vectorization of $\mathbf{W}$, it then holds that

$$\nabla f_{\mathbf{W}}(\mathbf{x}) = \frac{1}{\sqrt{md}}\mathrm{Vec}\Big(\big\{\sigma'\big(\frac{1}{\sqrt{d}}\mathbf{w}_i^\top \mathbf{h}(\mathbf{x})\big)\langle \mathbf{c}_i, \mathbf{h}'(u_i(\mathbf{x}))\rangle \mathbf{h}(\mathbf{x})\big\}_{i=1}^m\Big) \in \mathbb{R}^{mpd}.$$

Hence, we get

$$\begin{aligned}
\|\nabla f_{\mathbf{W}}(\mathbf{x})\|_2 &= \frac{1}{\sqrt{md}}\Big(\sum_{i=1}^m \sigma'\big(\frac{1}{\sqrt{d}}\mathbf{w}_i^\top \mathbf{h}(\mathbf{x})\big)^2 \big|\langle \mathbf{c}_i, \mathbf{h}'(u_i(\mathbf{x}))\rangle\big|^2 \|\mathbf{h}(\mathbf{x})\|_2^2\Big)^{\frac{1}{2}} \\
&\le B_\sigma' B_b \sqrt{\frac{p}{m}}\Big(\sum_{i=1}^m \big|\langle \mathbf{c}_i, \mathbf{h}'(u_i(\mathbf{x}))\rangle\big|^2\Big)^{\frac{1}{2}} \\
&\le B_\sigma' B_b B_b' \sqrt{\frac{p^2}{m}}\Big(\sum_{i=1}^m \|\mathbf{c}_i\|_2^2\Big)^{\frac{1}{2}} \le B_\sigma' B_b B_b'\, p\left(4\sqrt{p} + 2\sqrt{\frac{\log(\frac{2}{\delta})}{m}}\right),
\end{aligned} \tag{3}$$

where the last inequality used equation 1. The first part of the lemma is proved.

Now, we turn to estimate the Hessian of $f_{\mathbf{W}}(\mathbf{x})$, i.e.,

$$\nabla^2 f_{\mathbf{W}}(\mathbf{x}) = \begin{bmatrix} \partial_{\mathbf{w}_1}^2 f_{\mathbf{W}}(\mathbf{x}) & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \partial_{\mathbf{w}_m}^2 f_{\mathbf{W}}(\mathbf{x}) \end{bmatrix} \in \mathbb{R}^{mpd\times mpd},$$

where

$$\begin{aligned}
\partial_{\mathbf{w}_i}^2 f_{\mathbf{W}}(\mathbf{x}) =& \frac{1}{d\sqrt{m}}\Big(\sigma''\big(\frac{1}{\sqrt{d}}\mathbf{w}_i^\top \mathbf{h}(\mathbf{x})\big)\big\langle \mathbf{c}_i, \mathbf{h}'\big(\sigma\big(\frac{1}{\sqrt{d}}\mathbf{w}_i^\top \mathbf{h}(\mathbf{x})\big)\big)\big\rangle \\
&+ \big(\sigma'\big(\frac{1}{\sqrt{d}}\mathbf{w}_i^\top \mathbf{h}(\mathbf{x})\big)^2\big\langle \mathbf{c}_i, \mathbf{h}''\big(\sigma\big(\frac{1}{\sqrt{d}}\mathbf{w}_i^\top \mathbf{h}(\mathbf{x})\big)\big)\big\rangle\Big)\mathbf{h}(\mathbf{x})\mathbf{h}(\mathbf{x})^\top \in \mathbb{R}^{pd\times pd}.
\end{aligned}$$

We rewrite $\nabla^2 f_{\mathbf{W}}(\mathbf{x})$ as

$$\partial_{\mathbf{w}_i}^2 f_{\mathbf{W}}(\mathbf{x}) = \frac{1}{d\sqrt{m}}\langle \mathbf{c}_i, \mathbf{v}_i\rangle \mathbf{h}(\mathbf{x})\mathbf{h}(\mathbf{x})^\top$$

with $\mathbf{v}_i = \sigma''\big(\frac{1}{\sqrt{d}}\mathbf{w}_i^\top \mathbf{h}(\mathbf{x})\big)\mathbf{h}'\big(\sigma\big(\frac{1}{\sqrt{d}}\mathbf{w}_i^\top \mathbf{h}(\mathbf{x})\big)\big) + \sigma'\big(\frac{1}{\sqrt{d}}\mathbf{w}_i^\top \mathbf{h}(\mathbf{x})\big)^2 \mathbf{h}''\big(\sigma\big(\frac{1}{\sqrt{d}}\mathbf{w}_i^\top \mathbf{h}(\mathbf{x})\big)\big)$.

According to Assumption 1, we can control $\|\mathbf{v}_i\|_2$ as

$$\begin{aligned}\|\mathbf{v}_i\|_2 &= \big\|\sigma''\big(\frac{1}{\sqrt{d}}\mathbf{w}_i^\top \mathbf{h}(\mathbf{x})\big)\,\mathbf{h}'\big(\sigma\big(\frac{1}{\sqrt{d}}\mathbf{w}_i^\top \mathbf{h}(\mathbf{x})\big)\big) + \sigma'\big(\frac{1}{\sqrt{d}}\mathbf{w}_i^\top \mathbf{h}(\mathbf{x})\big)^2 \mathbf{h}''\big(\sigma\big(\frac{1}{\sqrt{d}}\mathbf{w}_i^\top \mathbf{h}(\mathbf{x})\big)\big)\big\|_2 \\ &\le \|\sigma''\|_\infty \sqrt{p}\|b'\|_\infty + \|\sigma'\|_\infty^2 \sqrt{p}\|b''\|_\infty \le \sqrt{p}(B_\sigma'' B_b' + B_\sigma'^2 B_b'').\end{aligned}$$

Combining the estimate of $\|\mathbf{v}_i\|_2$ and the fact that $\nabla^2 f_{\mathbf{W}}(\mathbf{x})$ is a block diagonal matrix, we can use Lemma 3 to get

$$\begin{aligned}\big\|\nabla^2 f_{\mathbf{W}}(\mathbf{x})\big\|_2 &= \max_{i\in[m]} \big\|\nabla^2_{\mathbf{w}_i} f_{\mathbf{W}}(\mathbf{x})\big\|_2 = \max_{i\in[m]} \sup_{\|\mathbf{a}\|_2=1} \big|\mathbf{a}^\top \partial^2_{\mathbf{w}_i} f_{\mathbf{W}}(\mathbf{x})\,\mathbf{a}\big| \\ &= \frac{1}{d\sqrt{m}} \max_{i\in[m]} \sup_{\|\mathbf{a}\|_2=1} \big|\langle \mathbf{c}_i, \mathbf{v}_i\rangle \langle \mathbf{h}(\mathbf{x}), \mathbf{a}\rangle^2\big| \le \frac{1}{d\sqrt{m}} \max_{i\in[m]} \big\|\mathbf{c}_i\big\|_2 \big\|\mathbf{v}_i\big\|_2 \big\|\mathbf{h}(\mathbf{x})\big\|_2^2 \\ &\le B_b^2 \frac{p}{\sqrt{m}}\Big(4\sqrt{p} + 2\sqrt{\log(\frac{2m}{\delta})}\Big) \max_{i\in[m]} \big\|\mathbf{v}_i\big\|_2 \\ &\le B_b^2 (B_\sigma'' B_b' + B_\sigma'^2 B_b'') \frac{p^{\frac{3}{2}}}{\sqrt{m}}\Big(4\sqrt{p} + 2\sqrt{\log(\frac{2m}{\delta})}\Big), \qquad (4)\end{aligned}$$

where the second equality used the fact that $\partial^2_{\mathbf{w}_i} f_{\mathbf{W}}(\mathbf{x})$ is symmetric, the first inequality used Cauchy-Schwarz inequality, the second inequality used Assumption 1 that $\sup_{t\in\mathbb{R}} |b(t)| \le B_b$ and equation 1. The proof is complete. □

The following lemma shows that the largest and smallest eigenvalues of $\nabla^2 \ell(y f_{\mathbf{W}}(\mathbf{x}))$ admit well-controlled upper and lower bounds, respectively. As a consequence, the loss function $\ell(y f_{\mathbf{W}}(\mathbf{x}))$ is weakly convex and smooth with respect to $\mathbf{W}$.

Lemma 11 (Smoothness and Curvature)**.** *Let $\delta \in (0,1)$. Suppose Assumption 1 and the event $\mathcal{E}_\delta$ occurs. Assume $m \gtrsim \log(m/\delta)$. It holds for any $\mathbf{W}$ and any data point $(\mathbf{x}, y) \in \mathcal{Z}$, that*

$$\lambda_{\min}\big(\nabla^2 \ell(y f_{\mathbf{W}}(\mathbf{x}))\big) \ge -\frac{C_{\sigma,b}\, p^{\frac{3}{2}}\big(\sqrt{\log(\frac{m}{\delta})} + \sqrt{p}\big)}{\sqrt{m}} \ell(y f_{\mathbf{W}}(\mathbf{x})),$$

$$\lambda_{\max}\big(\nabla^2 \ell(y f_{\mathbf{W}}(\mathbf{x}))\big) \le C_{\sigma,b}\, p^3$$

and

$$\big\|\nabla \mathcal{L}_S(\mathbf{W})\big\|_2^2 \le 4 C_{\sigma,b}\, p^3 \mathcal{L}_S(\mathbf{W}).$$

Proof. From the chain rule, the gradient of loss is given as

$$\nabla \ell(y f_{\mathbf{W}}(\mathbf{x})) = \ell'(y f_{\mathbf{W}}(\mathbf{x})) y \nabla f_{\mathbf{W}}(\mathbf{x}).$$

Note that

$$\nabla^2 \ell(y f_{\mathbf{W}}(\mathbf{x})) = \ell''(y f_{\mathbf{W}}(\mathbf{x})) y^2 \nabla f_{\mathbf{W}}(\mathbf{x}) \nabla f_{\mathbf{W}}(\mathbf{x})^\top + \ell'(y f_{\mathbf{W}}(\mathbf{x})) y \nabla^2 f_{\mathbf{W}}(\mathbf{x}).$$

Since ℓ is convex, $\ell''(a) \ge 0$ for all $a \in \mathbb{R}$. Then, $\ell''(y f_{\mathbf{W}}(\mathbf{x})) \nabla f_{\mathbf{W}}(\mathbf{x}) \nabla f_{\mathbf{W}}(\mathbf{x})^\top$ is a PSD matrix. By further noting that $|\ell'(a)| \le 1$ and $|\ell''(a)| \le 1/4$ for all $a \in \mathbb{R}$, we have

$$-|\ell'(y f_{\mathbf{W}}(\mathbf{x}))| \big\|\nabla^2 f_{\mathbf{W}}(\mathbf{x})\big\|_2 \le \lambda_{\min}\big(\nabla^2 \ell(y f_{\mathbf{W}}(\mathbf{x}))\big)$$

and

$$\lambda_{\max}\big(\nabla^2 \ell(y f_{\mathbf{W}}(\mathbf{x}))\big) \le \frac{1}{4}\big\|\nabla f_{\mathbf{W}}(\mathbf{x})\big\|_2^2 + \big\|\nabla^2 f_{\mathbf{W}}(\mathbf{x})\big\|_2.$$

Plugging the estimates of $\big\|\nabla f_{\mathbf{W}}(\mathbf{x})\big\|_2$ and $\big\|\nabla^2 f_{\mathbf{W}}(\mathbf{x})\big\|_2$ in Lemma 10 back and noting that $|\ell'(y f_{\mathbf{W}}(\mathbf{x}))| \le \ell(y f_{\mathbf{W}}(\mathbf{x}))$, we know

$$\lambda_{\min}\big(\nabla^2 \ell(y f_{\mathbf{W}}(\mathbf{x}))\big) \ge -\frac{C_{\sigma,b}\, p^{\frac{3}{2}}\big(\sqrt{\log(\frac{m}{\delta})} + \sqrt{p}\big)}{\sqrt{m}} \ell(y f_{\mathbf{W}}(\mathbf{x})),$$

and

$$\lambda_{\max}\big(\nabla^2 \ell(y f_{\mathbf{W}}(\mathbf{x}))\big) \le C_{\sigma,b}\Big[p^2\Big(p + \frac{\log(\frac{1}{\delta})}{m}\Big) + \frac{p^{\frac{3}{2}}\big(\sqrt{p} + \sqrt{\log(\frac{m}{\delta})}\big)}{\sqrt{m}}\Big].$$

Furthermore, if $m \gtrsim \log(m/\delta)$, we have

$$\lambda_{\max}\big(\nabla^2 \ell(y f_{\mathbf{W}}(\mathbf{x}))\big) \le C_{\sigma,b}\, p^3,$$

which completes the first two inequalities of the lemma.

Note $C_{\sigma,b}\, p^3$ is an upper bound for the term $\sup_{\mathbf{x}\in\mathcal{X}} \frac{1}{4}\|\nabla f_{\mathbf{W}}(\mathbf{x})\|_2^2 + \|\nabla^2 f_{\mathbf{W}}(\mathbf{x})\|_2$ and $|y| = 1$ for any $y \in \mathcal{Y}$. Then, it holds for any $\mathbf{W}$ that

$$\begin{aligned}\big\|\nabla \mathcal{L}_S(\mathbf{W})\big\|_2^2 &= \Big\|\frac{1}{n}\sum_{i=1}^n \ell'(y_i f_{\mathbf{W}}(\mathbf{x}_i)) y_i \nabla f_{\mathbf{W}}(\mathbf{x}_i)\Big\|_2^2 \le \Big|\frac{1}{n}\sum_{i=1}^n \ell'(y_i f_{\mathbf{W}}(\mathbf{x}_i))\Big|^2 \sup_{i\in[n]} \big\|\nabla f_{\mathbf{W}}(\mathbf{x}_i)\big\|_2^2 \\ &\le 4C_{\sigma,b}\, p^3 \Big|\frac{1}{n}\sum_{i=1}^n \ell(y_i f_{\mathbf{W}}(\mathbf{x}_i))\Big| = 4C_{\sigma,b}\, p^3 \mathcal{L}_S(\mathbf{W}),\end{aligned}$$

where the second inequality used the self-boundedness property and $|\ell'(\cdot)| \le 1$ of the logistic loss. This completes the proof of the lemma. □

C Proofs for optimization of DP-SGD with correlated noise

Throughout the paper, we use $\mathbf{W}$ to denote the vectorized parameter $(\mathbf{w}_1^\top, \dots, \mathbf{w}_m^\top)^\top \in \mathbb{R}^{mdp}$, unless otherwise specified. We occasionally use the same symbol for the matrix representation in the definition of $f_{\mathbf{W}}(\mathbf{x}) = \frac{1}{\sqrt{m}}\mathbf{c}^\top \mathbf{h}\big(\sigma\big(\frac{1}{\sqrt{d}}\mathbf{W}\mathbf{h}(\mathbf{x})\big)\big)$, when the meaning is clear from context.

Moreover, under $\mathcal{E}_\delta$, equation 3 gives

$$\|\nabla f_{\mathbf{W}}(\mathbf{x})\|_2 \le G_\delta := B'_\sigma B_b B'_b\, p\Big(4\sqrt{p} + 2\sqrt{\frac{\log(2/\delta)}{m}}\Big), \qquad \forall \mathbf{x} \in \mathcal{X},\ \forall \mathbf{W}.$$

Since $|\ell'| \le 1$, this further yields a uniform upper bound on the per-example loss gradient:

$$\|\nabla \ell(y f_{\mathbf{W}}(\mathbf{x}))\|_2 \le G_\delta, \qquad \forall (\mathbf{x}, y) \in \mathcal{Z},\ \forall \mathbf{W}. \tag{5}$$

Under $m \gtrsim \log(m/\delta)$, by enlarging $C_{\sigma,b}$ if necessary, we further have $G_\delta^2 \le 4C_{\sigma,b}p^3$. For convenience, denote $\beta := C_{\sigma,b}p^3$.

For an arbitrary clipping threshold $C_{\mathrm{clip}} > 0$, define

$$\rho_{\mathrm{clip}} := \min\Big\{1, \frac{C_{\mathrm{clip}}}{G_\delta}\Big\} \in (0,1], \quad G_{\mathrm{clip}} := \min\{C_{\mathrm{clip}}, G_\delta\} = \rho_{\mathrm{clip}} G_\delta, \quad \beta_{\mathrm{clip}} := \rho_{\mathrm{clip}}\beta.$$

For each $(\mathbf{x}, y) \in \mathcal{Z}$ and $\mathbf{W}$, let

$$\alpha(\mathbf{W}; \mathbf{x}, y) := \min\Big\{1, \frac{C_{\mathrm{clip}}}{\|\nabla \ell(y f_{\mathbf{W}}(\mathbf{x}))\|_2}\Big\},$$

with the convention $\alpha(\mathbf{W}; \mathbf{x}, y) = 1$ when $\nabla \ell(y f_{\mathbf{W}}(\mathbf{x})) = 0$. Then, on $\mathcal{E}_\delta$,

$$\rho_{\mathrm{clip}} \le \alpha(\mathbf{W}; \mathbf{x}, y) \le 1, \qquad \big\|\alpha(\mathbf{W}; \mathbf{x}, y)\nabla \ell(y f_{\mathbf{W}}(\mathbf{x}))\big\|_2 \le G_{\mathrm{clip}}.$$

In particular, when $C_{\mathrm{clip}} = G_\delta$, we have $\rho_{\mathrm{clip}} = 1$ and recover the inactive-clipping setting as a special case.

The key idea of the proof is to combine the cancellation structure of the correlated perturbations with the localization argument for KANs. Following the proof sketch in Section 4.3, we organize the detailed proof into the following six steps.

Step 1: Reduce to an auxiliary unprojected dynamics.

The correlated perturbations break the conditional-centering argument, and the projection in Algorithm 1 further obstructs the cross-iteration cancellation structure of the noise. To analyze this case, we introduce an auxiliary unprojected sequence $\{\widetilde{\mathbf{W}}_t\}_{t=0}^T$, defined recursively by

$$\widetilde{\mathbf{W}}_0 = \mathbf{W}_0, \qquad \widetilde{\mathbf{W}}_t = \widetilde{\mathbf{W}}_{t-1} - \eta\Big(\frac{1}{B}\sum_{i\in\mathcal{B}_t} g_{t,i}^{\mathrm{aux}} + \frac{C_{\mathrm{clip}}}{B}\xi_t\Big),$$

where, for each $t \in [T]$ and $i \in [n]$, $\alpha_{t,i} := \min\big\{1, \frac{C_{\mathrm{clip}}}{\|\nabla \ell(y_i f_{\widetilde{\mathbf{W}}_{t-1}}(x_i))\|_2}\big\}$, with the convention $\alpha_{t,i} = 1$ when the gradient vanishes, and $g_{t,i}^{\mathrm{aux}} := \alpha_{t,i}\nabla\ell\big(y_i f_{\widetilde{\mathbf{W}}_{t-1}}(x_i)\big)$.

Define the full-batch clipped gradient at the auxiliary iterate by $\bar{g}_t := \frac{1}{n}\sum_{i=1}^n g_{t,i}^{\mathrm{aux}}$, and the corresponding mini-batch fluctuation by

$$\widetilde{\Delta}_t := \frac{1}{B}\sum_{i\in\mathcal{B}_t} g_{t,i}^{\mathrm{aux}} - \bar{g}_t,$$

which measures the discrepancy between the mini-batch gradient and the full empirical gradient. Thus, conditioned on the history up to iteration $t-1$, $\mathbb{E}\big[\widetilde{\Delta}_t \mid \mathcal{F}_{t-1}\big] = 0$.

Starting from $\widetilde{\mathbf{W}}_0 = \mathbf{W}_0$ and $Z_0 = \mathbf{0}$, the auxiliary unprojected iterate can therefore be rewritten as

$$\widetilde{\mathbf{W}}_t = \widetilde{\mathbf{W}}_{t-1} - \eta\big(\bar{g}_t + \widetilde{\Delta}_t + \frac{C_{\mathrm{clip}}\kappa}{B}(Z_t - \lambda Z_{t-1})\big), \qquad t \in [T].$$

STEP 2: INTRODUCE THE SHIFTED ITERATE.

To obtain the high-probability estimate above, one would naturally try to apply a comparator recursion to the auxiliary trajectory. However, the correlated perturbation $Z_t - \lambda Z_{t-1}$ is not conditionally centered with respect to the natural filtration. The term involving Z_{t-1} is correlated with the current iterate and creates a first-order drift term.

To handle this obstruction, we define the shifted iterate

$$\mathbf{U}_t = \widetilde{\mathbf{W}}_t + \eta c_{\mathrm{priv}} Z_t, \qquad t = 0, 1, \ldots, T.$$

Here, $c_{\mathrm{priv}} = \frac{C_{\mathrm{clip}}\kappa}{B}$. The role of $\mathbf{U}_t$ is to absorb the current Gaussian perturbation into the state variable, which allows us to rewrite the correlated-noise dynamics in a form that is more amenable to one-step analysis.

STEP 3: DERIVE A SHIFTED POTENTIAL RECURSION USING THE LOCAL KAN GEOMETRY.

We now start the main proof route for $\lambda > 0$. The key step is to absorb the current Gaussian noise into a shifted iterate so that the correlated perturbation can be handled at the *single-step* level.

Lemma 12 (Shifted one-step identity)**.** *For any comparator $\mathbf{W}^* \in \mathbb{R}^{mdp}$ and any $t \in [T-1]$,*

$$\begin{aligned}\|\mathbf{U}_t - \mathbf{W}^*\|_2^2 &= \|\mathbf{U}_{t-1} - \mathbf{W}^*\|_2^2 - 2\eta\big\langle \bar{g}_t, \widetilde{\mathbf{W}}_{t-1} - \mathbf{W}^*\big\rangle - 2\eta\big\langle \widetilde{\Delta}_t, \mathbf{U}_{t-1} - \mathbf{W}^*\big\rangle \\ &\quad - 2(1-\lambda)\eta c_{\mathrm{priv}}\big\langle Z_{t-1}, \mathbf{U}_{t-1} - \mathbf{W}^*\big\rangle + \eta^2\|\bar{g}_t + \widetilde{\Delta}_t\|_2^2 + (1-\lambda)^2\eta^2 c_{\mathrm{priv}}^2\|Z_{t-1}\|_2^2 \\ &\quad + 2(1-\lambda)\eta^2 c_{\mathrm{priv}}\big\langle \widetilde{\Delta}_t, Z_{t-1}\big\rangle - 2\lambda\eta^2 c_{\mathrm{priv}}\big\langle \bar{g}_t, Z_{t-1}\big\rangle. \end{aligned} \tag{6}$$

Proof. By the definitions of $\widetilde{\mathbf{W}}_t$ and $\mathbf{U}_t$, we have

$$\begin{aligned}\mathbf{U}_t &= \widetilde{\mathbf{W}}_t + \eta c_{\mathrm{priv}} Z_t = \widetilde{\mathbf{W}}_{t-1} - \eta\big(\bar{g}_t + \widetilde{\Delta}_t + c_{\mathrm{priv}}(Z_t - \lambda Z_{t-1})\big) + \eta c_{\mathrm{priv}} Z_t \\ &= \widetilde{\mathbf{W}}_{t-1} - \eta(\bar{g}_t + \widetilde{\Delta}_t) + \eta\lambda c_{\mathrm{priv}} Z_{t-1}.\end{aligned}$$

Plugging $\widetilde{\mathbf{W}}_{t-1} = \mathbf{U}_{t-1} - \eta c_{\mathrm{priv}} Z_{t-1}$ into the above equality, we get

$$\begin{aligned}\mathbf{U}_t &= \mathbf{U}_{t-1} - \eta c_{\mathrm{priv}} Z_{t-1} - \eta(\bar{g}_t + \widetilde{\Delta}_t) + \eta\lambda c_{\mathrm{priv}} Z_{t-1} \\ &= \mathbf{U}_{t-1} - \eta\big(\bar{g}_t + \widetilde{\Delta}_t + (1-\lambda)c_{\mathrm{priv}} Z_{t-1}\big).\end{aligned}$$

Hence,

$$\mathbf{U}_t - \mathbf{W}^* = \mathbf{U}_{t-1} - \mathbf{W}^* - \eta\big(\bar{g}_t + \widetilde{\Delta}_t + (1-\lambda)c_{\mathrm{priv}} Z_{t-1}\big).$$

It then follows

$$\begin{aligned}
&\|\mathbf{U}_t - \mathbf{W}^*\|_2^2 \\
&= \|\mathbf{U}_{t-1} - \mathbf{W}^*\|_2^2 - 2\eta\big\langle \bar{g}_t + \widetilde{\Delta}_t + (1-\lambda)c_{\mathrm{priv}} Z_{t-1}, \mathbf{U}_{t-1} - \mathbf{W}^*\big\rangle + \eta^2 \big\|\bar{g}_t + \widetilde{\Delta}_t + (1-\lambda)c_{\mathrm{priv}} Z_{t-1}\big\|_2^2 \\
&= \|\mathbf{U}_{t-1} - \mathbf{W}^*\|_2^2 - 2\eta\big\langle \bar{g}_t, \widetilde{\mathbf{W}}_{t-1} - \mathbf{W}^*\big\rangle - 2\eta\big\langle \widetilde{\Delta}_t, \mathbf{U}_{t-1} - \mathbf{W}^*\big\rangle - 2(1-\lambda)\eta c_{\mathrm{priv}}\big\langle Z_{t-1}, \mathbf{U}_{t-1} - \mathbf{W}^*\big\rangle \\
&\quad + \eta^2 \|\bar{g}_t + \widetilde{\Delta}_t\|_2^2 + (1-\lambda)^2 \eta^2 c_{\mathrm{priv}}^2 \|Z_{t-1}\|_2^2 + 2(1-\lambda)\eta^2 c_{\mathrm{priv}}\big\langle \widetilde{\Delta}_t, Z_{t-1}\big\rangle - 2\lambda\eta^2 c_{\mathrm{priv}}\big\langle \bar{g}_t, Z_{t-1}\big\rangle,
\end{aligned}$$

where we have used $-2\eta\big\langle \bar{g}_t, \mathbf{U}_{t-1} - \mathbf{W}^*\big\rangle = -2\eta\big\langle \bar{g}_t, \widetilde{\mathbf{W}}_{t-1} - \mathbf{W}^*\big\rangle - 2\eta^2 c_{\mathrm{priv}}\big\langle \bar{g}_t, Z_{t-1}\big\rangle$ implied by $\mathbf{U}_{t-1} = \widetilde{\mathbf{W}}_{t-1} + \eta c_{\mathrm{priv}} Z_{t-1}$. The proof is complete. □

For a fixed comparator $\mathbf{W}^* \in \mathbb{R}^{mdp}$ and a shifted radius $\bar{R} > 0$, define

$$A_{\bar{R}}^U(t) = \Big\{ \max_{0 \le s \le t} \|\mathbf{U}_s - \mathbf{W}^*\|_2 \le \bar{R} \Big\} \qquad \text{and} \qquad A_{\bar{R}}^U = A_{\bar{R}}^U(T).$$

For $z > 0$ and $V_Z > 0$, define the Gaussian event

$$G_Z(z, V_Z) = \Big\{ \max_{0 \le t \le T-1} \|Z_t\|_2 \le z \quad \text{and} \quad \sum_{t=1}^{T} \|Z_{t-1}\|_2^2 \le V_Z \Big\}.$$

For $V_\Delta > 0$, define the stopped mini-batch quadratic event

$$G_{\Delta^2}(\bar{R}, V_\Delta) = \Big\{ \sum_{t=1}^{T} \|\widetilde{\Delta}_t\|_2^2 \mathbf{1}_{A_{\bar{R}}^U(t-1)} \le V_\Delta \Big\}.$$

Define the stopped potential fluctuation process

$$\begin{aligned}
M_k = \sum_{t=1}^{k} \mathbf{1}_{A_{\bar{R}}^U(t-1)} \Big[&- 2\eta\big\langle \widetilde{\Delta}_t, \mathbf{U}_{t-1} - \mathbf{W}^*\big\rangle - 2(1-\lambda)\eta c_{\mathrm{priv}}\big\langle Z_{t-1}, \mathbf{U}_{t-1} - \mathbf{W}^*\big\rangle \\
&+ 2(1-\lambda)\eta^2 c_{\mathrm{priv}}\big\langle \widetilde{\Delta}_t, Z_{t-1}\big\rangle \Big],
\end{aligned}$$

and define

$$G_{\mathrm{pot}}(\bar{R}, M) := \Big\{ \max_{1 \le k \le T} M_k \le M \Big\}.$$

Finally, define the *shifted good event*

$$\mathcal{E}_{\mathrm{good}} = G_Z(z, V_Z) \cap G_{\Delta^2}(\bar{R}, V_\Delta) \cap G_{\mathrm{pot}}(\bar{R}, M)$$

and the frozen weighted empirical loss

$$\mathcal{L}_{S,t}^{\alpha}(\mathbf{W}) := \frac{1}{n} \sum_{i=1}^{n} \alpha_{t,i}\, \ell(y_i f_{\mathbf{W}}(x_i)),$$

where the coefficients $\{\alpha_{t,i}\}_{i=1}^n$ are those defined in Step 1 at the auxiliary iterate $\widetilde{\mathbf{W}}_{t-1}$ and are treated as fixed with respect to $\mathbf{W}$.

Lemma 13 (Localized comparator under shifted localization)**.** *Suppose Assumption 1 holds and the event $\mathcal{E}_\delta$ occurs. Let $\mathbf{W}^* \in \mathbb{R}^{mdp}$ be fixed. On the events $A_{\bar{R}}^U(t-1)$ and $G_Z(z, V_Z)$, assume*

$$m \gtrsim C_{\sigma,b}^2 p^3 \big(\log(m/\delta) + p\big)\big(\bar{R} + \eta c_{\mathrm{priv}} z\big)^4. \tag{7}$$

Then

$$\frac{2}{3}\, \mathcal{L}_{S,t}^{\alpha}(\widetilde{\mathbf{W}}_{t-1}) \le \big\langle \bar{g}_t, \widetilde{\mathbf{W}}_{t-1} - \mathbf{W}^*\big\rangle + \frac{4}{3}\, \mathcal{L}_S(\mathbf{W}^*).$$

Furthermore,

$$\frac{2\rho_{\mathrm{clip}}}{3}\, \mathcal{L}_S(\widetilde{\mathbf{W}}_{t-1}) \le \big\langle \bar{g}_t, \widetilde{\mathbf{W}}_{t-1} - \mathbf{W}^*\big\rangle + \frac{4}{3}\, \mathcal{L}_S(\mathbf{W}^*).$$

Proof. First, by Lemma 11, for every data point $(\mathbf{x}_i, y_i)$ and every $\mathbf{V} \in \mathbb{R}^{mdp}$,

$$\lambda_{\min}\left(\nabla^2 \ell(y_i f_{\mathbf{V}}(\mathbf{x}_i))\right) \geq -\nu\, \ell(y_i f_{\mathbf{V}}(\mathbf{x}_i))$$

with $\nu = \frac{C_{\sigma,b} p^{\frac{3}{2}}\left(\sqrt{\log(m/\delta)}+\sqrt{p}\right)}{\sqrt{m}}$. Since the coefficients $\{\alpha_{t,i}\}_{i=1}^n$ are fixed and nonnegative,

$$\lambda_{\min}\left(\nabla^2 \mathcal{L}^{\alpha}_{S,t}(\mathbf{V})\right) \geq -\nu\, \mathcal{L}^{\alpha}_{S,t}(\mathbf{V}).$$

Moreover,

$$\nabla \mathcal{L}^{\alpha}_{S,t}(\widetilde{\mathbf{W}}_{t-1}) = \frac{1}{n}\sum_{i=1}^{n} \alpha_{t,i} \nabla \ell(y_i f_{\widetilde{\mathbf{W}}_{t-1}}(x_i)) = \bar{g}_t.$$

Since $\widetilde{\mathbf{W}}_{t-1} = \mathbf{U}_{t-1} - \eta c_{\text{priv}} Z_{t-1}$, we have $\|\widetilde{\mathbf{W}}_{t-1} - \mathbf{W}^*\|_2 \leq \bar{R} + \eta c_{\text{priv}} z$. Taking $D = \bar{R} + \eta c_{\text{priv}} z$, the width condition guarantees $\nu D^2 \leq \frac{1}{2}$.

Recall that we defined the line segment $[\widetilde{\mathbf{W}}_{t-1}, \mathbf{W}^*] = \{\alpha \widetilde{\mathbf{W}}_{t-1} + (1-\alpha)\mathbf{W}^* : \alpha \in [0,1]\}$. Applying Taylor's theorem, there exists a point $\mathbf{V} \in [\widetilde{\mathbf{W}}_{t-1}, \mathbf{W}^*]$ such that

$$\mathcal{L}^{\alpha}_{S,t}(\mathbf{W}^*) \geq \mathcal{L}^{\alpha}_{S,t}(\widetilde{\mathbf{W}}_{t-1}) + \left\langle \nabla \mathcal{L}^{\alpha}_{S,t}(\widetilde{\mathbf{W}}_{t-1}), \mathbf{W}^* - \widetilde{\mathbf{W}}_{t-1} \right\rangle - \frac{\nu}{2} \mathcal{L}^{\alpha}_{S,t}(\mathbf{V}) \|\widetilde{\mathbf{W}}_{t-1} - \mathbf{W}^*\|_2^2.$$

Hence,

$$\mathcal{L}^{\alpha}_{S,t}(\widetilde{\mathbf{W}}_{t-1}) - \mathcal{L}^{\alpha}_{S,t}(\mathbf{W}^*) \leq \left\langle \bar{g}_t, \widetilde{\mathbf{W}}_{t-1} - \mathbf{W}^* \right\rangle + \frac{\nu D^2}{2} \max_{\mathbf{V} \in [\widetilde{\mathbf{W}}_{t-1}, \mathbf{W}^*]} \mathcal{L}^{\alpha}_{S,t}(\mathbf{V}). \tag{8}$$

On the other hand, Lemma 8 with $\lambda_{\min}\left(\nabla^2 \mathcal{L}^{\alpha}_{S,t}(\mathbf{V})\right) \geq -\nu \mathcal{L}^{\alpha}_{S,t}(\mathbf{V})$ gives

$$\max_{\mathbf{V} \in [\widetilde{\mathbf{W}}_{t-1}, \mathbf{W}^*]} \mathcal{L}^{\alpha}_{S,t}(\mathbf{V}) \leq \rho \max\{\mathcal{L}^{\alpha}_{S,t}(\widetilde{\mathbf{W}}_{t-1}), \mathcal{L}^{\alpha}_{S,t}(\mathbf{W}^*)\},$$

where $\rho = \left(1 - \frac{\nu D^2}{2}\right)^{-1}$. Note $\nu D^2 \leq 1/2$, then $\frac{\nu D^2}{2}\rho = \frac{\nu D^2}{2 - \nu D^2} \leq \frac{1}{3}$. Therefore, combining the above inequality with equation 8 yields

$$\mathcal{L}^{\alpha}_{S,t}(\widetilde{\mathbf{W}}_{t-1}) - \mathcal{L}^{\alpha}_{S,t}(\mathbf{W}^*) \leq \left\langle \nabla \mathcal{L}^{\alpha}_{S,t}(\widetilde{\mathbf{W}}_{t-1}), \widetilde{\mathbf{W}}_{t-1} - \mathbf{W}^* \right\rangle + \frac{1}{3}\left(\mathcal{L}^{\alpha}_{S,t}(\widetilde{\mathbf{W}}_{t-1}) + \mathcal{L}^{\alpha}_{S,t}(\mathbf{W}^*)\right).$$

Rearranging gives

$$\frac{2}{3}\, \mathcal{L}^{\alpha}_{S,t}(\widetilde{\mathbf{W}}_{t-1}) \leq \left\langle \nabla \mathcal{L}^{\alpha}_{S,t}(\widetilde{\mathbf{W}}_{t-1}), \widetilde{\mathbf{W}}_{t-1} - \mathbf{W}^* \right\rangle + \frac{4}{3}\, \mathcal{L}^{\alpha}_{S,t}(\mathbf{W}^*). \tag{9}$$

Since $\alpha_{t,i} \leq 1$ and the loss is nonnegative, it holds $\mathcal{L}^{\alpha}_{S,t}(\mathbf{W}^*) \leq \mathcal{L}_S(\mathbf{W}^*)$. Moreover, on $\mathcal{E}_\delta$, $\alpha_{t,i} \geq \rho_{\text{clip}}$ for every $i \in [n]$, and hence $\mathcal{L}^{\alpha}_{S,t}(\widetilde{\mathbf{W}}_{t-1}) \geq \rho_{\text{clip}}\, \mathcal{L}_S(\widetilde{\mathbf{W}}_{t-1})$. Combining the above inequalities proves

$$\frac{2\rho_{\text{clip}}}{3} \mathcal{L}_S(\widetilde{\mathbf{W}}_{t-1}) \leq \left\langle \bar{g}_t, \widetilde{\mathbf{W}}_{t-1} - \mathbf{W}^* \right\rangle + \frac{4}{3} \mathcal{L}_S(\mathbf{W}^*).$$

This completes the proof. □

We next show that, on the shifted good event $\mathcal{E}_{\text{good}}$, the shifted iterate remains localized and yields a pathwise empirical loss bound. In particular, the projection becomes inactive under a suitable choice of R_*.

Lemma 14 (Conditional shifted-potential bootstrap and loss bound)**.** *Suppose $0 < \lambda < 1$, Assumption 1 holds and the event $\mathcal{E}_\delta$ occurs. Let $\beta = C_{\sigma,b} p^3$ and $\beta_{\text{clip}} = \rho_{\text{clip}} \beta$. Fix a comparator $\mathbf{W}^*$ and a shifted radius $\bar{R} > 0$, and assume $\eta \leq \frac{1}{12\beta_{\text{clip}}}$, equation 7, and*

$$\|\mathbf{W}_0 - \mathbf{W}^*\|_2^2 + \frac{8}{3}\eta T \mathcal{L}_S(\mathbf{W}^*) + M + 2\eta^2 V_\Delta + \left((1-\lambda)^2 \eta^2 c_{\text{priv}}^2 + 12\beta_{\text{clip}} \lambda^2 \eta^3 c_{\text{priv}}^2\right) V_Z \leq \bar{R}^2. \tag{10}$$

Then, on the good event $\mathcal{E}_{\text{good}}$, the localization event $A^U_{\bar{R}}$ also occurs and

$$\frac{1}{T}\sum_{t=1}^{T} \mathcal{L}^{\alpha}_{S,t}(\widetilde{\mathbf{W}}_{t-1}) \leq 8\mathcal{L}_S(\mathbf{W}^*) + \frac{3(\|\mathbf{W}_0 - \mathbf{W}^*\|_2^2 + M)}{\eta T} + \frac{6\eta V_\Delta}{T} + 3\left((1-\lambda)^2 \eta + 12\beta_{\text{clip}} \lambda^2 \eta^2\right) \frac{c_{\text{priv}}^2 V_Z}{T}.$$

If we further assume $R_* \geq \bar{R} + \|\mathbf{W}^* - \mathbf{W}_0\|_2 + \eta c_{\mathrm{priv}} z$*, then*

$$\frac{1}{T}\sum_{t=1}^{T}\mathcal{L}_S(\mathbf{W}_{t-1}) \leq \frac{1}{\rho_{\mathrm{clip}}}\Big[8\mathcal{L}_S(\mathbf{W}^*) + \frac{3(\|\mathbf{W}_0 - \mathbf{W}^*\|_2^2 + M)}{\eta T} + \frac{6\eta V_\Delta}{T} + 3\Big((1-\lambda)^2\eta + 12\beta_{\mathrm{clip}}\lambda^2\eta^2\Big)c_{\mathrm{priv}}^2\frac{V_Z}{T}\Big].$$

Proof. Fix an outcome in $\mathcal{E}_{\mathrm{good}}$. We first show that the shifted localization event $A_{\bar{R}}^U$ holds. The proof is done by contradiction. Suppose $A_{\bar{R}}^U$ does not hold. Since $\|\mathbf{U}_0 - \mathbf{W}^*\|_2^2 = \|\mathbf{W}_0 - \mathbf{W}^*\|_2^2 \leq \bar{R}^2$ by equation 10, the first exit time

$$\tau = \min\Big\{t \in [T] : \|\mathbf{U}_t - \mathbf{W}^*\|_2 > \bar{R}\Big\}$$

is well-defined. Then $\tau \in [T]$ and $A_{\bar{R}}^U(\tau - 1)$ holds. Hence, for every $1 \leq t \leq \tau$, $\mathbf{1}_{A_{\bar{R}}^U(t-1)} = 1$.

For every $t \leq \tau$, since $A_{\bar{R}}^U(t-1)$ and $G_Z(z, V_Z)$ hold, we have

$$\|\widetilde{\mathbf{W}}_{t-1} - \mathbf{W}^*\|_2 = \|\mathbf{U}_{t-1} - \eta c_{\mathrm{priv}} Z_{t-1} - \mathbf{W}^*\|_2 \leq \bar{R} + \eta c_{\mathrm{priv}} z.$$

By Lemma 12, for every $t \leq \tau$,

$$\begin{aligned}\|\mathbf{U}_t - \mathbf{W}^*\|_2^2 = \|\mathbf{U}_{t-1} - \mathbf{W}^*\|_2^2 &- 2\eta\langle\bar{g}_t, \widetilde{\mathbf{W}}_{t-1} - \mathbf{W}^*\rangle - 2\eta\langle\widetilde{\Delta}_t, \mathbf{U}_{t-1} - \mathbf{W}^*\rangle \\ &- 2(1-\lambda)\eta c_{\mathrm{priv}}\langle Z_{t-1}, \mathbf{U}_{t-1} - \mathbf{W}^*\rangle + \eta^2\|\bar{g}_t + \widetilde{\Delta}_t\|_2^2 + (1-\lambda)^2\eta^2 c_{\mathrm{priv}}^2\|Z_{t-1}\|_2^2 \\ &+ 2(1-\lambda)\eta^2 c_{\mathrm{priv}}\langle\widetilde{\Delta}_t, Z_{t-1}\rangle - 2\lambda\eta^2 c_{\mathrm{priv}}\langle\bar{g}_t, Z_{t-1}\rangle.\end{aligned}$$

According to Lemma 13, it holds

$$-2\eta\langle\bar{g}_t, \widetilde{\mathbf{W}}_{t-1} - \mathbf{W}^*\rangle \leq -\frac{4\eta}{3}\mathcal{L}_{S,t}^{\alpha}(\widetilde{\mathbf{W}}_{t-1}) + \frac{8\eta}{3}\mathcal{L}_S(\mathbf{W}^*).$$

Moreover, by Jensen's inequality,

$$\|\bar{g}_t\|_2^2 = \Big\|\frac{1}{n}\sum_{i=1}^{n}\alpha_{t,i}\nabla\ell\big(y_i f_{\widetilde{\mathbf{W}}_{t-1}}(x_i)\big)\Big\|_2^2 \leq \frac{1}{n}\sum_{i=1}^{n}\alpha_{t,i}^2\Big\|\nabla\ell\big(y_i f_{\widetilde{\mathbf{W}}_{t-1}}(x_i)\big)\Big\|_2^2.$$

For each $i \in [n]$, denote $\nabla\ell_{t,i} := \nabla\ell\big(y_i f_{\widetilde{\mathbf{W}}_{t-1}}(x_i)\big)$. By the definition of $\alpha_{t,i}$ and equation 5,

$$\alpha_{t,i}\|\nabla\ell_{t,i}\|_2 = \min\big\{\|\nabla\ell_{t,i}\|_2, C_{\mathrm{clip}}\big\} \leq G_{\mathrm{clip}}.$$

Hence,

$$\alpha_{t,i}^2\|\nabla\ell_{t,i}\|_2^2 \leq G_{\mathrm{clip}}\alpha_{t,i}\|\nabla\ell_{t,i}\|_2.$$

Moreover, using $|\ell'| \leq \ell$ and equation 3,

$$\|\nabla\ell_{t,i}\|_2 = \big|\ell'\big(y_i f_{\widetilde{\mathbf{W}}_{t-1}}(x_i)\big)\big|\big\|\nabla f_{\widetilde{\mathbf{W}}_{t-1}}(x_i)\big\|_2 \leq G_\delta\,\ell\big(y_i f_{\widetilde{\mathbf{W}}_{t-1}}(x_i)\big).$$

Combining the above estimates, $G_{\mathrm{clip}} = \rho_{\mathrm{clip}}G_\delta$, and $G_\delta^2 \leq 4\beta$, we obtain

$$\|\bar{g}_t\|_2^2 \leq G_{\mathrm{clip}}G_\delta\frac{1}{n}\sum_{i=1}^{n}\alpha_{t,i}\ell\big(y_i f_{\widetilde{\mathbf{W}}_{t-1}}(x_i)\big) = \rho_{\mathrm{clip}}G_\delta^2\mathcal{L}_{S,t}^{\alpha}(\widetilde{\mathbf{W}}_{t-1}) \leq 4\beta_{\mathrm{clip}}\mathcal{L}_{S,t}^{\alpha}(\widetilde{\mathbf{W}}_{t-1}).$$

Hence,

$$\eta^2\|\bar{g}_t + \widetilde{\Delta}_t\|_2^2 \leq 2\eta^2\|\bar{g}_t\|_2^2 + 2\eta^2\|\widetilde{\Delta}_t\|_2^2 \leq 8\beta_{\mathrm{clip}}\eta^2\mathcal{L}_{S,t}^{\alpha}(\widetilde{\mathbf{W}}_{t-1}) + 2\eta^2\|\widetilde{\Delta}_t\|_2^2.$$

Furthermore, from Young's inequality $2ab \leq \rho a^2 + \rho^{-1}b^2$ with $a = \sqrt{\eta}\,\|\bar{g}_t\|_2, b = \lambda\eta^{3/2}c_{\mathrm{priv}}\|Z_{t-1}\|_2$ and $\rho = \frac{1}{12\beta_{\mathrm{clip}}}$, we obtain

$$\begin{aligned}-2\lambda\eta^2 c_{\mathrm{priv}}\langle\bar{g}_t, Z_{t-1}\rangle &\leq 2\lambda\eta^2 c_{\mathrm{priv}}\big|\langle\bar{g}_t, Z_{t-1}\rangle\big| \leq \frac{\eta}{12\beta_{\mathrm{clip}}}\|\bar{g}_t\|_2^2 + 12\beta_{\mathrm{clip}}\lambda^2\eta^3 c_{\mathrm{priv}}^2\|Z_{t-1}\|_2^2 \\ &\leq \frac{\eta}{3}\mathcal{L}_{S,t}^{\alpha}(\widetilde{\mathbf{W}}_{t-1}) + 12\beta_{\mathrm{clip}}\lambda^2\eta^3 c_{\mathrm{priv}}^2\|Z_{t-1}\|_2^2.\end{aligned}$$

Combining the above estimates with $-\frac{4\eta}{3} + 8\beta_{\text{clip}}\eta^2 + \frac{\eta}{3} = -\eta + 8\beta_{\text{clip}}\eta^2 \le -\frac{\eta}{3}$, where the last inequality follows from $\eta \le \frac{1}{12\beta_{\text{clip}}}$, gives

$$\begin{aligned}\|\mathbf{U}_t - \mathbf{W}^*\|_2^2 \le \|\mathbf{U}_{t-1} - \mathbf{W}^*\|_2^2 - \frac{\eta}{3}\mathcal{L}_{S,t}^{\alpha}(\widetilde{\mathbf{W}}_{t-1}) + \frac{8\eta}{3}\mathcal{L}_S(\mathbf{W}^*) - 2\eta\langle\widetilde{\Delta}_t, \mathbf{U}_{t-1} - \mathbf{W}^*\rangle \\ - 2(1-\lambda)\eta c_{\text{priv}}\langle Z_{t-1}, \mathbf{U}_{t-1} - \mathbf{W}^*\rangle + 2(1-\lambda)\eta^2 c_{\text{priv}}\langle\widetilde{\Delta}_t, Z_{t-1}\rangle \\ + 2\eta^2\|\widetilde{\Delta}_t\|_2^2 + \Big((1-\lambda)^2\eta^2 c_{\text{priv}}^2 + 12\beta_{\text{clip}}\lambda^2\eta^3 c_{\text{priv}}^2\Big)\|Z_{t-1}\|_2^2.\end{aligned}$$

Summing from $t = 1$ to τ and using $\mathbf{1}_{A_{\bar{R}}^U(t-1)} = 1$ for every $1 \le t \le \tau$, we obtain

$$\begin{aligned}\|\mathbf{U}_\tau - \mathbf{W}^*\|_2^2 + \frac{\eta}{3}\sum_{t=1}^{\tau}\mathcal{L}_{S,t}^{\alpha}(\widetilde{\mathbf{W}}_{t-1}) \le \|\mathbf{W}_0 - \mathbf{W}^*\|_2^2 + \frac{8}{3}\eta T\mathcal{L}_S(\mathbf{W}^*) + M_\tau + 2\eta^2\sum_{t=1}^{\tau}\|\widetilde{\Delta}_t\|_2^2 \\ + \Big((1-\lambda)^2\eta^2 c_{\text{priv}}^2 + 12\beta_{\text{clip}}\lambda^2\eta^3 c_{\text{priv}}^2\Big)\sum_{t=1}^{\tau}\|Z_{t-1}\|_2^2. \quad (11)\end{aligned}$$

On $G_{\text{pot}}(\bar{R}, M)$, $G_{\Delta^2}(\bar{R}, V_\Delta)$, and $G_Z(z, V_Z)$, since the stopping indicators equal one for all $t \le \tau$, it holds that

$$M_\tau \le M \quad \text{and} \quad \sum_{t=1}^{\tau}\|\widetilde{\Delta}_t\|_2^2 \le \sum_{t=1}^{T}\|\widetilde{\Delta}_t\|_2^2\mathbf{1}_{A_{\bar{R}}^U(t-1)} \le V_\Delta \quad \text{and} \quad \sum_{t=1}^{\tau}\|Z_{t-1}\|_2^2 \le V_Z.$$

Hence,

$$\begin{aligned}&\|\mathbf{U}_\tau - \mathbf{W}^*\|_2^2 + \frac{\eta}{3}\sum_{t=1}^{\tau}\mathcal{L}_{S,t}^{\alpha}(\widetilde{\mathbf{W}}_{t-1}) \\ &\le \|\mathbf{W}_0 - \mathbf{W}^*\|_2^2 + \frac{8}{3}\eta T\mathcal{L}_S(\mathbf{W}^*) + M + 2\eta^2 V_\Delta + \Big((1-\lambda)^2\eta^2 c_{\text{priv}}^2 + 12\beta_{\text{clip}}\lambda^2\eta^3 c_{\text{priv}}^2\Big)V_Z \le \bar{R}^2.\end{aligned}$$

Therefore, $\|\mathbf{U}_\tau - \mathbf{W}^*\|_2^2 \le \bar{R}^2$, which contradicts the definition of τ. Hence, $A_{\bar{R}}^U$ holds.

Under $A_{\bar{R}}^U$, using equation 11 with $\tau = T$ and dropping the nonnegative term $\|\mathbf{U}_T - \mathbf{W}^*\|_2^2$, we obtain

$$\begin{aligned}\frac{\eta}{3}\sum_{t=1}^{T}\mathcal{L}_{S,t}^{\alpha}(\widetilde{\mathbf{W}}_{t-1}) \le \|\mathbf{W}_0 - \mathbf{W}^*\|_2^2 + \frac{8}{3}\eta T\mathcal{L}_S(\mathbf{W}^*) + M + 2\eta^2 V_\Delta \\ + \Big((1-\lambda)^2\eta^2 c_{\text{priv}}^2 + 12\beta_{\text{clip}}\lambda^2\eta^3 c_{\text{priv}}^2\Big)V_Z.\end{aligned}$$

Dividing both sides by $\eta T/3$ yields

$$\frac{1}{T}\sum_{t=1}^{T}\mathcal{L}_{S,t}^{\alpha}(\widetilde{\mathbf{W}}_{t-1}) \le 8\mathcal{L}_S(\mathbf{W}^*) + \frac{3(\|\mathbf{W}_0 - \mathbf{W}^*\|_2^2 + M)}{\eta T} + \frac{6\eta V_\Delta}{T} + 3\Big((1-\lambda)^2\eta + 12\beta_{\text{clip}}\lambda^2\eta^2\Big)\frac{c_{\text{priv}}^2 V_Z}{T}.$$

It remains to prove the empirical loss bound for the projected iterates. On $A_{\bar{R}}^U \cap G_Z(z, V_Z)$, for every $t = 0, 1, \ldots, T$,

$$\begin{aligned}\|\widetilde{\mathbf{W}}_t - \mathbf{W}_0\|_2 &= \|\mathbf{U}_t - \eta c_{\text{priv}} Z_t - \mathbf{W}_0\|_2 \le \|\mathbf{U}_t - \mathbf{W}^*\|_2 + \|\mathbf{W}^* - \mathbf{W}_0\|_2 + \eta c_{\text{priv}}\|Z_t\|_2 \\ &\le \bar{R} + \|\mathbf{W}^* - \mathbf{W}_0\|_2 + \eta c_{\text{priv}} z.\end{aligned}$$

Thus, if $R_* \ge \bar{R} + \|\mathbf{W}^* - \mathbf{W}_0\|_2 + \eta c_{\text{priv}} z$, then $\widetilde{\mathbf{W}}_t \in \mathcal{K} = \mathcal{B}(\mathbf{W}_0, R_*)$ for all $t = 0, 1, \ldots, T$.

By induction, the projected and auxiliary unprojected iterates coincide. The claim is trivial at $t = 0$. Suppose $\mathbf{W}_{t-1} = \widetilde{\mathbf{W}}_{t-1}$. Then the clipping coefficients, and hence the clipped mini-batch gradients, in the projected and auxiliary updates coincide. Therefore, the two updates have the same pre-projection point, namely $\widetilde{\mathbf{W}}_t$. Since $\widetilde{\mathbf{W}}_t \in \mathcal{K}$, projection leaves it unchanged:

$$\mathbf{W}_t = \Pi_{\mathcal{K}}(\widetilde{\mathbf{W}}_t) = \widetilde{\mathbf{W}}_t \quad \text{and further} \quad \mathbf{W}_t = \widetilde{\mathbf{W}}_t, \qquad \forall\, t = 0, 1, \ldots, T.$$

Since $\alpha_{t,i} \geq \rho_{\text{clip}}$ on $\mathcal{E}_\delta$ for every $i \in [n]$, we have

$$\mathcal{L}^{\alpha}_{S,t}(\widetilde{\mathbf{W}}_{t-1}) \geq \rho_{\text{clip}}\, \mathcal{L}_S(\widetilde{\mathbf{W}}_{t-1}) = \rho_{\text{clip}}\, \mathcal{L}_S(\mathbf{W}_{t-1}).$$

Therefore, the preceding bound on the weighted empirical loss yields

$$\frac{1}{T}\sum_{t=1}^{T} \mathcal{L}_S(\mathbf{W}_{t-1}) \leq \frac{1}{\rho_{\text{clip}}}\Big[8\mathcal{L}_S(\mathbf{W}^*) + \frac{3(\|\mathbf{W}_0 - \mathbf{W}^*\|_2^2 + M)}{\eta T} + \frac{6\eta V_\Delta}{T} + 3\Big((1-\lambda)^2\eta + 12\beta_{\text{clip}}\lambda^2\eta^2\Big)c_{\text{priv}}^2\frac{V_Z}{T}\Big].$$

This completes the proof. □

STEP 4: HIGH-PROBABILITY CONTROL AND A STOPPED OPTIMIZATION BOUND.

Now, we show that the good event $\mathcal{E}_{\text{good}}$ holds with high probability. We need the following Laurent–Massart chi-square tail bound Laurent & Massart (2000).

Lemma 15. *Let $G \sim \mathcal{N}(0, \mathbf{I}_d)$. Then $\|G\|_2^2 \sim \chi^2_d$, and for every $\delta \in (0,1)$,*

$$\mathbb{P}\left(\|G\|_2^2 \leq d + 2\sqrt{d\log(1/\delta)} + 2\log(1/\delta)\right) \geq 1 - \delta.$$

Lemma 16 (High-probability control of the Gaussian noises)**.** *Suppose $Z_1, \ldots, Z_T$ are i.i.d. standard Gaussian vectors in $\mathbb{R}^{mdp}$, and let $Z_0 = \mathbf{0}$. For any $\delta_Z \in (0,1)$, define $z_{\delta_Z} = \sqrt{mdp} + \sqrt{2\log\left(\frac{2T}{\delta_Z}\right)}$ and $V_{Z,\delta_Z} = (T-1)mdp + 2\sqrt{(T-1)mdp\log\left(\frac{2}{\delta_Z}\right)} + 2\log\left(\frac{2}{\delta_Z}\right)$. Then,*

$$\mathbb{P}\left(G_Z(z_{\delta_Z}, V_{Z,\delta_Z})\right) \geq 1 - \delta_Z.$$

Proof. We first control the maximum Gaussian norm. For a standard Gaussian vector $Z_t \sim \mathcal{N}(0, \mathbf{I}_{mdp})$, the standard Gaussian norm concentration inequality gives, for each $t = 1, \ldots, T$,

$$\mathbb{P}\Big(\|Z_t\|_2 > \sqrt{mdp} + \sqrt{2\log\big(\frac{2T}{\delta_Z}\big)}\Big) \leq \frac{\delta_Z}{2T}.$$

Since $Z_0 = \mathbf{0}$, taking a union bound over $t = 1, \ldots, T$ yields

$$\mathbb{P}\big(\max_{0\leq t\leq T}\|Z_t\|_2 \leq z_{\delta_Z}\big) \geq 1 - \frac{\delta_Z}{2}.$$

Next, we control the quadratic sum. Note $Z_0 = \mathbf{0}$, it holds

$$\sum_{t=1}^{T}\|Z_{t-1}\|_2^2 = \sum_{s=0}^{T-1}\|Z_s\|_2^2 = \sum_{s=1}^{T-1}\|Z_s\|_2^2.$$

The random variable on the right-hand side follows a chi-square distribution with $K = (T-1)mdp$ degrees of freedom. By the Laurent–Massart chi-square tail bound (see Lemma 15), it holds

$$\mathbb{P}\Big(\sum_{t=1}^{T}\|Z_{t-1}\|_2^2 \leq (T-1)mdp + 2\sqrt{(T-1)mdp\log\big(\frac{2}{\delta_Z}\big)} + 2\log\big(\frac{2}{\delta_Z}\big)\Big) \geq 1 - \frac{\delta_Z}{2}.$$

Hence,

$$\mathbb{P}\Big(\sum_{t=1}^{T}\|Z_{t-1}\|_2^2 \leq V_{Z,\delta_Z}\Big) \geq 1 - \frac{\delta_Z}{2}.$$

Combining the two estimates by a union bound gives

$$\mathbb{P}\left(G_Z(z_{\delta_Z}, V_{Z,\delta_Z})\right) \geq 1 - \delta_Z.$$

This completes the proof. □

Lemma 17 (High-probability control of the stopped mini-batch quadratic variation)**.** *Suppose Assumption 1 holds and the event $\mathcal{E}_\delta$ occurs. Fix $\bar{R} > 0$ and $\delta_\Delta \in (0,1)$. Define $V_{\Delta,\delta_\Delta} = \frac{2TG_{\text{clip}}^2}{B} + 8G_{\text{clip}}^2 \log\left(\frac{1}{\delta_\Delta}\right)$. Then, conditioned on the dataset and the initialization satisfying $\mathcal{E}_\delta$, it holds that*

$$\mathbb{P}_{\mathcal{A}}\left(G_{\Delta^2}(\bar{R}, V_{\Delta,\delta_\Delta})\right) \geq 1 - \delta_\Delta.$$

Proof. Condition on the dataset S and on the initialization satisfying $\mathcal{E}_\delta$. Let $\mathscr{F}_t = \sigma\big(\mathbf{W}_0, \mathbf{c}, \mathcal{B}_1, \ldots, \mathcal{B}_t, Z_1, \ldots, Z_t\big)$ be the natural filtration generated by the algorithmic randomness up to time t. For each $t \in [T]$, define $X_t = \|\widetilde{\Delta}_t\|_2^2 \mathbf{1}_{A_{\bar{R}}^U(t-1)}$.

We first bound the conditional expectation of X_t. Conditioned on $\mathscr{F}_{t-1}$, the iterate $\widetilde{\mathbf{W}}_{t-1}$ is fixed. Hence the clipped gradients $\{g_{t,i}^{\text{aux}}\}_{i=1}^n$ are deterministic. Moreover, on $\mathcal{E}_\delta$,

$$\|g_{t,i}^{\text{aux}}\|_2 = \left\|\alpha_{t,i} \nabla \ell\big(y_i f_{\widetilde{\mathbf{W}}_{t-1}}(\mathbf{x}_i)\big)\right\|_2 \leq G_{\text{clip}}, \qquad \forall i \in [n].$$

Therefore, by Lemma 4, it holds that $\mathbb{E}\big[\|\widetilde{\Delta}_t\|_2^2 \,\big|\, \mathscr{F}_{t-1}\big] \leq \frac{G_{\text{clip}}^2}{B}$. Thus,

$$\mathbb{E}[X_t \mid \mathscr{F}_{t-1}] = \mathbf{1}_{A_{\bar{R}}^U(t-1)} \mathbb{E}\big[\|\widetilde{\Delta}_t\|_2^2 \,\big|\, \mathscr{F}_{t-1}\big] \leq \frac{G_{\text{clip}}^2}{B}.$$

Summing over t, we get

$$\sum_{t=1}^T \mathbb{E}[X_t \mid \mathscr{F}_{t-1}] \leq \frac{TG_{\text{clip}}^2}{B} =: V.$$

Next, we bound X_t almost surely. From the definition, we know $\left\|\frac{1}{B}\sum_{i\in\mathcal{B}_t} g_{t,i}^{\text{aux}}\right\|_2 \leq G_{\text{clip}}$ and $\|\bar{g}_t\|_2 = \left\|\frac{1}{n}\sum_{i=1}^n g_{t,i}^{\text{aux}}\right\|_2 \leq G_{\text{clip}}$, then

$$\|\widetilde{\Delta}_t\|_2 \leq 2G_{\text{clip}}, \qquad X_t \leq 4G_{\text{clip}}^2 =: L.$$

According to Lemma 5 with $V = \frac{TG_{\text{clip}}^2}{B}$ and $L = 4G_{\text{clip}}^2$, with probability at least $1 - \delta_\Delta$,

$$\sum_{t=1}^T \|\widetilde{\Delta}_t\|_2^2 \mathbf{1}_{A_{\bar{R}}^U(t-1)} \leq \frac{2TG_{\text{clip}}^2}{B} + 8G_{\text{clip}}^2 \log\left(\frac{1}{\delta_\Delta}\right) = V_{\Delta,\delta_\Delta}.$$

This completes the proof. □

We next estimate the three terms in the stopped potential fluctuation M_k separately.

Lemma 18. *Suppose Assumption 1 holds and the event $\mathcal{E}_\delta$ occurs. Fix $\bar{R} > 0$ and $\delta_{\text{pot},1} \in (0,1)$. Define*

$$M_k^{(1)} = -2\eta \sum_{t=1}^k \mathbf{1}_{A_{\bar{R}}^U(t-1)} \langle \widetilde{\Delta}_t, \mathbf{U}_{t-1} - \mathbf{W}^* \rangle, \qquad k \in [T].$$

Then, conditioned on the dataset and the initialization satisfying $\mathcal{E}_\delta$, it holds that

$$\mathbb{P}_{\mathcal{A}}\Big(\max_{1\leq k\leq T} |M_k^{(1)}| \leq 4\sqrt{2}\,\eta G_{\text{clip}} \bar{R} \sqrt{\frac{T\log(2/\delta_{\text{pot},1})}{B}}\Big) \geq 1 - \delta_{\text{pot},1}.$$

Proof. Condition on the dataset S and on the initialization satisfying $\mathcal{E}_\delta$. For each $t \in [T]$, define $X_t = -2\eta \mathbf{1}_{A_{\bar{R}}^U(t-1)} \langle \widetilde{\Delta}_t, \mathbf{U}_{t-1} - \mathbf{W}^* \rangle$. Then $M_k^{(1)} = \sum_{t=1}^k X_t$. Since $\mathbb{E}[\widetilde{\Delta}_t \mid \mathscr{F}_{t-1}] = 0$, we have $\mathbb{E}[X_t \mid \mathscr{F}_{t-1}] = 0$.

We now show that X_t is conditionally sub-Gaussian. Conditioned on $\mathscr{F}_{t-1}$, define $h_t = \mathbf{1}_{A_{\bar{R}}^U(t-1)}(\mathbf{U}_{t-1} - \mathbf{W}^*)$. Then $\|h_t\|_2 \leq \bar{R}$. Also define the deterministic scalars $a_i = \langle g_{t,i}^{\text{aux}}, h_t \rangle$ for $i \in [n]$. On the event $\mathcal{E}_\delta$, by the definition of $g_{t,i}^{\text{aux}}$,

$$|a_i| \leq \|g_{t,i}^{\text{aux}}\|_2 \|h_t\|_2 \leq G_{\text{clip}} \bar{R}, \qquad \forall i \in [n].$$

Let $\bar{a} = \frac{1}{n}\sum_{i=1}^{n} a_i$. Conditioned on $\mathscr{F}_{t-1}$, the values $a_1, \dots, a_n$ are fixed, and $\langle \widetilde{\Delta}_t, h_t \rangle = \frac{1}{B}\sum_{i\in\mathcal{B}_t} a_i - \bar{a}$. By Lemma 6, conditionally on $\mathscr{F}_{t-1}$, with $a_i = \langle g_{t,i}^{\mathrm{aux}}, h_t \rangle$ and $|a_i| \le G_{\mathrm{clip}}\bar{R}$, we have, for every $\theta \in \mathbb{R}$,

$$\mathbb{E}\Big[\exp\Big(\theta\langle\widetilde{\Delta}_t, h_t\rangle\Big)\Big|\mathscr{F}_{t-1}\Big] \le \exp\Big(\frac{\theta^2 G_{\mathrm{clip}}^2\bar{R}^2}{2B}\Big).$$

Since $X_t = -2\eta\langle\widetilde{\Delta}_t, h_t\rangle$, it follows that

$$\mathbb{E}\left[e^{\theta X_t}\middle|\mathscr{F}_{t-1}\right] \le \exp\Big(\frac{2\theta^2\eta^2 G_{\mathrm{clip}}^2\bar{R}^2}{B}\Big).$$

Therefore, $\{X_t\}_{t=1}^T$ is a conditionally sub-Gaussian martingale difference sequence with variance proxy $\sigma_1^2 = \frac{4\eta^2 G_{\mathrm{clip}}^2\bar{R}^2}{B}$.

By Theorem 1(a) of Howard et al. (2020), applied to the sub-Gaussian process with $S_k = \sum_{t=1}^k X_t$ and $V_k = k\sigma_1^2$, and then to $-S_k$, we obtain, for every $u > 0$,

$$\mathbb{P}_{\mathcal{A}}\Big(\max_{1\le k\le T}\Big|\sum_{t=1}^{k} X_t\Big| \ge \sqrt{2T\sigma_1^2 u}\Big) \le 2e^{-u}.$$

Taking $u = \log\big(\frac{2}{\delta_{\mathrm{pot},1}}\big)$, we obtain

$$\mathbb{P}_{\mathcal{A}}\Big(\max_{1\le k\le T}|M_k^{(1)}| \ge \eta 2G_{\mathrm{clip}}\bar{R}\sqrt{\frac{2T\log(2/\delta_{\mathrm{pot},1})}{B}}\Big) \le \delta_{\mathrm{pot},1}.$$

Enlarging the numerical constant to $4\sqrt{2}$ proves the stated bound. □

Lemma 19. *Suppose $0 < \lambda < 1$. Fix $\bar{R} > 0$ and $\delta_{\mathrm{pot},2} \in (0,1)$. Define*

$$M_k^{(2)} = -2(1-\lambda)\eta c_{\mathrm{priv}}\sum_{t=1}^{k}\mathbf{1}_{A_{\bar{R}}^U(t-1)}\big\langle Z_{t-1}, \mathbf{U}_{t-1} - \mathbf{W}^*\big\rangle, \qquad k \in [T].$$

Then, conditioned on the dataset and the initialization satisfying the event $\mathcal{E}_\delta$, it holds that

$$\mathbb{P}_{\mathcal{A}}\Big(\max_{1\le k\le T}|M_k^{(2)}| \le 4(1-\lambda)\eta c_{\mathrm{priv}}\,\bar{R}\sqrt{T\log(2/\delta_{\mathrm{pot},2})}\Big) \ge 1 - \delta_{\mathrm{pot},2}.$$

Proof. Condition on the dataset S, the comparator $\mathbf{W}^*$, and the initialization $(\mathbf{W}_0, \mathbf{c})$ satisfying the event $\mathcal{E}_\delta$. For $t \ge 2$, define the lagged filtration $\mathscr{H}_{t-1} = \sigma\big(S, \mathbf{W}_0, \mathbf{c}, \mathbf{W}^*, \mathcal{B}_1, \dots, \mathcal{B}_{t-1}, Z_1, \dots, Z_{t-2}\big)$. For $t = 1$, the summand is zero since $Z_0 = \mathbf{0}$. For $t \ge 2$, the shifted iterate $\mathbf{U}_{t-1}$ is $\mathscr{H}_{t-1}$-measurable, and so is $\mathbf{1}_{A_{\bar{R}}^U(t-1)}$. Moreover, $Z_{t-1} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{mdp})$ is independent of $\mathscr{H}_{t-1}$.

Define

$$Y_t = -2(1-\lambda)\eta c_{\mathrm{priv}}\,\mathbf{1}_{A_{\bar{R}}^U(t-1)}\big\langle Z_{t-1}, \mathbf{U}_{t-1} - \mathbf{W}^*\big\rangle, \qquad t = 1, \dots, T.$$

Then, $M_k^{(2)} = \sum_{t=1}^k Y_t$. For $t = 1$, $Y_1 = 0$. For $t \ge 2$, $\mathbb{E}[Y_t \mid \mathscr{H}_{t-1}] = 0$.

Furthermore, conditioned on $\mathscr{H}_{t-1}$, Y_t is Gaussian with variance

$$4(1-\lambda)^2\eta^2 c_{\mathrm{priv}}^2\mathbf{1}_{A_{\bar{R}}^U(t-1)}\|\mathbf{U}_{t-1} - \mathbf{W}^*\|_2^2 \le 4(1-\lambda)^2\eta^2 c_{\mathrm{priv}}^2\bar{R}^2.$$

Therefore, for every $\theta \in \mathbb{R}$,

$$\mathbb{E}\left[\exp(\theta Y_t)\,|\,\mathscr{H}_{t-1}\right] \le \exp\left(2\theta^2(1-\lambda)^2\eta^2 c_{\mathrm{priv}}^2\bar{R}^2\right).$$

That is, $\{Y_t\}_{t=1}^T$ is a conditionally sub-Gaussian martingale difference sequence with variance proxy $\sigma_2^2 = 4(1-\lambda)^2\eta^2 c_{\mathrm{priv}}^2\bar{R}^2$. Then, by Theorem 1(a) of Howard et al. (2020), applied to the conditionally sub-Gaussian martingale difference sequence $\{Y_t\}_{t=1}^T$ and to its negative,

$$\mathbb{P}_{\mathcal{A}}\Big(\max_{1\le k\le T}\Big|\sum_{t=1}^{k} Y_t\Big| \ge \sqrt{2T\sigma_2^2 u}\Big) \le 2e^{-u}.$$

Taking $u = \log(\frac{2}{\delta_{\mathrm{pot},2}})$, we get with probability at least $1 - \delta_{\mathrm{pot},2}$,

$$\max_{1\le k\le T} |M_k^{(2)}| \le 2\sqrt{2}(1-\lambda)\eta c_{\mathrm{priv}}\,\bar{R}\sqrt{T\log(2/\delta_{\mathrm{pot},2})}.$$

Enlarging the constant to 4 proves the claim. □

Lemma 20. *Suppose Assumption 1 holds and the event $\mathcal{E}_\delta$ occurs. Fix $\bar{R} > 0$, $V_Z > 0$, and $\delta_{\mathrm{pot},3} \in (0,1)$. Define*

$$M_k^{(3)} = 2(1-\lambda)\eta^2 c_{\mathrm{priv}} \sum_{t=1}^{k} \mathbf{1}_{A_{\bar{R}}^U(t-1)} \big\langle \widetilde{\Delta}_t, Z_{t-1} \big\rangle, \qquad k \in [T].$$

Then, conditioned on the dataset and the initialization satisfying $\mathcal{E}_\delta$, it holds that

$$\mathbb{P}_{\mathcal{A}}\Big(G_Z(z,V_Z) \cap \Big\{ \max_{1\le k\le T} |M_k^{(3)}| > 4(1-\lambda)\eta^2 c_{\mathrm{priv}} G_{\mathrm{clip}} \sqrt{\frac{V_Z \log(2/\delta_{\mathrm{pot},3})}{B}} \Big\} \Big) \le \delta_{\mathrm{pot},3}.$$

Proof. Condition on the dataset S, the comparator $\mathbf{W}^*$, and the initialization $(\mathbf{W}_0, \mathbf{c})$ satisfying $\mathcal{E}_\delta$. Conditioned on $\mathscr{F}_{t-1}$, the vectors Z_{t-1}, $\widetilde{\mathbf{W}}_{t-1}$, and $\{g_{t,i}^{\mathrm{aux}}\}_{i=1}^n$ are fixed, while the randomness in $\widetilde{\Delta}_t$ comes only from the fresh mini-batch $\mathcal{B}_t$.

For each $t \in [T]$, define $Y_t = 2(1-\lambda)\eta^2 c_{\mathrm{priv}} \mathbf{1}_{A_{\bar{R}}^U(t-1)} \big\langle \widetilde{\Delta}_t, Z_{t-1} \big\rangle$. Then $M_k^{(3)} = \sum_{t=1}^k Y_t$. Since $\mathbb{E}[\widetilde{\Delta}_t \mid \mathscr{F}_{t-1}] = 0$, we have $\mathbb{E}[Y_t \mid \mathscr{F}_{t-1}] = 0$.

Conditioned on $\mathscr{F}_{t-1}$, set $h_t = \mathbf{1}_{A_{\bar{R}}^U(t-1)} Z_{t-1}$. Since $\|g_{t,i}^{\mathrm{aux}}\|_2 \le G_{\mathrm{clip}}$ for $i \in [n]$, Lemma 6 gives, for every $\theta \in \mathbb{R}$,

$$\mathbb{E}\Big[\exp\big(\theta \langle \widetilde{\Delta}_t, h_t \rangle\big) \big| \mathscr{F}_{t-1} \Big] \le \exp\Big(\frac{\theta^2 G_{\mathrm{clip}}^2 \mathbf{1}_{A_{\bar{R}}^U(t-1)} \|Z_{t-1}\|_2^2}{2B} \Big).$$

Since $Y_t = 2(1-\lambda)\eta^2 c_{\mathrm{priv}} \langle \widetilde{\Delta}_t, h_t \rangle$, we obtain

$$\mathbb{E}\left[\exp(\theta Y_t) | \mathscr{F}_{t-1}\right] \le \exp\Big(\frac{\theta^2 \sigma_t^2}{2}\Big),$$

where $\sigma_t^2 = 4(1-\lambda)^2\eta^4 c_{\mathrm{priv}}^2 G_{\mathrm{clip}}^2 \mathbf{1}_{A_{\bar{R}}^U(t-1)} \|Z_{t-1}\|_2^2 / B$. Thus, $\{Y_t\}_{t=1}^T$ is a conditionally sub-Gaussian martingale difference sequence with predictable variance process $V_k^{(3)} = \sum_{t=1}^k \sigma_t^2$. On $G_Z(z, V_Z)$, we have

$$V_T^{(3)} \le 4(1-\lambda)^2\eta^4 c_{\mathrm{priv}}^2 G_{\mathrm{clip}}^2 \frac{V_Z}{B} =: v.$$

By Theorem 1(a) of Howard et al. (2020), applied to $\sum_{t=1}^k Y_t$ and its negative, for every $u > 0$,

$$\mathbb{P}_{\mathcal{A}}\Big(\max_{1\le k\le T} \Big| \sum_{t=1}^k Y_t \Big| \ge \sqrt{2vu}, \quad V_T^{(3)} \le v \Big) \le 2e^{-u}.$$

Taking $u = \log(\frac{2}{\delta_{\mathrm{pot},3}})$ and using the fact that $G_Z(z, V_Z)$ implies $V_T^{(3)} \le v$, we obtain

$$\mathbb{P}_{\mathcal{A}}\Big(G_Z(z,V_Z) \cap \Big\{ \max_{1\le k\le T} |M_k^{(3)}| \ge 2\sqrt{2}(1-\lambda)\eta^2 c_{\mathrm{priv}} G_{\mathrm{clip}} \sqrt{\frac{V_Z \log(2/\delta_{\mathrm{pot},3})}{B}} \Big\} \Big) \le \delta_{\mathrm{pot},3}.$$

Enlarging the numerical constant from $2\sqrt{2}$ to 4 proves the stated bound. □

We now combine the above three fluctuation bounds to control the stopped potential-fluctuation event G_{pot}.

Lemma 21 (High-probability control of the stopped potential fluctuation)**.** *Suppose Assumption 1 holds and the event $\mathcal{E}_\delta$ occurs. Fix $\bar{R} > 0$, $z > 0$, $V_Z > 0$, and $\delta_{\mathrm{pot}} \in (0,1)$. Define*

$$M_{\mathrm{pot}}(\bar{R}, z, V_Z; \delta_{\mathrm{pot}}) = 4\sqrt{2}\,\eta G_{\mathrm{clip}} \bar{R}\sqrt{\frac{T\log(6/\delta_{\mathrm{pot}})}{B}} + 4(1-\lambda)\eta c_{\mathrm{priv}}\,\bar{R}\sqrt{T\log\big(6/\delta_{\mathrm{pot}}\big)} + 4(1-\lambda)\eta^2 c_{\mathrm{priv}} G_{\mathrm{clip}}\sqrt{\frac{V_Z\log(6/\delta_{\mathrm{pot}})}{B}}. \tag{12}$$

Then, conditioned on the dataset and the initialization satisfying $\mathcal{E}_\delta$, it holds that

$$\mathbb{P}_{\mathcal{A}}\left(G_Z(z,V_Z) \cap G_{\mathrm{pot}}\big(\bar{R}, M_{\mathrm{pot}}(\bar{R}, z, V_Z; \delta_{\mathrm{pot}})\big)\right) \geq \mathbb{P}_{\mathcal{A}}\big(G_Z(z,V_Z)\big) - \delta_{\mathrm{pot}}.$$

Proof. Note that $M_k = M_k^{(1)} + M_k^{(2)} + M_k^{(3)}$ for all $k \in [T]$. Define the following three events

$$E_1 = \Big\{ \max_{1\leq k\leq T} |M_k^{(1)}| \leq 4\sqrt{2}\,\eta G_{\mathrm{clip}}\bar{R}\sqrt{\frac{T\log(6/\delta_{\mathrm{pot}})}{B}}\Big\},$$

$$E_2 = \Big\{ \max_{1\leq k\leq T} |M_k^{(2)}| \leq 4(1-\lambda)\eta c_{\mathrm{priv}}\,\bar{R}\sqrt{T\log\big(\frac{6}{\delta_{\mathrm{pot}}}\big)}\Big\},$$

$$E_3 = \Big\{ \max_{1\leq k\leq T} |M_k^{(3)}| \leq 4(1-\lambda)\eta^2 c_{\mathrm{priv}} G_{\mathrm{clip}}\sqrt{\frac{V_Z\log(6/\delta_{\mathrm{pot}})}{B}}\Big\}.$$

By Lemmas 18, 19 and 20 with $\delta_{\mathrm{pot},1} = \delta_{\mathrm{pot},2} = \delta_{\mathrm{pot},3} = \frac{\delta_{\mathrm{pot}}}{3}$, we have

$$\mathbb{P}_{\mathcal{A}}(E_1) \geq 1 - \frac{\delta_{\mathrm{pot}}}{3}, \qquad \mathbb{P}_{\mathcal{A}}(E_2) \geq 1 - \frac{\delta_{\mathrm{pot}}}{3} \quad \text{and} \quad \mathbb{P}_{\mathcal{A}}\big(G_Z(z,V_Z)\cap E_3^c\big) \leq \frac{\delta_{\mathrm{pot}}}{3}.$$

On $E_1 \cap E_2 \cap E_3$, for every $k \in [T]$, it holds that $|M_k| \leq |M_k^{(1)}| + |M_k^{(2)}| + |M_k^{(3)}| \leq M_{\mathrm{pot}}(\bar{R}, z, V_Z; \delta_{\mathrm{pot}})$. Hence,

$$E_1 \cap E_2 \cap E_3 \subseteq G_{\mathrm{pot}}\big(\bar{R}, M_{\mathrm{pot}}(\bar{R}, z, V_Z; \delta_{\mathrm{pot}})\big).$$

Therefore,

$$\mathbb{P}_{\mathcal{A}}\left(G_Z(z,V_Z) \cap \big(G_{\mathrm{pot}}\big(\bar{R}, M_{\mathrm{pot}}(\bar{R}, z, V_Z; \delta_{\mathrm{pot}})\big)\big)^c\right) \leq \mathbb{P}_{\mathcal{A}}(E_1^c) + \mathbb{P}_{\mathcal{A}}(E_2^c) + \mathbb{P}_{\mathcal{A}}\big(G_Z(z,V_Z)\cap E_3^c\big) \leq \delta_{\mathrm{pot}}.$$

Rearranging proves the claim. □

Recall that

$$z_{\delta_Z} = \sqrt{mdp} + \sqrt{2\log(2T/\delta_Z)}, \qquad V_{\Delta,\delta_\Delta} = \frac{2TG_{\mathrm{clip}}^2}{B} + 8G_{\mathrm{clip}}^2\log(1/\delta_\Delta),$$

$$V_{Z,\delta_Z} = (T-1)mdp + 2\sqrt{(T-1)mdp\log(2/\delta_Z)} + 2\log(2/\delta_Z),$$

and

$$M_{\delta_{\mathrm{pot}}} = 4\sqrt{2}\,\eta G_{\mathrm{clip}}\bar{R}\sqrt{\frac{T\log(6/\delta_{\mathrm{pot}})}{B}} + 4(1-\lambda)\eta c_{\mathrm{priv}}\bar{R}\sqrt{T\log(6/\delta_{\mathrm{pot}})} + 4(1-\lambda)\eta^2 c_{\mathrm{priv}} G_{\mathrm{clip}}\sqrt{\frac{V_{Z,\delta_Z}\log(6/\delta_{\mathrm{pot}})}{B}}. \tag{13}$$

We now combine the high-probability bounds for G_Z, G_{Δ^2}, and G_{pot} with the shifted bootstrap lemma to obtain a conditional optimization bound for the projected iterates.

Lemma 22 (High-probability shifted bootstrap)**.** *Suppose $0 < \lambda < 1$, Assumption 1 holds and the event $\mathcal{E}_\delta$ occurs. Fix a comparator $\mathbf{W}^*$ and a shifted radius $\bar{R} > 0$. Let $\delta_Z, \delta_\Delta, \delta_{\mathrm{pot}} \in (0,1)$. Assume $\eta \leq \frac{1}{12\beta_{\mathrm{clip}}}$, equation 7 holds, and equation 10 holds with $M = M_{\delta_{\mathrm{pot}}}, V_\Delta = V_{\Delta,\delta_\Delta}$ and $V_Z = V_{Z,\delta_Z}$. If $R_* \geq \bar{R} + \|\mathbf{W}^* - \mathbf{W}_0\|_2 + \eta c_{\mathrm{priv}} z_{\delta_Z}$, then, conditional on the dataset and the*

initialization satisfying $\mathcal{E}_\delta$, with probability at least $1-(\delta_Z+\delta_\Delta+\delta_{\mathrm{pot}})$ over the randomness of the algorithm,

$$\frac{1}{T}\sum_{t=1}^{T}\mathcal{L}_S(\mathbf{W}_{t-1}) \le \frac{1}{\rho_{\mathrm{clip}}}\Big[8\mathcal{L}_S(\mathbf{W}^*) + \frac{3\big(\|\mathbf{W}_0-\mathbf{W}^*\|_2^2 + M_{\delta_{\mathrm{pot}}}\big)}{\eta T} + \frac{6\eta V_{\Delta,\delta_\Delta}}{T} + 3\Big((1-\lambda)^2\eta + 12\beta_{\mathrm{clip}}\lambda^2\eta^2\Big)\frac{c_{\mathrm{priv}}^2 V_{Z,\delta_Z}}{T}\Big].$$

Proof. By Lemma 16, Lemma 17 and Lemma 21, we have

$$\mathbb{P}_{\mathcal{A}}\Big(G_Z(z_{\delta_Z},V_{Z,\delta_Z})\cap G_{\Delta^2}\big(\bar{R},V_{\Delta,\delta_\Delta}\big)\cap G_{\mathrm{pot}}\big(\bar{R},M_{\delta_{\mathrm{pot}}}\big)\Big) \ge 1-\big(\delta_Z+\delta_\Delta+\delta_{\mathrm{pot}}\big).$$

Combining this with Lemma 14 completes the proof. □

STEP 5 AND 6: CLOSE THE BOOTSTRAP, SHOW PROJECTION INACTIVITY AND CHOOSE THE COMPARATOR USING INITIALIZATION SEPARABILITY.

Compared with the proof sketch in Section 4.3, the last two steps are treated together in the detailed proof. In the sketch, Step 5 closes the localization bootstrap for a fixed comparator, while Step 6 chooses the NTK-separability comparator. In the rigorous argument, these two steps are coupled, since the shifted localization event, the bootstrap condition, and the projection-inactivity argument all depend on the chosen comparator $\mathbf{W}^*$. We therefore first establish the existence of a suitable comparator under the initialization separability assumption, and then use this comparator to close the bootstrap and transfer the bound to the projected iterates.

Lemma 23 (Comparator under NTK separability)**.** *Suppose Assumption 2 holds. Assume*

$$m \gtrsim \log(m/\delta)\big(\log^2(T/\rho_{\mathrm{clip}})+\log(n/\delta)\big)/\gamma^4.$$

Let $\tau \asymp \frac{1}{\gamma}\big(\log(T/\rho_{\mathrm{clip}})+\sqrt{\log(n/\delta)}\big)$. Assume $R_* \ge \tau$. Then, with probability at least $1-\delta$ over the randomness of the initialization, there exists a comparator $\mathbf{W}^* = \mathbf{W}_0+\tau\mathbf{u}\in\mathcal{K}$ such that

$$L_S(\mathbf{W}^*) \le \frac{\rho_{\mathrm{clip}}}{T} \quad \textit{and} \quad \|\mathbf{W}^*-\mathbf{W}_0\|_2^2 = \tau^2 \asymp \frac{\log^2(T/\rho_{\mathrm{clip}})+\log(n/\delta)}{\gamma^2}.$$

Proof. Let $\mathbf{u}$ be the unit vector in Assumption 2, and define $\mathbf{W}^* = \mathbf{W}_0+\tau\mathbf{u}$. Since $\|\mathbf{u}\|_2=1$ and $\tau\le R_*$, we have

$$\mathbf{W}^*\in\mathcal{K}, \qquad \|\mathbf{W}^*-\mathbf{W}_0\|_2=\tau.$$

For any fixed $i\in[n]$, by Taylor's theorem, there exists $\alpha_i\in[0,1]$ with $\mathbf{W}_i' = \alpha_i\mathbf{W}_0+(1-\alpha_i)\mathbf{W}^*$ such that

$$f_{\mathbf{W}^*}(\mathbf{x}_i) = f_{\mathbf{W}_0}(\mathbf{x}_i)+\big\langle\nabla f_{\mathbf{W}_0}(\mathbf{x}_i),\mathbf{W}^*-\mathbf{W}_0\big\rangle+\frac{1}{2}(\mathbf{W}^*-\mathbf{W}_0)^\top\nabla^2 f_{\mathbf{W}_i'}(\mathbf{x}_i)(\mathbf{W}^*-\mathbf{W}_0).$$

Multiplying both sides by y_i and using $\mathbf{W}^*-\mathbf{W}_0=\tau\mathbf{u}$, we obtain

$$y_i f_{\mathbf{W}^*}(\mathbf{x}_i) = y_i f_{\mathbf{W}_0}(\mathbf{x}_i)+\tau\, y_i\big\langle\nabla f_{\mathbf{W}_0}(\mathbf{x}_i),\mathbf{u}\big\rangle+\frac{\tau^2}{2}y_i\,\mathbf{u}^\top\nabla^2 f_{\mathbf{W}_i'}(\mathbf{x}_i)\mathbf{u}.$$

Since $|y_i|\le 1$, it holds that

$$y_i f_{\mathbf{W}^*}(\mathbf{x}_i) \ge -|f_{\mathbf{W}_0}(\mathbf{x}_i)|+\tau\, y_i\big\langle\nabla f_{\mathbf{W}_0}(\mathbf{x}_i),\mathbf{u}\big\rangle-\frac{\tau^2}{2}\big\|\nabla^2 f_{\mathbf{W}_i'}(\mathbf{x}_i)\big\|_2.$$

From Assumption 2, we know that

$$y_i\big\langle\nabla f_{\mathbf{W}_0}(\mathbf{x}_i),\mathbf{u}\big\rangle\ge\gamma,$$

which implies

$$y_i f_{\mathbf{W}^*}(\mathbf{x}_i) \ge \tau\gamma-|f_{\mathbf{W}_0}(\mathbf{x}_i)|-\frac{\tau^2}{2}\big\|\nabla^2 f_{\mathbf{W}_i'}(\mathbf{x}_i)\big\|_2. \tag{14}$$

Now, we estimate the lower bound of the right hand side of equation 14. From Eq. (27) in Wang et al. (2026), with probability at least $1-\delta/2$ over the initialization,

$$|f_{\mathbf{W}_0}(\mathbf{x}_i)| \le B_b\sqrt{2p\log(2n/\delta)}, \qquad \forall i \in [n]. \tag{15}$$

Moreover, $\mathcal{E}_{\delta/2}$ occurs with probability at least $1-\delta/2$. On this event, Lemma 10 gives

$$\left\|\nabla^2 f_{\mathbf{W}_i'}(\mathbf{x}_i)\right\|_2 \le \frac{C_{\sigma,b}p^{3/2}\big(\sqrt{p}+\sqrt{\log(2m/\delta)}\big)}{\sqrt{m}}.$$

Plugging the above two bounds into equation 14, we get

$$\begin{aligned} y_i f_{\mathbf{W}^*}(\mathbf{x}_i) &\ge \tau\gamma - B_b\sqrt{2p\log(\frac{2n}{\delta})} - \frac{\tau^2 C_{\sigma,b}\,p^{\frac{3}{2}}\big(\sqrt{p}+\sqrt{\log(2m/\delta)}\big)}{2\sqrt{m}} \\ &\ge \frac{\tau\gamma}{2} - \frac{\tau^2 C_{\sigma,b}\,p^{\frac{3}{2}}\big(\sqrt{p}+\sqrt{\log(2m/\delta)}\big)}{2\sqrt{m}} \ge \frac{\tau\gamma}{4} \gtrsim \log(T/\rho_{\mathrm{clip}}), \end{aligned}$$

where the second inequality uses $\tau \asymp \frac{\log(T/\rho_{\mathrm{clip}})+\sqrt{\log(n/\delta)}}{\gamma}$, and the third inequality uses $m \gtrsim \frac{\log(m/\delta)(\log^2(T/\rho_{\mathrm{clip}})+\log(n/\delta))}{\gamma^4} \gtrsim \frac{\log(m/\delta)\tau^2}{\gamma^2}$.

It then follows that

$$\begin{aligned} \ell\big(y_i f_{\mathbf{W}^*}(\mathbf{x}_i)\big) &= \log\big(1+\exp\big(-y_i f_{\mathbf{W}^*}(\mathbf{x}_i)\big)\big) \le \exp\big(-y_i f_{\mathbf{W}^*}(\mathbf{x}_i)\big) \\ &\le \exp\big(-\log(T/\rho_{\mathrm{clip}})\big) = \frac{\rho_{\mathrm{clip}}}{T}. \end{aligned}$$

Averaging over $i \in [n]$ yields

$$\mathcal{L}_S(\mathbf{W}^*) \le \frac{\rho_{\mathrm{clip}}}{T},$$

which proves the first result of the lemma.

The second result follows directly from the definition of τ. This completes the proof. □

Recall that $c_{\mathrm{priv}} = \frac{C_{\mathrm{clip}}\kappa}{B}$, $\beta = C_{\sigma,b}p^3$, $G_{\mathrm{clip}} = \min\{C_{\mathrm{clip}}, G_\delta\}$ and $\rho_{\mathrm{clip}} = \min\big\{1, \frac{C_{\mathrm{clip}}}{G_\delta}\big\}$. Define $\eta_{\mathrm{eff}} := \eta\rho_{\mathrm{clip}}$ and $c_{\mathrm{eff}} := \frac{c_{\mathrm{priv}}}{\rho_{\mathrm{clip}}}$. Now, we give our main result.

Theorem 24 (Restatement of Theorem 1). *Let $\delta \in (0,1)$, $0<\lambda<1$, $\kappa>0$, and suppose Assumptions 1 and 2 hold. Set $R_* \asymp \frac{\log(T/\rho_{\mathrm{clip}})+\sqrt{\log(n/\delta)}}{\gamma}$. Let $\{\mathbf{W}_t\}_{t=0}^{T-1}$ be generated by Algorithm 1 with $\eta_{\mathrm{eff}} \le \frac{1}{12C_{\sigma,b}p^3}$ and $C_{\mathrm{clip}} > 0$. Assume*

$$\eta_{\mathrm{eff}}\gamma\sqrt{T}\Big(\frac{1}{\sqrt{B}} + (1-\lambda)c_{\mathrm{eff}}\Big) + \eta_{\mathrm{eff}}^2\gamma^2\Big(\frac{T}{B}+1\Big) \lesssim 1$$

and

$$\widetilde{\Omega}(\gamma^{-4}) \le m \le \widetilde{\mathcal{O}}\Big((\eta_{\mathrm{eff}}^2 c_{\mathrm{eff}}^2\gamma^2 d)^{-1}\min\big\{1, \big(T((1-\lambda)^2+\lambda^2\eta_{\mathrm{eff}})\big)^{-1}\big\}\Big),$$

where $\widetilde{\mathcal{O}}$ and $\widetilde{\Omega}$ suppress polylogarithmic factors in $m, n, T, \delta^{-1}, \rho_{\mathrm{clip}}^{-1}$. Then, with probability at least $1-\delta$ over the initialization and the algorithmic randomness,

$$\begin{aligned} \frac{1}{T}\sum_{t=0}^{T-1}\mathcal{L}_S(\mathbf{W}_t) \lesssim \mathrm{polylog}\big(\frac{nT}{\delta\rho_{\mathrm{clip}}}\big)\cdot\Big[\,&\frac{1}{\gamma^2\eta_{\mathrm{eff}}T} + \frac{1}{\gamma\sqrt{BT}} + \eta_{\mathrm{eff}}\big(\frac{1}{B}+\frac{1}{T}\big) + \frac{(1-\lambda)c_{\mathrm{eff}}}{\gamma\sqrt{T}} \\ &+ \big((1-\lambda)^2+\lambda^2\eta_{\mathrm{eff}}\big)\,\eta_{\mathrm{eff}}c_{\mathrm{eff}}^2 md\Big] =: A_{\mathrm{corr}}. \end{aligned}$$

Proof. Fix auxiliary failure probabilities $\delta_{\mathrm{init}} = \delta_{\mathrm{ntk}} = \delta_Z = \delta_\Delta = \delta_{\mathrm{pot}} = \frac{\delta}{5}$. Since replacing δ by a constant fraction only affects logarithmic factors by absolute constants, we suppress this distinction below. Set

$$\tau_\gamma^2 \asymp \frac{\log^2(T/\rho_{\mathrm{clip}}) + \log(n/\delta)}{\gamma^2}.$$

We choose the shifted localization radius as $\bar{R} = C_{\bar{R}}\tau_\gamma$, where $C_{\bar{R}} > 0$ is a sufficiently large universal constant. Our proof consists of the following steps.

(I). COMPARATOR CONSTRUCTION UNDER INITIALIZATION SEPARABILITY.

By Lemma 23, under the width condition

$$m \gtrsim \frac{\log(m/\delta_{\mathrm{ntk}})\big(\log^2(T/\rho_{\mathrm{clip}}) + \log(n/\delta_{\mathrm{ntk}})\big)}{\gamma^4},$$

there exists a comparator $\mathbf{W}^* = \mathbf{W}_0 + \tau_{\mathrm{ntk}}\mathbf{u}$ such that, with probability at least $1 - \delta_{\mathrm{ntk}}$ over the initialization,

$$\mathcal{L}_S(\mathbf{W}^*) \le \frac{\rho_{\mathrm{clip}}}{T} \quad \text{and} \quad \|\mathbf{W}^* - \mathbf{W}_0\|_2 \le \tau_{\mathrm{ntk}},$$

where $\tau_{\mathrm{ntk}}^2 \lesssim \frac{\log^2(T/\rho_{\mathrm{clip}})+\log(n/\delta)}{\gamma^2}$. By taking $C_{\bar{R}}$ sufficiently large, we may assume $\tau_{\mathrm{ntk}} \le \tau_\gamma$.

(II). VERIFICATION OF THE SHIFTED-BOOTSTRAP ASSUMPTIONS.

We now verify the assumptions of Lemma 22. Throughout this step, we use the same convention as in the theorem statement and suppress polylogarithmic factors in $m, n, T, \delta^{-1}, \rho_{\mathrm{clip}}^{-1}$ and polynomial factors in p and $C_{\sigma,b}$.

By the definitions of z_{δ_Z}, V_{Z,δ_Z}, V_{Δ,δ_Δ}, and $M_{\delta_{\mathrm{pot}}}$, and since all auxiliary failure probabilities are constant fractions of δ, we have

$$z_{\delta_Z} \lesssim \sqrt{md + \log(nT/\delta)}, \quad \frac{V_{Z,\delta_Z}}{T} \lesssim md + \frac{\log(nT/\delta)}{T}, \quad V_{\Delta,\delta_\Delta} \lesssim G_{\mathrm{clip}}^2\Big(\frac{T}{B} + \log\big(\frac{nT}{\delta}\big)\Big).$$

Moreover,

$$M_{\delta_{\mathrm{pot}}} \lesssim \eta G_{\mathrm{clip}}\bar{R}\sqrt{\frac{T\log(nT/\delta)}{B}} + (1-\lambda)\eta c_{\mathrm{priv}}\bar{R}\sqrt{T\log\Big(\frac{nT}{\delta}\Big)} + (1-\lambda)\eta^2 c_{\mathrm{priv}} G_{\mathrm{clip}}\sqrt{\frac{Tmd\log(nT/\delta)}{B}}.$$

Using $G_{\mathrm{clip}} = \rho_{\mathrm{clip}}G_\delta$, $c_{\mathrm{priv}} = \rho_{\mathrm{clip}}c_{\mathrm{eff}}$, $\eta_{\mathrm{eff}} = \eta\rho_{\mathrm{clip}}$, and $\bar{R} \asymp \tau_\gamma$, we obtain

$$M_{\delta_{\mathrm{pot}}} \lesssim \eta_{\mathrm{eff}}\tau_\gamma\sqrt{\frac{T\log(nT/\delta)}{B}} + (1-\lambda)\eta_{\mathrm{eff}}c_{\mathrm{eff}}\tau_\gamma\sqrt{T\log\big(\frac{nT}{\delta}\big)} + (1-\lambda)\eta_{\mathrm{eff}}^2 c_{\mathrm{eff}}\sqrt{\frac{Tmd\log(nT/\delta)}{B}},$$

up to the suppressed factors.

We next verify the localization condition required by Lemma 13. The lemma requires

$$m \gtrsim C_{\sigma,b}^2 p^3\big(\log(m/\delta) + p\big)\big(\bar{R} + \eta c_{\mathrm{priv}} z_{\delta_Z}\big)^4.$$

By the upper bound on m in the theorem, together with the bound on z_{δ_Z} and $\eta c_{\mathrm{priv}} = \eta_{\mathrm{eff}}c_{\mathrm{eff}}$, we have $\eta c_{\mathrm{priv}} z_{\delta_Z} = \eta_{\mathrm{eff}}c_{\mathrm{eff}}z_{\delta_Z} \lesssim \tau_\gamma$ up to the suppressed factors. Since $\bar{R} \asymp \tau_\gamma$, it follows that $\bar{R} + \eta c_{\mathrm{priv}} z_{\delta_Z} \lesssim \tau_\gamma$. Since $\tau_\gamma^4 \asymp \frac{(\log^2(T/\rho_{\mathrm{clip}})+\log(n/\delta))^2}{\gamma^4}$, the stated lower bound on m implies that the localization condition in Lemma 13 holds.

It remains to verify the self-consistency condition equation 10 in Lemma 14. Using $\|\mathbf{W}^* - \mathbf{W}_0\|_2^2 \le \tau_\gamma^2$, $\mathcal{L}_S(\mathbf{W}^*) \le \rho_{\mathrm{clip}}/T$, $\beta_{\mathrm{clip}} = \rho_{\mathrm{clip}}\beta$, and the definitions of η_{eff} and c_{eff}, the left-hand side of equation 10 is bounded, up to the suppressed factors, by

$$\tau_\gamma^2 + \eta_{\mathrm{eff}} + M_{\delta_{\mathrm{pot}}} + \eta_{\mathrm{eff}}^2\Big(\frac{T}{B} + 1\Big) + \eta_{\mathrm{eff}}^2 c_{\mathrm{eff}}^2\Big((1-\lambda)^2 + \lambda^2\eta_{\mathrm{eff}}\Big)Tmd.$$

Substituting the preceding estimates and using the step-size and width conditions in the theorem gives

$$\|\mathbf{W}_0 - \mathbf{W}^*\|_2^2 + \frac{8}{3}\eta T\mathcal{L}_S(\mathbf{W}^*) + M_{\delta_{\mathrm{pot}}} + 2\eta^2 V_{\Delta,\delta_\Delta} + \big((1-\lambda)^2\eta^2 c_{\mathrm{priv}}^2 + 12\beta_{\mathrm{clip}}\lambda^2\eta^3 c_{\mathrm{priv}}^2\big)V_{Z,\delta_Z} \lesssim \tau_\gamma^2.$$

Since $\bar{R} = C_{\bar{R}}\tau_\gamma$, choosing $C_{\bar{R}}$ sufficiently large verifies equation 10.

Finally, the same bound $\eta c_{\mathrm{priv}} z_{\delta_Z} = \eta_{\mathrm{eff}} c_{\mathrm{eff}} z_{\delta_Z} \lesssim \tau_\gamma$, together with $\bar{R} \asymp \tau_\gamma$ and $\|\mathbf{W}^* - \mathbf{W}_0\|_2 \le \tau_\gamma$, gives

$$\bar{R} + \|\mathbf{W}^* - \mathbf{W}_0\|_2 + \eta c_{\mathrm{priv}} z_{\delta_Z} \lesssim \tau_\gamma.$$

Choosing the constant in $R_* = C\tau_\gamma$ sufficiently large ensures

$$R_* \ge \bar{R} + \|\mathbf{W}^* - \mathbf{W}_0\|_2 + \eta c_{\mathrm{priv}} z_{\delta_Z}.$$

Therefore, all assumptions of Lemma 22 are verified.

(III). APPLICATION OF THE SHIFTED-BOOTSTRAP LEMMA.

Applying Lemma 22 and using $\mathcal{L}_S(\mathbf{W}^*) \le \rho_{\mathrm{clip}}/T$ and $\|\mathbf{W}_0 - \mathbf{W}^*\|_2^2 \le \tau_\gamma^2$, we obtain that, conditioned on the initialization and comparator events, with probability at least $1 - (\delta_Z + \delta_\Delta + \delta_{\mathrm{pot}})$ over the algorithmic randomness,

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathcal{L}_S(\mathbf{W}_t) \lesssim \frac{1}{T} + \frac{\tau_\gamma^2}{\eta_{\mathrm{eff}} T} + \frac{M_{\delta_{\mathrm{pot}}}}{\eta_{\mathrm{eff}} T} + \eta_{\mathrm{eff}}\Big(\frac{1}{B} + \frac{1}{T}\Big) + \Big((1-\lambda)^2\eta_{\mathrm{eff}} + \lambda^2\eta_{\mathrm{eff}}^2\Big) c_{\mathrm{eff}}^2 \Big(md + \frac{\log(nT/\delta)}{T}\Big). \tag{16}$$

Substituting the estimate of $M_{\delta_{\mathrm{pot}}}$ from Step (ii) into equation 16 and using $\tau_\gamma^2 \asymp (\log^2(T/\rho_{\mathrm{clip}}) + \log(n/\delta))/\gamma^2$, we obtain, up to the suppressed logarithmic and p-dependent factors,

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathcal{L}_S(\mathbf{W}_t) \lesssim \mathrm{polylog}\Big(\frac{nT}{\delta\rho_{\mathrm{clip}}}\Big) \cdot \Big[\frac{1}{\gamma^2\eta_{\mathrm{eff}} T} + \frac{1}{\gamma\sqrt{BT}} + \eta_{\mathrm{eff}}\Big(\frac{1}{B} + \frac{1}{T}\Big) + \frac{(1-\lambda)c_{\mathrm{eff}}}{\sqrt{T}}\Big(\frac{1}{\gamma} + \frac{\eta_{\mathrm{eff}}\sqrt{md}}{\sqrt{B}}\Big) + \Big((1-\lambda)^2 + \lambda^2\eta_{\mathrm{eff}}\Big)\eta_{\mathrm{eff}} c_{\mathrm{eff}}^2 md\Big] =: A_{\mathrm{corr}}.$$

Here we also used $m \gtrsim \log(nT/\delta)/(Td)$ and $1/T \lesssim 1/(\gamma^2\eta_{\mathrm{eff}} T)$. This is exactly the claimed bound.

(IV). PROBABILITY ESTIMATE.

The above argument holds on the intersection of the initialization event $\mathcal{E}_{\delta_{\mathrm{init}}}$, the comparator event from Lemma 23, and the algorithmic good event from Lemma 22. Their total failure probability is at most $\delta_{\mathrm{init}} + \delta_{\mathrm{ntk}} + \delta_Z + \delta_\Delta + \delta_{\mathrm{pot}} = \delta$. Therefore, the stated optimization bound holds with probability at least $1 - \delta$. □

D PROOFS FOR POPULATION OF DP-SGD WITH CORRELATED NOISE

We introduce the concept of algorithmic stability to control generalization gap, and then combining generalization gap with optimization risk bound to get population risk bound.

For a randomized algorithm $\mathcal{A}$, let $\mathcal{A}(S) \in \mathbb{R}^{mdp}$ be the output of $\mathcal{A}$ based on dataset S. The on-average argument stability measures the on-average sensitivity of the output up to the perturbation of the dataset.

Definition 2 (On-average argument stability Lei & Ying (2020))**.** Let $S = \{z_1, \ldots, z_n\}$ and $\widetilde{S} = \{z_1', \ldots, z_n'\}$ be drawn independently from $\mathcal{P}$. For any $i \in [n]$, define $S^{(i)} = \{z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_n\}$. Let $\mathcal{A}(S)$ and $\mathcal{A}(S^{(i)})$ be produced by an randomized algorithm $\mathcal{A}$ based on S and $S^{(i)}$ respectively. We say $\mathcal{A}$ is on-average argument ϵ-stable if

$$\mathbb{E}_{S,\widetilde{S},\mathcal{A}}\Big[\frac{1}{n}\sum_{i=1}^{n} \|\mathcal{A}(S) - \mathcal{A}(S^{(i)})\|_2\Big] \le \epsilon.$$

We consider using the connection between the on-average argument stability and generalization error bounds Lei & Ying (2020).

Lemma 25 (Lei & Ying (2020))**.** *If $\mathcal{A}$ is on-average argument ϵ-stable and the loss ℓ is L-Lipschitz with respect to $\mathcal{A}(S)$, then*

$$\mathbb{E}_{S,\mathcal{A}}\big[\mathcal{L}(\mathcal{A}(S)) - \mathcal{L}_S(\mathcal{A}(S))\big] \le 2L\epsilon.$$

To compare the outputs on two neighboring datasets S and $S^{(i)}$, we couple two runs of Algorithm 1 by using the same initialization, the same sequence of mini-batches, and the same Gaussian noise sequence. Under this coupling, the additive perturbations appear identically in the two updates and cancel in the difference of the iterates. The resulting stability recursion therefore depends only on the gradient part of the projected update. It is important, however, that the empirical losses appearing below are still evaluated along the two coupled private trajectories.

Let

$$\Delta_t^{(i)} = \|\mathbf{W}_t - \mathbf{W}_t^{(i)}\|_2, \qquad \mathbf{W}_{\alpha,t}^{(i)} = \alpha\mathbf{W}_t + (1-\alpha)\mathbf{W}_t^{(i)}, \quad \alpha \in [0,1].$$

For each $t \ge 0$, define the shared mini-batch loss

$$\mathcal{L}_{t,i}^{\mathrm{sh}}(\mathbf{W}) = \frac{1}{B}\sum_{j\in\mathcal{B}_t\setminus\{i\}} \ell\left(y_j f_{\mathbf{W}}(\mathbf{x}_j)\right),$$

where the sum is interpreted as 0 if $\mathcal{B}_t \setminus \{i\} = \varnothing$. We also define

$$\mathcal{L}_{\mathcal{B}_t}(\mathbf{W}) = \frac{1}{B}\sum_{j\in\mathcal{B}_t} \ell\left(y_j f_{\mathbf{W}}(\mathbf{x}_j)\right),$$

$$\mathcal{L}_{\mathcal{B}_t}^{(i)}(\mathbf{W}) = \frac{1}{B}\sum_{j\in\mathcal{B}_t} \ell\left(y_j^{(i)} f_{\mathbf{W}}(\mathbf{x}_j^{(i)})\right) = \mathcal{L}_{t,i}^{\mathrm{sh}}(\mathbf{W}) + \frac{1}{B}\mathbf{1}_{\{i\in\mathcal{B}_t\}}\ell\left(y_i' f_{\mathbf{W}}(\mathbf{x}_i')\right).$$

Further, let

$$a_{\delta,m} = \frac{2C_{\sigma,b}\, p^{\frac{3}{2}}\left(\sqrt{\log(\frac{m}{\delta})} + \sqrt{p}\right)}{\sqrt{m}}, \qquad b_{\delta,m} = C_{\sigma,b}\, p\Big(\sqrt{p} + \sqrt{\tfrac{\log(1/\delta)}{m}}\Big).$$

Lemma 26. *Suppose Assumption 1 holds and the event $\mathcal{E}_\delta$ occurs. Let $\{\mathbf{W}_t\}_{t\ge0}$ and $\{\mathbf{W}_t^{(i)}\}_{t\ge0}$ be two coupled runs of Algorithm 1 with $\lambda > 0$ on S and $S^{(i)}$, respectively, driven by the same initialization, the same mini-batch sequence, and the same Gaussian noise sequence. Assume $\eta_{\mathrm{eff}} \le \frac{1}{12C_{\sigma,b}p^3}$ and $m \gtrsim C_{\sigma,b}^2 p^3(\log(m/\delta)+p)R_*^4$. Then, for all $t \ge 0$, it holds that*

$$\Delta_{t+1}^{(i)} \le \big(1 + a_{\delta,m}\eta(\mathcal{L}_{\mathcal{B}_t}(\mathbf{W}_t) + \mathcal{L}_{\mathcal{B}_t}^{(i)}(\mathbf{W}_t^{(i)}))\big)\Delta_t^{(i)} + \frac{\eta b_{\delta,m}}{B}\mathbf{1}_{\{i\in\mathcal{B}_t\}}\big(\ell(y_i f_{\mathbf{W}_t}(\mathbf{x}_i)) + \ell(y_i' f_{\mathbf{W}_t^{(i)}}(\mathbf{x}_i'))\big).$$

Proof. For $z = (\mathbf{x}, y)$, define $\phi_z(\mathbf{W}) = \mathrm{clip}_{C_{\mathrm{clip}}}\big(\nabla\ell(y f_{\mathbf{W}}(\mathbf{x}))\big)$. Let $\zeta_t = \frac{C_{\mathrm{clip}}}{B}\xi_t$ denote the common additive Gaussian perturbation in the two coupled runs. Since the two algorithms use the same initialization, the same mini-batches, and the same Gaussian noise sequence, the projected updates can be written as

$$\mathbf{W}_{t+1} = \Pi_{\mathcal{K}}\big(\mathbf{W}_t - \eta(v_t + \zeta_t)\big), \qquad \mathbf{W}_{t+1}^{(i)} = \Pi_{\mathcal{K}}\big(\mathbf{W}_t^{(i)} - \eta(v_t^{(i)} + \zeta_t)\big),$$

where $v_t = \frac{1}{B}\sum_{j\in\mathcal{B}_t}\phi_{z_j}(\mathbf{W}_t)$ and $v_t^{(i)} = \frac{1}{B}\sum_{j\in\mathcal{B}_t}\phi_{z_j^{(i)}}(\mathbf{W}_t^{(i)})$.

By the non-expansiveness of the Euclidean projection,

$$\Delta_{t+1}^{(i)} \le \big\|\mathbf{W}_t - \eta v_t - \big(\mathbf{W}_t^{(i)} - \eta v_t^{(i)}\big)\big\|_2, \tag{17}$$

where the common noise term cancels exactly.

We split off the possibly replaced sample. Define $v_{t,i}^{\mathrm{sh}}(\mathbf{W}) = \frac{1}{B}\sum_{j\in\mathcal{B}_t\setminus\{i\}}\phi_{z_j}(\mathbf{W})$. Then

$$v_t = v_{t,i}^{\mathrm{sh}}(\mathbf{W}_t) + \frac{1}{B}\mathbf{1}_{\{i\in\mathcal{B}_t\}}\phi_{z_i}(\mathbf{W}_t), \qquad v_t^{(i)} = v_{t,i}^{\mathrm{sh}}(\mathbf{W}_t^{(i)}) + \frac{1}{B}\mathbf{1}_{\{i\in\mathcal{B}_t\}}\phi_{z_i'}(\mathbf{W}_t^{(i)}).$$

Substituting these decompositions into equation 17 gives

$$\Delta_{t+1}^{(i)} \le \big\|\mathbf{W}_t - \eta v_{t,i}^{\mathrm{sh}}(\mathbf{W}_t) - \big(\mathbf{W}_t^{(i)} - \eta v_{t,i}^{\mathrm{sh}}(\mathbf{W}_t^{(i)})\big)\big\|_2 + \frac{\eta}{B}\mathbf{1}_{\{i\in\mathcal{B}_t\}}\big(\|\phi_{z_i}(\mathbf{W}_t)\|_2 + \|\phi_{z_i'}(\mathbf{W}_t^{(i)})\|_2\big). \tag{18}$$

We first bound the first term in equation 18. For a fixed $z = (\mathbf{x}, y)$, consider the clipped-gradient map ϕ_z defined above. This map is locally Lipschitz and hence differentiable almost everywhere. We use $\nabla\phi_z(\mathbf{W})$ to denote its Jacobian with respect to $\mathbf{W}$. When clipping is inactive,

$$\nabla\phi_z(\mathbf{W}) = \ell''(yf_{\mathbf{W}}(\mathbf{x}))\nabla f_{\mathbf{W}}(\mathbf{x})\nabla f_{\mathbf{W}}(\mathbf{x})^\top + y\ell'(yf_{\mathbf{W}}(\mathbf{x}))\nabla^2 f_{\mathbf{W}}(\mathbf{x}).$$

When clipping is active, writing $\alpha_z(\mathbf{W}) = \frac{C_{\mathrm{clip}}}{\|\nabla\ell(yf_{\mathbf{W}}(\mathbf{x}))\|_2}$ and $u_z(\mathbf{W}) = \frac{\nabla\ell(yf_{\mathbf{W}}(\mathbf{x}))}{\|\nabla\ell(yf_{\mathbf{W}}(\mathbf{x}))\|_2}$, we have

$$\nabla\phi_z(\mathbf{W}) = \alpha_z(\mathbf{W})\big(\mathbf{I} - u_z(\mathbf{W})u_z(\mathbf{W})^\top\big)\nabla^2\ell(yf_{\mathbf{W}}(\mathbf{x})).$$

Since $u_z(\mathbf{W})$ is parallel to $\nabla f_{\mathbf{W}}(\mathbf{x})$, it holds $\big(\mathbf{I} - u_z(\mathbf{W})u_z(\mathbf{W})^\top\big)\nabla f_{\mathbf{W}}(\mathbf{x})\nabla f_{\mathbf{W}}(\mathbf{x})^\top = \mathbf{0}$. Thus, in both cases we can write

$$\nabla\phi_z(\mathbf{W}) = P_z(\mathbf{W}) + E_z(\mathbf{W}),$$

where $P_z(\mathbf{W})$ is the positive-semidefinite rank-one term $\ell''(yf_{\mathbf{W}}(\mathbf{x}))\nabla f_{\mathbf{W}}(\mathbf{x})\nabla f_{\mathbf{W}}(\mathbf{x})^\top$ when clipping is inactive and is zero when clipping is active. The remaining term is denoted by $E_z(\mathbf{W})$.

On $\mathcal{E}_\delta$, we have

$$P_z(\mathbf{W}) \succeq 0, \qquad \|P_z(\mathbf{W})\|_{\mathrm{op}} \le \rho_{\mathrm{clip}}G_\delta^2, \qquad \|E_z(\mathbf{W})\|_{\mathrm{op}} \le \frac{a_{\delta,m}}{2}\ell(yf_{\mathbf{W}}(\mathbf{x})).$$

Indeed, in the inactive regime, using $\ell''(u) \le |\ell'(u)|$ gives

$$\|P_z(\mathbf{W})\|_{\mathrm{op}} \le \min\{C_{\mathrm{clip}}, G_\delta\}G_\delta = \rho_{\mathrm{clip}}G_\delta^2,$$

while in the active regime $P_z(\mathbf{W}) = 0$. Moreover, in both regimes, $\alpha_z(\mathbf{W}) \le 1$, $|\ell'(u)| \le \ell(u)$, and Lemma 10 imply the stated bound on $E_z(\mathbf{W})$.

Averaging over the shared samples gives

$$\nabla v_{t,i}^{\mathrm{sh}}(\mathbf{W}) = P_{t,i}(\mathbf{W}) + E_{t,i}(\mathbf{W}),$$

where $P_{t,i}(\mathbf{W}) \succeq 0$, $\|P_{t,i}(\mathbf{W})\|_{\mathrm{op}} \le \rho_{\mathrm{clip}}G_\delta^2$ and $\|E_{t,i}(\mathbf{W})\|_{\mathrm{op}} \le \frac{a_{\delta,m}}{2}\mathcal{L}_{t,i}^{\mathrm{sh}}(\mathbf{W})$.

By absolute continuity along the line segment,

$$v_{t,i}^{\mathrm{sh}}(\mathbf{W}_t) - v_{t,i}^{\mathrm{sh}}(\mathbf{W}_t^{(i)}) = (\bar{P}_{t,i} + \bar{E}_{t,i})(\mathbf{W}_t - \mathbf{W}_t^{(i)}),$$

where $\bar{P}_{t,i} = \int_0^1 P_{t,i}(\mathbf{W}_{\alpha,t}^{(i)})\,d\alpha$ and $\bar{E}_{t,i} = \int_0^1 E_{t,i}(\mathbf{W}_{\alpha,t}^{(i)})\,d\alpha$.

Since $\bar{P}_{t,i} \succeq 0$ and $\|\bar{P}_{t,i}\|_{\mathrm{op}} \le \rho_{\mathrm{clip}}G_\delta^2$, using $G_\delta^2 \le 4C_{\sigma,b}p^3$ gives

$$\eta\|\bar{P}_{t,i}\|_{\mathrm{op}} \le \eta_{\mathrm{eff}}G_\delta^2 \le \frac{1}{3}, \qquad \|\mathbf{I} - \eta\bar{P}_{t,i}\|_{\mathrm{op}} \le 1.$$

We next control $\bar{E}_{t,i}$. Both iterates lie in $\mathcal{K} = \mathcal{B}(\mathbf{W}_0, R_*)$, hence the whole segment $\{\mathbf{W}_{\alpha,t}^{(i)} : \alpha \in [0,1]\}$ lies in $\mathcal{K}$, and $\|\mathbf{W}_t - \mathbf{W}_t^{(i)}\|_2 \le 2R_*$. Under the width condition $m \gtrsim C_{\sigma,b}^2 p^3(\log(m/\delta) + p)R_*^4$, applying Lemma 8 with $G = \mathcal{L}_{t,i}^{\mathrm{sh}}$ and $D = 2R_*$ to the shared mini-batch loss $\mathcal{L}_{t,i}^{\mathrm{sh}}$ yields

$$\max_{\alpha\in[0,1]} \mathcal{L}_{t,i}^{\mathrm{sh}}(\mathbf{W}_{\alpha,t}^{(i)}) \le 2\max\big\{\mathcal{L}_{t,i}^{\mathrm{sh}}(\mathbf{W}_t), \mathcal{L}_{t,i}^{\mathrm{sh}}(\mathbf{W}_t^{(i)})\big\}.$$

Since the loss is nonnegative, then $\mathcal{L}_{t,i}^{\mathrm{sh}}(\mathbf{W}_t) \le \mathcal{L}_{\mathcal{B}_t}(\mathbf{W}_t)$ and $\mathcal{L}_{t,i}^{\mathrm{sh}}(\mathbf{W}_t^{(i)}) \le \mathcal{L}_{\mathcal{B}_t}^{(i)}(\mathbf{W}_t^{(i)})$. Therefore,

$$\|\bar{E}_{t,i}\|_{\mathrm{op}} \le a_{\delta,m}\big(\mathcal{L}_{\mathcal{B}_t}(\mathbf{W}_t) + \mathcal{L}_{\mathcal{B}_t}^{(i)}(\mathbf{W}_t^{(i)})\big).$$

Therefore,

$$\big\|\mathbf{W}_t - \eta v_{t,i}^{\mathrm{sh}}(\mathbf{W}_t) - \big(\mathbf{W}_t^{(i)} - \eta v_{t,i}^{\mathrm{sh}}(\mathbf{W}_t^{(i)})\big)\big\|_2 \le \big(1 + a_{\delta,m}\eta(\mathcal{L}_{\mathcal{B}_t}(\mathbf{W}_t) + \mathcal{L}_{\mathcal{B}_t}^{(i)}(\mathbf{W}_t^{(i)}))\big)\Delta_t^{(i)}.$$

It remains to bound the two single-sample terms in equation 18. Since clipping cannot increase the norm,

$$\|\phi_z(\mathbf{W})\|_2 \le \|\nabla \ell(y f_{\mathbf{W}}(\mathbf{x}))\|_2 \le b_{\delta,m}\ell(y f_{\mathbf{W}}(\mathbf{x})),$$

where we used $|\ell'(u)| \le \ell(u)$ and Lemma 10. Hence

$$\|\phi_{z_i}(\mathbf{W}_t)\|_2 \le b_{\delta,m}\ell(y_i f_{\mathbf{W}_t}(\mathbf{x}_i)), \qquad \|\phi_{z_i'}(\mathbf{W}_t^{(i)})\|_2 \le b_{\delta,m}\ell(y_i' f_{\mathbf{W}_t^{(i)}}(\mathbf{x}_i')).$$

Substituting these estimates into equation 18 yields

$$\Delta_{t+1}^{(i)} \le \big(1+a_{\delta,m}\eta(\mathcal{L}_{\mathcal{B}_t}(\mathbf{W}_t)+\mathcal{L}_{\mathcal{B}_t}^{(i)}(\mathbf{W}_t^{(i)}))\big)\Delta_t^{(i)} + \frac{\eta b_{\delta,m}}{B}\mathbf{1}_{\{i\in\mathcal{B}_t\}}\big(\ell(y_i f_{\mathbf{W}_t}(\mathbf{x}_i))+\ell(y_i' f_{\mathbf{W}_t^{(i)}}(\mathbf{x}_i'))\big).$$

This completes the proof. □

We first control the cumulative mini-batch losses along the two coupled trajectories. For $\delta_f \in (0,1)$, define the initialization-output event

$$\mathcal{E}_{\delta_f}^{\text{out}} = \Big\{ \max_{1\le j\le n} |f_{\mathbf{W}_0}(\mathbf{x}_j)| \le B_b\sqrt{2p\log\big(4n/\delta_f\big)} \text{ and } \max_{1\le j\le n} |f_{\mathbf{W}_0}(\mathbf{x}_j')| \le B_b\sqrt{2p\log\big(4n/\delta_f\big)}\Big\},$$

where $\widetilde{S} = \{(\mathbf{x}_j', y_j')\}_{j=1}^n$ is an independent copy of S.

Lemma 27. *Suppose Assumption 1 holds. Then $\mathbb{P}\big(\mathcal{E}_{\delta_f}^{\text{out}}\big) \ge 1-\delta_f$. Moreover, on the event $\mathcal{E}_\delta \cap \mathcal{E}_{\delta_f}^{\text{out}}$, for every $t \ge 0$ and every $i, j \in [n]$,*

$$\max\big\{|f_{\mathbf{W}_t}(\mathbf{x}_j)|, |f_{\mathbf{W}_t^{(i)}}(\mathbf{x}_j^{(i)})|\big\} \le F_{\delta,\delta_f} \text{ and } \max\Big\{\ell\big(y_j f_{\mathbf{W}_t}(\mathbf{x}_j)\big), \ell\big(y_j^{(i)} f_{\mathbf{W}_t^{(i)}}(\mathbf{x}_j^{(i)})\big)\Big\} \le U_{\delta,\delta_f},$$

where $F_{\delta,\delta_f} = B_b\sqrt{2p\log\Big(\frac{4n}{\delta_f}\Big)} + b_{\delta,m}R_*$ and $U_{\delta,\delta_f} = \log\big(1+e^{F_{\delta,\delta_f}}\big) \le \log 2 + F_{\delta,\delta_f}$.

Proof. We first control the initial outputs. Fix any input $\mathbf{x} \in \mathbb{R}^d$. Recall that

$$f_{\mathbf{W}_0}(\mathbf{x}) = \frac{1}{\sqrt{m}}\mathbf{c}^\top \mathbf{h}\Big(\sigma\Big(\frac{1}{\sqrt{d}}\mathbf{W}_0\mathbf{h}(\mathbf{x})\Big)\Big).$$

Conditioned on $(\mathbf{W}_0, \mathbf{x})$, the only randomness comes from $\mathbf{c} \sim \mathcal{N}(0, \mathbf{I}_{mp})$. Hence,

$$f_{\mathbf{W}_0}(\mathbf{x}) \sim \mathcal{N}\Big(0, \frac{1}{m}\Big\|\mathbf{h}\Big(\sigma\Big(\frac{1}{\sqrt{d}}\mathbf{W}_0\mathbf{h}(\mathbf{x})\Big)\Big)\Big\|_2^2\Big).$$

Since each spline basis function is bounded by B_b on the range of σ, then we have

$$\Big\|\mathbf{h}\Big(\sigma\Big(\frac{1}{\sqrt{d}}\mathbf{W}_0\mathbf{h}(\mathbf{x})\Big)\Big)\Big\|_2^2 \le mpB_b^2.$$

Applying the Gaussian tail bound to the $2n$ points $\mathbf{x}_1, \dots, \mathbf{x}_n, \mathbf{x}_1', \dots, \mathbf{x}_n'$ and taking a union bound yields $\mathbb{P}\big(\mathcal{E}_{\delta_f}^{\text{out}}\big) \ge 1-\delta_f$.

We next extend the bound from the initialization to the whole projected trajectories. On the event $\mathcal{E}_\delta$ (see equation 1), Lemma 10 implies that

$$\|\nabla f_{\mathbf{W}}(\mathbf{x})\|_2 \le b_{\delta,m}, \qquad \forall\, \mathbf{W} \in \mathbb{R}^{mdp},\ \mathbf{x} \in \mathcal{X}.$$

Since both trajectories are projected onto $\mathcal{K} = \mathcal{B}(\mathbf{W}_0, R_*)$, we have

$$\|\mathbf{W}_t - \mathbf{W}_0\|_2 \le R_*, \qquad \|\mathbf{W}_t^{(i)} - \mathbf{W}_0\|_2 \le R_*.$$

Therefore, by the mean-value theorem,

$$\max\big\{|f_{\mathbf{W}_t}(\mathbf{x}_j) - f_{\mathbf{W}_0}(\mathbf{x}_j)|, |f_{\mathbf{W}_t^{(i)}}(\mathbf{x}_j^{(i)}) - f_{\mathbf{W}_0}(\mathbf{x}_j^{(i)})|\big\} \le b_{\delta,m}R_*.$$

Since $\mathbf{x}_j^{(i)}$ is either $\mathbf{x}_j$ or $\mathbf{x}_i'$, combining this estimate with $\mathcal{E}_{\delta_f}^{\text{out}}$ gives

$$\max\big\{|f_{\mathbf{W}_t}(\mathbf{x}_j)|, |f_{\mathbf{W}_t^{(i)}}(\mathbf{x}_j^{(i)})|\big\} \le B_b\sqrt{2p\log\Big(\frac{4n}{\delta_f}\Big)} + b_{\delta,m}R_* = F_{\delta,\delta_f}.$$

Finally, the logistic loss satisfies $\ell(u) = \log(1+e^{-u}) \le \log(1+e^{|u|}) \le \log 2 + |u|$. Hence,

$$\max\Big\{\ell\big(y_j f_{\mathbf{W}_t}(\mathbf{x}_j)\big), \ell\big(y_j^{(i)} f_{\mathbf{W}_t^{(i)}}(\mathbf{x}_j^{(i)})\big)\Big\} \le U_{\delta,\delta_f},$$

which completes the proof. □

Lemma 28 (Freedman's inequality)**.** *Let $\{X_t, \mathscr{F}_t\}_{t=1}^T$ be a martingale difference sequence, that is, X_t is $\mathscr{F}_t$-measurable and $\mathbb{E}[X_t \mid \mathscr{F}_{t-1}] = 0$ a.s. for all $t \in [T]$. If $|X_t| \le L$ a.s. for all $t \in [T]$, and the predictable quadratic variation $V_T = \sum_{t=1}^T \mathbb{E}[X_t^2 \mid \mathscr{F}_{t-1}] \le v$ a.s., then for any $\delta \in (0,1)$,*

$$\mathbb{P}\Big(\sum_{t=1}^T X_t \ge \sqrt{2v\log(1/\delta)} + \frac{2L}{3}\log(1/\delta)\Big) \le \delta.$$

Lemma 29. *Suppose the assumptions of Lemma 26 hold, and let $\delta_{\mathrm{mb}} \in (0,1)$. Then, conditioned on S, $S^{(i)}$, and the initialization satisfying $\mathcal{E}_\delta \cap \mathcal{E}_{\delta_f}^{\mathrm{out}}$, with probability at least $1 - 2\delta_{\mathrm{mb}}$ over the remaining algorithmic randomness,*

$$\sum_{t=0}^{T-1} \mathcal{L}_{\mathcal{B}_t}(\mathbf{W}_t) \le \sum_{t=0}^{T-1} \mathcal{L}_S(\mathbf{W}_t) + C\,U_{\delta,\delta_f}\Big(\sqrt{\frac{T\log(1/\delta_{\mathrm{mb}})}{B}} + \log\Big(\frac{1}{\delta_{\mathrm{mb}}}\Big)\Big),$$
$$\sum_{t=0}^{T-1} \mathcal{L}^{(i)}_{\mathcal{B}_t}(\mathbf{W}^{(i)}_t) \le \sum_{t=0}^{T-1} \mathcal{L}_{S^{(i)}}(\mathbf{W}^{(i)}_t) + C\,U_{\delta,\delta_f}\Big(\sqrt{\frac{T\log(1/\delta_{\mathrm{mb}})}{B}} + \log\Big(\frac{1}{\delta_{\mathrm{mb}}}\Big)\Big),$$

where $C > 0$ is an absolute constant.

Proof. We prove the first inequality, and the second one follows in the same way. Define the pre-sampling filtration

$$\mathscr{F}_t = \sigma\big(S, S^{(i)}, \mathbf{W}_0, \mathcal{B}_0, \dots, \mathcal{B}_{t-1}, Z_0, \dots, Z_{t-1}\big), \qquad t \in [T]\cup\{0\}.$$

Thus, $\mathscr{F}_t$ contains all randomness revealed before the fresh mini-batch $\mathcal{B}_t$ is drawn.

For each $t = 0, \dots, T-1$, define $X_t = \mathcal{L}_{\mathcal{B}_t}(\mathbf{W}_t) - \mathcal{L}_S(\mathbf{W}_t)$. Conditioned on $\mathscr{F}_t$, the iterate $\mathbf{W}_t$ is fixed and the only randomness in X_t comes from the fresh mini-batch $\mathcal{B}_t$. Since $\mathcal{B}_t$ is sampled uniformly without replacement, it holds $\mathbb{E}_{\mathcal{A}}[X_t \mid \mathscr{F}_t] = 0$. Moreover, on the event $\mathcal{E}_\delta \cap \mathcal{E}_{\delta_f}^{\mathrm{out}}$, Lemma 27 gives

$$|X_t| \le \sup_{z\in S} |\ell(y f_{\mathbf{W}_t}(\mathbf{x}))| \le U_{\delta,\delta_f}.$$

We next bound the conditional variance. Conditioned on $\mathscr{F}_t$, let $u_j = \ell(y_j f_{\mathbf{W}_t}(\mathbf{x}_j))$, so that $0 \le u_j \le U_{\delta,\delta_f}$ for all $j \in [n]$. By Lemma 4,

$$\mathbb{E}_A[X_t^2 \mid \mathscr{F}_t] \le \frac{U_{\delta,\delta_f}^2}{B}.$$

Thus, regarding X_t as the $(t+1)$-st martingale difference with respect to the filtration $\{\mathscr{F}_t\}_{t=0}^T$, Lemma 28 with $v = TU_{\delta,\delta_f}^2/B$ implies that, with probability at least $1-\delta_{\mathrm{mb}}$,

$$\sum_{t=0}^{T-1} X_t \le \sqrt{\frac{2TU_{\delta,\delta_f}^2}{B}\log\Big(\frac{1}{\delta_{\mathrm{mb}}}\Big)} + \frac{2U_{\delta,\delta_f}}{3}\log\Big(\frac{1}{\delta_{\mathrm{mb}}}\Big).$$

Using

$$\sum_{t=0}^{T-1} \mathcal{L}_{\mathcal{B}_t}(\mathbf{W}_t) = \sum_{t=0}^{T-1} \mathcal{L}_S(\mathbf{W}_t) + \sum_{t=0}^{T-1} X_t$$

proves the first inequality.

The second inequality follows in the same way. A union bound over the two estimates yields probability at least $1 - 2\delta_{\mathrm{mb}}$. □

For each $i \in [n]$ and $t \ge 0$, define $M_t^{(i)} = \eta a_{\delta,m}\big(\mathcal{L}_{\mathcal{B}_t}(\mathbf{W}_t) + \mathcal{L}^{(i)}_{\mathcal{B}_t}(\mathbf{W}^{(i)}_t)\big)$. Let $\mathcal{E}^{(i)}_{\mathrm{opt}}$ denote the event on which $\frac{1}{T}\sum_{t=0}^{T-1}\mathcal{L}_S(\mathbf{W}_t) \le A_{\mathrm{corr}}$ and $\frac{1}{T}\sum_{t=0}^{T-1}\mathcal{L}_{S^{(i)}}(\mathbf{W}^{(i)}_t) \le A_{\mathrm{corr}}$ hold.

For each $i \in [n]$, let $\mathcal{E}_{\mathrm{mb}}^{(i)}$ be the event on which both inequalities in Lemma 29 hold, namely,

$$\sum_{t=0}^{T-1} \mathcal{L}_{\mathcal{B}_t}(\mathbf{W}_t) \le \sum_{t=0}^{T-1} \mathcal{L}_S(\mathbf{W}_t) + C\, U_{\delta,\delta_f}\Big(\sqrt{\frac{T\log(1/\delta_{\mathrm{mb}})}{B}} + \log\Big(\frac{1}{\delta_{\mathrm{mb}}}\Big)\Big),$$
$$\sum_{t=0}^{T-1} \mathcal{L}^{(i)}_{\mathcal{B}_t}(\mathbf{W}_t^{(i)}) \le \sum_{t=0}^{T-1} \mathcal{L}_{S^{(i)}}(\mathbf{W}_t^{(i)}) + C\, U_{\delta,\delta_f}\Big(\sqrt{\frac{T\log(1/\delta_{\mathrm{mb}})}{B}} + \log\Big(\frac{1}{\delta_{\mathrm{mb}}}\Big)\Big).$$

Lemma 30. *Suppose the assumptions of Lemma 29 hold. Assume*

$$m \gtrsim C_{\sigma,b}^2 p^3\big(\log(m/\delta)+p\big)\eta^2\Big(TA_{\mathrm{corr}} + U_{\delta,\delta_f}\sqrt{\frac{T\log(1/\delta_{\mathrm{mb}})}{B}} + U_{\delta,\delta_f}\log\Big(\frac{1}{\delta_{\mathrm{mb}}}\Big)\Big)^2.$$

Then, on the event $\mathcal{E}_{\mathrm{opt}}^{(i)} \cap \mathcal{E}_{\mathrm{mb}}^{(i)} \cap \mathcal{E}_{\delta_f}^{\mathrm{out}}$, it holds $\sum_{t=0}^{T-1} M_t^{(i)} \le 1$.

Proof. On the event $\mathcal{E}_{\mathrm{mb}}^{(i)}$, Lemma 29 implies

$$\sum_{t=0}^{T-1} M_t^{(i)} = \eta a_{\delta,m} \sum_{t=0}^{T-1}\big(\mathcal{L}_{\mathcal{B}_t}(\mathbf{W}_t) + \mathcal{L}^{(i)}_{\mathcal{B}_t}(\mathbf{W}_t^{(i)})\big)$$
$$\le \eta a_{\delta,m}\Big[\sum_{t=0}^{T-1}\mathcal{L}_S(\mathbf{W}_t) + \sum_{t=0}^{T-1}\mathcal{L}_{S^{(i)}}(\mathbf{W}_t^{(i)}) + 2C\, U_{\delta,\delta_f}\Big(\sqrt{\frac{T\log(1/\delta_{\mathrm{mb}})}{B}} + \log\Big(\frac{1}{\delta_{\mathrm{mb}}}\Big)\Big)\Big].$$

On $\mathcal{E}_{\mathrm{opt}}^{(i)}$, each of the first two sums is at most TA_{corr}. Hence,

$$\sum_{t=0}^{T-1} M_t^{(i)} \le 2\eta a_{\delta,m}\Big(TA_{\mathrm{corr}} + C\, U_{\delta,\delta_f}\Big(\sqrt{\frac{T\log(1/\delta_{\mathrm{mb}})}{B}} + \log\Big(\frac{1}{\delta_{\mathrm{mb}}}\Big)\Big)\Big).$$

The claim follows by substituting the definition of $a_{\delta,m}$ and using the assumed lower bound on m. □

Define $R_t^{(i)} = \frac{\eta b_{\delta,m}}{B}\mathbf{1}_{\{i\in\mathcal{B}_t\}}\big(\ell(y_i f_{\mathbf{W}_t}(\mathbf{x}_i)) + \ell(y_i' f_{\mathbf{W}_t^{(i)}}(\mathbf{x}_i'))\big)$.

Lemma 31. *Under the assumptions of Lemma 30, define $\mathcal{E}_{\mathrm{stab}}^{(i)} = \mathcal{E}_{\mathrm{opt}}^{(i)} \cap \mathcal{E}_{\mathrm{mb}}^{(i)} \cap \mathcal{E}_{\delta_f}^{\mathrm{out}}$. Then, on the event $\mathcal{E}_{\mathrm{stab}}^{(i)}$, for every $t = 0,\dots,T-1$, it holds $\Delta_{t+1}^{(i)} \le e\sum_{k=0}^{t} R_k^{(i)}$.*

Proof. By Lemma 26, for every $t = 0,\dots,T-1$,

$$\Delta_{t+1}^{(i)} \le \big(1 + M_t^{(i)}\big)\Delta_t^{(i)} + R_t^{(i)}.$$

Iterating this recursion and using $\Delta_0^{(i)} = 0$, we obtain

$$\Delta_{t+1}^{(i)} \le \sum_{k=0}^{t} R_k^{(i)} \prod_{s=k+1}^{t}\big(1 + M_s^{(i)}\big).$$

On the event $\mathcal{E}_{\mathrm{stab}}^{(i)}$, Lemma 30 yields $\sum_{s=0}^{T-1} M_s^{(i)} \le 1$. Since $M_s^{(i)} \ge 0$, for every $0 \le k \le t \le T-1$,

$$\prod_{s=k+1}^{t}\big(1 + M_s^{(i)}\big) \le \exp\Big(\sum_{s=k+1}^{t} M_s^{(i)}\Big) \le e,$$

where we used $1 + x \le e^x$ for $x \ge 0$. Substituting this estimate into the previous display gives

$$\Delta_{t+1}^{(i)} \le e\sum_{k=0}^{t} R_k^{(i)},$$

which completes the proof. □

For each $i \in [n]$, define the overall good event $\widehat{\mathcal{E}}^{(i)} = \mathcal{E}_\delta \cap \mathcal{E}_{\mathrm{opt}}^{(i)} \cap \mathcal{E}_{\mathrm{mb}}^{(i)} \cap \mathcal{E}_{\delta_f}^{\mathrm{out}}$.

Theorem 32 (On-average argument stability)**.** *Under the assumptions of Lemmas 30 and 31, assume that $\mathbb{P}\big((\widehat{\mathcal{E}}^{(i)})^c\big) \le \delta_{\mathrm{stab}}$ for all $i \in [n]$. Then*

$$\mathbb{E}_{S,\widetilde{S},\mathcal{A}}\Big[\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{W}_T - \mathbf{W}_T^{(i)}\|_2\Big] \le \frac{2e\eta b_{\delta,m}}{n}\sum_{t=0}^{T-1}\mathbb{E}_{S,\mathcal{A}}\big[\mathcal{L}_S(\mathbf{W}_t)\big] + 2R_*\delta_{\mathrm{stab}}.$$

Proof. Fix $i \in [n]$. On the event $\widehat{\mathcal{E}}^{(i)}$, Lemma 31 yields $\Delta_T^{(i)} \le e\sum_{t=0}^{T-1} R_t^{(i)}$. Since both trajectories are projected onto $\mathcal{K} = \mathcal{B}(\mathbf{W}_0, R_*)$, we also have $\Delta_T^{(i)} \le 2R_*$. Therefore,

$$\mathbb{E}_{S,\widetilde{S},\mathcal{A}}[\Delta_T^{(i)}] \le e\,\mathbb{E}_{S,\widetilde{S},\mathcal{A}}\Big[\sum_{t=0}^{T-1} R_t^{(i)}\mathbf{1}_{\widehat{\mathcal{E}}^{(i)}}\Big] + 2R_*\mathbb{P}\big((\widehat{\mathcal{E}}^{(i)})^c\big) \le e\sum_{t=0}^{T-1}\mathbb{E}_{S,\widetilde{S},\mathcal{A}}[R_t^{(i)}] + 2R_*\delta_{\mathrm{stab}}.$$

Averaging over $i \in [n]$ gives

$$\mathbb{E}_{S,\widetilde{S},\mathcal{A}}\Big[\frac{1}{n}\sum_{i=1}^{n}\Delta_T^{(i)}\Big] \le \frac{e}{n}\sum_{t=0}^{T-1}\sum_{i=1}^{n}\mathbb{E}_{S,\widetilde{S},\mathcal{A}}[R_t^{(i)}] + 2R_*\delta_{\mathrm{stab}}.$$

We now estimate the first term. Since $\mathbf{W}_t$ and $\mathbf{W}_t^{(i)}$ are determined before the fresh mini-batch $\mathcal{B}_t$ is drawn, they are independent of $\mathcal{B}_t$. Hence,

$$\mathbb{E}\Big[\frac{1}{B}\mathbf{1}_{\{i\in\mathcal{B}_t\}}\;\Big|\; S,\widetilde{S},\mathbf{W}_t,\mathbf{W}_t^{(i)}\Big] = \frac{1}{n}.$$

It follows from the definition of $R_t^{(i)}$ that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{S,\widetilde{S},\mathcal{A}}[R_t^{(i)}] = \frac{\eta b_{\delta,m}}{n}\mathbb{E}_{S,\widetilde{S},\mathcal{A}}\Big[\frac{1}{n}\sum_{i=1}^{n}\big(\ell(y_i f_{\mathbf{W}_t}(\mathbf{x}_i)) + \ell(y_i' f_{\mathbf{W}_t^{(i)}}(\mathbf{x}_i'))\big)\Big].$$

Using $\frac{1}{n}\sum_{i=1}^{n}\ell(y_i f_{\mathbf{W}_t}(\mathbf{x}_i)) = \mathcal{L}_S(\mathbf{W}_t)$ and the fact that $S^{(i)}$ has the same distribution as S, we obtain

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{S,\widetilde{S},\mathcal{A}}[R_t^{(i)}] = \frac{2\eta b_{\delta,m}}{n}\,\mathbb{E}_{S,\mathcal{A}}\big[\mathcal{L}_S(\mathbf{W}_t)\big].$$

Summing over $t = 0,\dots,T-1$ proves

$$\mathbb{E}_{S,\widetilde{S},\mathcal{A}}\Big[\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{W}_T - \mathbf{W}_T^{(i)}\|_2\Big] \le \frac{2e\eta b_{\delta,m}}{n}\sum_{t=0}^{T-1}\mathbb{E}_{S,\mathcal{A}}\big[\mathcal{L}_S(\mathbf{W}_t)\big] + 2R_*\delta_{\mathrm{stab}},$$

which completes the proof. □

Lemma 33. *Let $\mathcal{I}$ denote the initialization and let $\mathcal{R}$ denote all remaining randomness. If an event $\mathcal{E}$ satisfies $\mathbb{P}_{\mathcal{I},\mathcal{R}}(\mathcal{E}^c) \le \alpha\beta$, then, with probability at least $1-\alpha$ over $\mathcal{I}$,*

$$\mathbb{P}_{\mathcal{R}}(\mathcal{E}^c \mid \mathcal{I}) \le \beta.$$

Proof. The claim follows from Markov's inequality applied to $\mathbb{P}_{\mathcal{R}}(\mathcal{E}^c \mid \mathcal{I})$. □

Recall that $U_{\delta,\delta_f} \lesssim \sqrt{p\log(n/\delta)} + p^{3/2}R_*$.

Theorem 34 (Restatement of Theorem 2)**.** *Suppose Assumption 1 holds and the parameter conditions of Theorem 1 are satisfied. Assume further that, conditional on the initialization, an i.i.d. sample $S \sim \mathcal{P}^n$ satisfies Assumption 2 with margin γ with probability at least $1-\delta_{\mathrm{sep}}$, where $\delta_{\mathrm{sep}} \le \delta/8$. Let $\{\mathbf{W}_t\}_{t=0}^T$ be generated by Algorithm 1 with $C_{\mathrm{clip}} > 0$ and $\eta_{\mathrm{eff}} \le \frac{1}{12C_{\sigma,b}p^3}$. Assume*

$$m \gtrsim R_*^2\log(m/\delta)\eta^2\Big(TA_{\mathrm{corr}} + (\log(n/\delta)+R_*)\Big(\sqrt{\frac{T\log(1/\delta)}{B}} + \log\Big(\frac{1}{\delta}\Big)\Big)\Big)^2.$$

Then, with probability at least $1-\delta$ over the initialization,

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_{S,\mathcal{A}}[\mathcal{L}(\mathbf{W}_t)] \lesssim \Big(1 + \frac{\eta T\log(1/\delta)}{n}\Big)\Big(A_{\mathrm{corr}} + \log\Big(\frac{n}{\delta}\Big)(\sqrt{m}+R_*)\delta\Big).$$

Proof. Let $\mathcal{E}_{\mathrm{sep}}(S;\gamma)$ denote the event that Assumption 2 holds for S with margin γ. Choose auxiliary failure probabilities $\delta_f = \delta_{\mathrm{mb}} = \frac{\delta}{8}$ and $\bar{\delta}_{\mathrm{opt}} = \frac{\delta^2}{64}$. Applying Theorem 1 with failure probability $\bar{\delta}_{\mathrm{opt}}$ and Lemma 33 with $\alpha = \beta = \delta/8$, there exists an initialization event $\mathcal{E}_{\mathrm{cond}}$ with $\mathbb{P}(\mathcal{E}_{\mathrm{cond}}^c) \le \delta/8$ such that, on $\mathcal{E}_{\mathrm{cond}}$, the conditional probability that the optimization bound in Theorem 1 fails for an i.i.d. sample satisfying Assumption 2 is at most $\delta/8$. Here, replacing the failure probability by $\bar{\delta}_{\mathrm{opt}}$ only changes logarithmic factors, which are absorbed in A_{corr}. We condition on the initialization event $\mathcal{E}_{\delta/8} \cap \mathcal{E}_{\mathrm{cond}}$. Since $S^{(i)} \overset{d}{=} S$, the sample-separability assumption gives

$$\mathbb{P}\big(\mathcal{E}_{\mathrm{sep}}(S;\gamma)^c\big) \le \delta_{\mathrm{sep}}, \qquad \mathbb{P}\big(\mathcal{E}_{\mathrm{sep}}(S^{(i)};\gamma)^c\big) \le \delta_{\mathrm{sep}}.$$

Hence, by a union bound, both S and $S^{(i)}$ satisfy Assumption 2 with the same margin lower bound γ with probability at least $1 - 2\delta_{\mathrm{sep}}$, possibly with different separating directions. Since $S^{(i)} \overset{d}{=} S$, the conditional optimization guarantee above applies to both coupled runs, yielding

$$\mathbb{P}\big((\mathcal{E}_{\mathrm{opt}}^{(i)})^c\big) \le 2\delta_{\mathrm{sep}} + \frac{\delta}{4}.$$

Moreover, Lemma 29 gives $\mathbb{P}\big(\mathcal{E}_{\delta/8} \cap \mathcal{E}_{\delta_f}^{\mathrm{out}} \cap (\mathcal{E}_{\mathrm{mb}}^{(i)})^c\big) \le 2\delta_{\mathrm{mb}}$. Together with $\mathbb{P}((\mathcal{E}_{\delta_f}^{\mathrm{out}})^c) \le \delta_f$ from Lemma 27, for $\widehat{\mathcal{E}}^{(i)} = \mathcal{E}_{\delta/8} \cap \mathcal{E}_{\delta_f}^{\mathrm{out}} \cap \mathcal{E}_{\mathrm{opt}}^{(i)} \cap \mathcal{E}_{\mathrm{mb}}^{(i)}$, we obtain

$$\mathbb{P}\big((\widehat{\mathcal{E}}^{(i)})^c\big) \le \delta_f + 2\delta_{\mathrm{sep}} + \frac{\delta}{4} + 2\delta_{\mathrm{mb}} \le \delta,$$

where we used $\delta_{\mathrm{sep}} \le \delta/8$. Thus, Theorem 32 applies with $\delta_{\mathrm{stab}} = \delta$ and yields

$$\mathbb{E}_{S,\widetilde{S},\mathcal{A}}\Big[\frac{1}{n}\sum_{i=1}^{n} \|\mathbf{W}_T - \mathbf{W}_T^{(i)}\|_2\Big] \le \frac{2e\eta b_{\delta/8,m}}{n} \sum_{t=0}^{T-1} \mathbb{E}_{S,\mathcal{A}}\big[\mathcal{L}_S(\mathbf{W}_t)\big] + 2R_*\delta.$$

We now pass from stability to generalization. On $\mathcal{E}_{\delta/8}$, Lemma 10 gives $\|\nabla f_{\mathbf{W}}(\mathbf{x})\|_2 \le b_{\delta/8,m}$ for all $\mathbf{W} \in \mathcal{K}, \mathbf{x} \in \mathcal{X}$. Since $|\ell'(u)| \le 1$ for the logistic loss, the loss is $b_{\delta/8,m}$-Lipschitz with respect to $\mathbf{W}$ on this event. Applying Lemma 25, we obtain

$$\begin{aligned}\mathbb{E}_{S,\mathcal{A}}\big[\mathcal{L}(\mathbf{W}_T) - \mathcal{L}_S(\mathbf{W}_T)\big] &\le 2b_{\delta/8,m}\Big(\frac{2e\eta b_{\delta/8,m}}{n} \sum_{t=0}^{T-1} \mathbb{E}_{S,\mathcal{A}}\big[\mathcal{L}_S(\mathbf{W}_t)\big] + 2R_*\delta\Big) \\ &\lesssim \frac{\eta \log(1/\delta)}{n} \sum_{t=0}^{T-1} \mathbb{E}_{S,\mathcal{A}}\big[\mathcal{L}_S(\mathbf{W}_t)\big] + \sqrt{\log(1/\delta)}\, R_*\delta,\end{aligned}$$

where we used $b_{\delta/8,m} \lesssim \sqrt{\log(1/\delta)}$.

We now derive the averaged population risk bound. The same stability argument applies to the algorithm stopped at any time t. Therefore, for every $t = 0, \dots, T-1$,

$$\mathbb{E}_{S,\mathcal{A}}\big[\mathcal{L}(\mathbf{W}_t) - \mathcal{L}_S(\mathbf{W}_t)\big] \lesssim \frac{\eta \log(1/\delta)}{n} \sum_{s=0}^{T-1} \mathbb{E}_{S,\mathcal{A}}\big[\mathcal{L}_S(\mathbf{W}_s)\big] + \sqrt{\log(1/\delta)}\, R_*\delta.$$

Averaging this bound over $t = 0, \dots, T-1$ gives

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}_{S,\mathcal{A}}\big[\mathcal{L}(\mathbf{W}_t)\big] &\le \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}_{S,\mathcal{A}}\big[\mathcal{L}_S(\mathbf{W}_t)\big] + \frac{\eta \log(1/\delta)}{n} \sum_{t=0}^{T-1} \mathbb{E}_{S,\mathcal{A}}\big[\mathcal{L}_S(\mathbf{W}_t)\big] + \sqrt{\log(1/\delta)}\, R_*\delta \\ &= \Big(1 + \frac{\eta T \log(1/\delta)}{n}\Big)\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}_{S,\mathcal{A}}\big[\mathcal{L}_S(\mathbf{W}_t)\big] + \sqrt{\log(1/\delta)}\, R_*\delta.\end{aligned}$$

Let $\mathcal{E}_{\mathrm{opt}}$ be the event on which S satisfies Assumption 2 with margin γ and $\frac{1}{T}\sum_{t=0}^{T-1} \mathcal{L}_S(\mathbf{W}_t) \le A_{\mathrm{corr}}$. Then, on $\mathcal{E}_{\mathrm{cond}}$, $\mathbb{P}(\mathcal{E}_{\mathrm{opt}}^c) \le \delta_{\mathrm{sep}} + \delta/8$. On the event $\mathcal{E}_{\delta/8}$, for any $\mathbf{x} \in \mathcal{X}$, $y \in \{-1, +1\}$, and $\mathbf{W} \in \mathcal{B}(\mathbf{W}_0, R_*)$, it holds that

$$|f_{\mathbf{W}}(\mathbf{x})| \le |f_{\mathbf{W}_0}(\mathbf{x})| + b_{\delta/8,m}R_* \le B_b\sqrt{p}\,\|\mathbf{c}\|_2 + b_{\delta/8,m}R_* \lesssim \log\Big(\frac{n}{\delta}\Big)(\sqrt{m} + R_*),$$

where we used $\|\mathbf{c}\|_2 \lesssim \sqrt{m}$ on $\mathcal{E}_{\delta/8}$. Consequently,

$$\ell(yf_{\mathbf{W}}(\mathbf{x})) = \log\bigl(1 + e^{-yf_{\mathbf{W}}(\mathbf{x})}\bigr) \le \log 2 + |f_{\mathbf{W}}(\mathbf{x})| \lesssim \log\Bigl(\frac{n}{\delta}\Bigr)(\sqrt{m} + R_*).$$

Since $\mathbf{W}_t \in \mathcal{B}(\mathbf{W}_0, R_*)$ for all t by construction of Algorithm 1, the same bound holds for $\mathcal{L}_S(\mathbf{W}_t)$ uniformly over t. Therefore, conditioning on the initialization event $\mathcal{E}_{\delta/8} \cap \mathcal{E}_{\mathrm{cond}}$ and decomposing over $\mathcal{E}_{\mathrm{opt}}$, we obtain

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}_{S,\mathcal{A}}\bigl[\mathcal{L}_S(\mathbf{W}_t)\bigr] &\le \mathbb{E}_S\Bigl[\mathbb{E}_{\mathcal{A}}\Bigl[\frac{1}{T}\sum_{t=0}^{T-1} \mathcal{L}_S(\mathbf{W}_t)\mathbf{1}_{\mathcal{E}_{\mathrm{opt}}} \,\Big|\, S, \mathbf{W}_0, \mathbf{c}\Bigr]\Bigr] \\ &\quad + \mathbb{E}_S\Bigl[\mathbb{E}_{\mathcal{A}}\Bigl[\frac{1}{T}\sum_{t=0}^{T-1} \mathcal{L}_S(\mathbf{W}_t)\mathbf{1}_{\mathcal{E}_{\mathrm{opt}}^c} \,\Big|\, S, \mathbf{W}_0, \mathbf{c}\Bigr]\Bigr] \\ &\lesssim A_{\mathrm{corr}} + \log\Bigl(\frac{n}{\delta}\Bigr)(\sqrt{m} + R_*)\delta.\end{aligned}$$

Substituting this estimate into the averaged population risk bound gives

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}_{S,\mathcal{A}}[\mathcal{L}(\mathbf{W}_t)] \lesssim \Bigl(1 + \frac{\eta T \log(1/\delta)}{n}\Bigr)\Bigl(A_{\mathrm{corr}} + \log\Bigl(\frac{n}{\delta}\Bigr)(\sqrt{m} + R_*)\delta\Bigr).$$

Finally, $\mathbb{P}(\mathcal{E}_{\delta/8} \cap \mathcal{E}_{\mathrm{cond}}) \ge 1 - \delta/4 \ge 1 - \delta$, so the above bound holds with probability at least $1 - \delta$ over the initialization. This completes the proof. □

E Privacy Calibration under Temporal Correlation

E.1 Example: Finite-horizon improvement under matched privacy budgets

We justify the finite-horizon example stated in Section 5. Consider the full-batch setting $B = n$ with $C_{\mathrm{clip}} = G_\delta$, any fixed finite horizon $T \ge 2$, and compare correlated and independent noise using the same step size and target privacy budget (ϵ, δ). Since $C_{\mathrm{clip}} = G_\delta$, we have $\rho_{\mathrm{clip}} = 1$ and $\eta_{\mathrm{eff}} = \eta$.

For full-batch training, the differing record participates at every iteration. In the dominating-pair argument, retaining the exact finite-horizon weights gives the normalized squared sensitivity bound

$$\frac{\Delta_\lambda^2}{C_{\mathrm{clip}}^2} \le \sum_{t=1}^{T}\left(\frac{1-\lambda^t}{1-\lambda}\right)^2.$$

Therefore, up to universal constants, a sufficient Gaussian calibration is

$$\kappa_\lambda^2 \asymp \frac{\log(1/\delta)}{\epsilon^2}\sum_{t=1}^{T}\left(\frac{1-\lambda^t}{1-\lambda}\right)^2.$$

When $\lambda = 0$, this reduces to the independent noise calibration

$$\kappa_0^2 \asymp \frac{T\log(1/\delta)}{\epsilon^2},$$

and hence, equivalently,

$$\kappa_\lambda^2 = \kappa_0^2 \frac{1}{T}\sum_{t=1}^{T}\left(\frac{1-\lambda^t}{1-\lambda}\right)^2$$

up to constant factors.

We now compare the privacy-dependent terms in the two population risk bounds. Relative to independent noise, the linear and quadratic privacy terms in A_{corr} are multiplied, respectively, by

$$r_1(\lambda) = (1-\lambda)\left(\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1-\lambda^t}{1-\lambda}\right)^2\right)^{1/2} = \left(\frac{1}{T}\sum_{t=1}^{T}(1-\lambda^t)^2\right)^{1/2},$$

and

$$r_2(\lambda) = \big((1-\lambda)^2 + \eta\lambda^2\big)\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1-\lambda^t}{1-\lambda}\right)^2.$$

For every $\lambda \in (0,1)$, $r_1(\lambda) < 1$. Moreover, $r_1(\lambda)$ is strictly decreasing in λ and converges to 0 as $\lambda \to 1$. For the quadratic term, as $\lambda \to 0^+$,

$$r_2(\lambda) = 1 - \frac{2\lambda}{T} + \left(\eta - \frac{1}{T}\right)\lambda^2 + \mathcal{O}(\lambda^3).$$

Hence, for every fixed finite $T \geq 2$, $r_2(\lambda) < 1$ for all sufficiently small positive λ. Therefore, there exists $\lambda_T \in (0,1)$ such that

$$r_1(\lambda_T) < 1, \qquad r_2(\lambda_T) < 1.$$

Consequently, under the common admissibility conditions of Theorem 2, the population risk bound for correlated noise satisfies

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_{S,\mathcal{A}}\big[\mathcal{L}(\mathbf{W}_t)\big] \lesssim \left(1+\frac{\eta T}{n}\right)\left(A_{\mathrm{com}} + \frac{r_1(\lambda_T)}{n\epsilon}\left(\frac{1}{\gamma} + \eta\sqrt{\frac{md}{n}}\right) + \frac{r_2(\lambda_T)\eta T m d}{n^2\epsilon^2}\right),$$

where A_{com} collects the terms common to the correlated and independent noise mechanisms. The corresponding independent noise bound has the same form with $r_1 = r_2 = 1$. Thus, the linear privacy-dependent term is strictly reduced for every positive correlation level, while for every fixed finite horizon one can choose a positive correlation level that also reduces the quadratic term. Consequently, temporal correlation yields a strictly smaller population risk upper bound under the same privacy budget in this setting.

E.2 Example: Representative rate under conservative worst-case privacy calibration

We consider a second representative setting based on a conservative worst-case privacy calibration for mini-batch training. To obtain an explicit (ϵ,δ)-DP population risk rate, we calibrate the noise multiplier κ using the following privacy guarantee for Algorithm 1. This calibration is sufficient to establish a vanishing population risk rate, although sharper privacy accounting may require a smaller noise multiplier.

Theorem 35 (Privacy guarantee)**.** *Let $\epsilon, \delta \in (0,1]$, let $\lambda \in (0,1)$, and define $r = B/n$. Assume $rT \geq 3\log(2/\delta)$. If*

$$\kappa^2 \geq 8\Big(\frac{1-\lambda^T}{1-\lambda}\Big)^2\Big(rT + \sqrt{3rT\log\big(\frac{2}{\delta}\big)}\Big)\frac{\log\big(\frac{2.5}{\delta}\big)}{\epsilon^2},$$

then Algorithm 1 satisfies (ϵ,δ)-DP. In particular, up to universal constants, it suffices to take

$$\kappa^2 \asymp \Big(\frac{1-\lambda^T}{1-\lambda}\Big)^2\Big(\frac{B}{n}T + \sqrt{\frac{B}{n}T\log\big(\frac{1}{\delta}\big)}\Big)\log\big(\frac{1}{\delta}\big)\epsilon^{-2}.$$

Combining Theorems 2 and 35, consider a representative regime with $\eta \asymp 1$, $C_{\mathrm{clip}} \asymp G_\delta$, and $\lambda \in (0,1)$ fixed and bounded away from 1. Let $0 < \rho < 1$ be fixed, take $m \asymp \gamma^{-6}\mathrm{polylog}(n/\delta)$, $\delta \leq \min\big\{\frac{1}{n\sqrt{m}}, \frac{\gamma}{n}\big\}$, and set $B = \rho n$ and $T \asymp \min\big\{\frac{\sqrt{n}}{\gamma}, \frac{n\epsilon\gamma^2}{\sqrt{d}}\big\}$. Then both the averaged empirical and population risks are of order $\mathcal{O}\big(\frac{1}{\gamma\sqrt{n}} + \frac{\sqrt{d}}{\gamma^4 n\epsilon}\big)$ up to logarithmic factors. Under this conservative worst-case privacy calibration, the reduction in the effective perturbation due to temporal correlation is offset by the corresponding increase in the required noise multiplier, so the resulting rate has the same leading dependence on n, d, and ϵ as independent noise DP-SGD. The particular choices of B, T, and C_{clip} are made to obtain a transparent rate and are not claimed to be optimal. This should be viewed as a worst-case guarantee rather than as evidence that temporal correlation cannot improve the privacy–utility tradeoff. Sharper privacy accounting or more favorable temporal structure may preserve part of the optimization gain. A formal statement is given below.

Corollary 36 (Representative worst-case rate)**.** *Suppose Assumptions 1 and 2 hold. Let $\epsilon, \delta \in (0, 1]$ and let $\{\mathbf{W}_t\}_{t=0}^{T}$ be generated by Algorithm 1 with $\eta \asymp 1$ and $C_{\text{clip}} \asymp G_\delta$. Assume $\lambda \in (0, 1)$ is a fixed constant bounded away from 1, and let $0 < \rho < 1$ be a fixed constant. Suppose $m \asymp \gamma^{-6}\text{polylog}(n/\delta)$ and $\delta \le \min\{\frac{1}{n\sqrt{m}}, \frac{\gamma}{n}\}$. Set $B = \rho n$, $T \asymp \min\{\frac{\sqrt{n}}{\gamma}, \frac{n\epsilon\gamma^2}{\sqrt{d}}\}$, assume $\rho T \ge 3\log(2/\delta)$, and choose κ according to Theorem 35. Then, with probability at least $1 - \delta$ over the initialization and the algorithmic randomness,*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathcal{L}_S(\mathbf{W}_t) \lesssim \text{polylog}\left(\frac{n}{\delta}\right)\left(\frac{1}{\gamma\sqrt{n}} + \frac{\sqrt{d}}{\gamma^4 n\epsilon}\right).$$

For the population risk statement, assume further the sample-separability condition of Theorem 2. With probability at least $1 - \delta$ over the initialization,

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}_{S,\mathcal{A}}\big[\mathcal{L}(\mathbf{W}_t)\big] \lesssim \text{polylog}\left(\frac{n}{\delta}\right)\left(\frac{1}{\gamma\sqrt{n}} + \frac{\sqrt{d}}{\gamma^4 n\epsilon}\right).$$

Proof of Theorem 35. We proceed in two steps. First, we conduct a general privacy analysis of Algorithm 1 in terms of *dominating pairs*. We then use this analysis to derive an analytical upper bound on the noise multiplier κ.

E.2.1 Step 1: Dominating pairs for arbitrary correlation matrices

It suffices to analyze the sequence of noisy clipped-gradient releases, since the iterates generated by Algorithm 1 are obtained from these releases by post-processing. We normalize the per-example clipped gradients by C_{clip} and multiply each mini-batch average by B/C_{clip}. Under this normalization, each per-example contribution has Euclidean norm at most one and the Gaussian noise has standard deviation κ.

The normalized mechanism can be written in the matrix form

$$\mathbf{X} + \mathbf{C}^{-1}\mathbf{Z}, \qquad \mathbf{X} = \sum_{i=1}^{n} \mathbf{G}^{(i)},$$

where $\mathbf{G}^{(i)} \in \mathbb{R}^{T\times mdp}$ denotes the normalized contribution of the i-th record across the training iterations, and

$$Z_{t,j} \sim \mathcal{N}(0, \kappa^2).$$

The t-th row of $\mathbf{G}^{(i)}$ satisfies

$$\|\mathbf{G}^{(i)}_{t,:}\|_2 \le 1$$

if record i is included in the t-th mini-batch, and $\mathbf{G}^{(i)}_{t,:} = 0$ otherwise. Note that $\mathbf{G}^{(i)}_{t,:}$ may be chosen adaptively based on earlier outcomes.

The matrix $\mathbf{C}^{-1} \in \mathbb{R}^{T\times T}$ is given by

$$C^{-1}_{t,u} = \begin{cases} 1, & u = t, \\ -\lambda, & u = t-1, \\ 0, & \text{otherwise}, \end{cases}$$

and its inverse is

$$C_{t,u} = \begin{cases} 1, & u = t, \\ \lambda^{t-u}, & u < t, \\ 0, & \text{otherwise}. \end{cases} \tag{19}$$

Analogously, for a neighboring dataset S', the mechanism can be represented as

$$\mathbf{X}' + \mathbf{C}^{-1}\mathbf{Z},$$

where $\mathbf{X}'$ differs from $\mathbf{X}$ only through the contribution of a single record.

To proceed with the privacy analysis, we formalize the phrase "differ in the contribution of one data record" used in Definition 1. DP-SGD with Poisson subsampling, where each record is independently

included with a fixed rate $r \in [0, 1]$, is typically analyzed under the insertion/removal relation. In contrast, Algorithm 1 uses mini-batches of fixed size B, sampled uniformly without replacement. For this fixed-size subsampling scheme, we use the standard zero-out relation; see Annamalai et al. (2025) for an overview of papers using this relation.

Definition 3 (Zero-out relation)**.** Datasets S and S' of equal size n are neighboring under the zero-out relation if there exists a pair of records (x_i, y_i) and (x'_i, y'_i) such that

$$S' = S \setminus \{(x_i, y_i)\} \cup \{(x'_i, y'_i)\},$$

and the normalized gradient contribution of one of the two records is zero, i.e.,

$$\mathbf{G}^{(i)} = 0 \qquad \text{or} \qquad \mathbf{G}'^{(i)} = 0.$$

To simplify the analysis, we further subdivide this relation into two asymmetric parts.

Definition 4 (Zero-out-remove relation)**.** Datasets S and S' of equal size n are neighboring under the zero-out-remove relation if there exists a pair of records (x_i, y_i) and (x'_i, y'_i) such that

$$S' = S \setminus \{(x_i, y_i)\} \cup \{(x'_i, y'_i)\},$$

and the normalized gradient contribution of the substituted record in S' is zero, i.e.,

$$\mathbf{G}'^{(i)} = 0.$$

Definition 5 (Zero-out-add relation)**.** Datasets S and S' of equal size n are neighboring under the zero-out-add relation if there exists a pair of records (x_i, y_i) and (x'_i, y'_i) such that

$$S' = S \setminus \{(x_i, y_i)\} \cup \{(x'_i, y'_i)\},$$

and the normalized gradient contribution of the substituted record in S is zero, i.e.,

$$\mathbf{G}^{(i)} = 0.$$

In the following, we use $S \simeq_0 S'$, $S \simeq_{0,r} S'$, and $S \simeq_{0,a} S'$ for datasets neighboring under the zero-out, zero-out-remove, and zero-out-add relations, respectively.

Finally, we recall the hockey-stick divergence and dominating pairs, which provide an alternative characterization of the privacy profile $\delta(\epsilon)$ Zhu et al. (2022).

Definition 6 (Hockey-stick divergence)**.** The hockey-stick divergence of order $\alpha \geq 0$ between two distributions P, Q on a measure space $(\Omega, \mathcal{F})$ is

$$H_\alpha(P\|Q) = \sup_{E \in \mathcal{F}} \big(P(E) - \alpha Q(E)\big).$$

Definition 7 (Dominating pair)**.** Consider a randomized algorithm $\mathcal{A}$ and two distributions P, Q. If

$$H_\alpha\big(\mathcal{A}(\cdot \mid S)\|\mathcal{A}(\cdot \mid S')\big) \leq H_\alpha(P\|Q)$$

for all neighboring datasets $S \simeq S'$ and all $\alpha \geq 0$, then (P, Q) is a dominating pair for $\mathcal{A}$.

If (P, Q) is a dominating pair for $\mathcal{A}$, then the corresponding hockey-stick divergence controls the privacy profile Barthe & Olmedo (2013). In particular, if the relevant divergences in both neighboring directions are bounded by δ at $\alpha = e^\epsilon$, then $\mathcal{A}$ satisfies (ϵ, δ)-DP.

We first recall the following symmetry relation. The proof is identical to that of Lemma 28 in Zhu et al. (2022).

Lemma 37 (Zhu et al. (2022))**.** *The following two statements are equivalent:*

1. (P, Q) *dominates algorithm* $\mathcal{A}$ *for zero-out-add neighbors.*
2. (Q, P) *dominates algorithm* $\mathcal{A}$ *for zero-out-remove neighbors.*

Proof. Let $T[P, Q] : [0, 1] \to [0, 1]$ be the trade-off function of distributions P and Q. We use that $T[P, Q]$ and $T[Q, P]$ are inverse functions and that

$$H_\alpha(P\|Q) = 1 + T[P, Q]^*(-\alpha),$$

where f^* denotes the convex conjugate of f. Consider two datasets $S \simeq_{0,a} S'$. Then

$$\begin{aligned}(P, Q) \text{ dominates } \mathcal{A}(S), \mathcal{A}(S') &\iff H_\alpha(\mathcal{A}(S)\|\mathcal{A}(S')) \le H_\alpha(P\|Q) \\ &\iff T[\mathcal{A}(S'), \mathcal{A}(S)] \ge T[Q, P] \\ &\iff T[\mathcal{A}(S), \mathcal{A}(S')] \ge T[P, Q] \\ &\iff H_\alpha(\mathcal{A}(S')\|\mathcal{A}(S)) \le H_\alpha(Q\|P),\end{aligned}$$

which proves the claim. □

We further use Lemma 4.5 of Choquette-Choo et al. (2024c), restated as Lemma 3.3 in Choquette-Choo et al. (2024b).

Lemma 38 (Choquette-Choo et al. (2024b))**.** *Let $\mathbf{c}_1, \dots, \mathbf{c}_k \in \mathbb{R}^{n\times p}$ and let $\mathbf{c}'_1, \dots, \mathbf{c}'_k \in \mathbb{R}^n$ satisfy*

$$\|\mathbf{c}_i[j, :]\|_2 \le \mathbf{c}'_i(j) \qquad \text{for all } i, j.$$

Let

$$P = \mathcal{N}(0, \sigma^2 I), \qquad Q = \sum_i p_i \mathcal{N}(\mathbf{c}_i, \sigma^2 I),$$

and

$$P' = \mathcal{N}(0, \sigma^2 I), \qquad Q' = \sum_i p_i \mathcal{N}(\mathbf{c}'_i, \sigma^2 I).$$

Then, for every $\alpha \ge 0$,

$$H_\alpha(P\|Q) \le H_\alpha(P'\|Q').$$

Furthermore, the conclusion continues to hold when the j-th row of each $\mathbf{c}_i$ is chosen as a function of the first $j-1$ rows, subject to the same row-norm bound, while the vectors $\mathbf{c}'_i$ remain fixed.

Using these results, we now derive a dominating pair for the normalized mechanism. The argument follows the proof strategy of Lemma 3.3 in Choquette-Choo et al. (2024b).

Theorem 39. *Given the number of iterations T, batch size B, and dataset size n, define the subsampling rate*

$$r = \frac{B}{n}.$$

Let

$$s(\mathbf{y}) = r^{\|\mathbf{y}\|_0}(1-r)^{T-\|\mathbf{y}\|_0}, \qquad \mathbf{y} \in \{0,1\}^T.$$

Define

$$P = \sum_{\mathbf{y}\in\{0,1\}^T} s(\mathbf{y})\mathcal{N}(\mathbf{C}\mathbf{y}, \kappa^2 I_T), \qquad Q = \mathcal{N}(0, \kappa^2 I_T).$$

Then the normalized mechanism associated with Algorithm 1 is dominated by (P, Q) under the zero-out-remove relation and by (Q, P) under the zero-out-add relation.

Proof. Consider datasets $S \simeq_{0,r} S'$ and assume without loss of generality that the n-th record is zeroed out in S'. Then

$$\mathcal{A}(S) = \mathbf{X} + \mathbf{C}^{-1}\mathbf{Z}, \qquad \mathbf{X} = \sum_{i=1}^{n} \mathbf{G}^{(i)},$$

whereas

$$\mathcal{A}(S') = \mathbf{X}' + \mathbf{C}^{-1}\mathbf{Z}, \qquad \mathbf{X}' = \sum_{i=1}^{n-1} \mathbf{G}^{(i)}.$$

Applying the invertible linear transformation $\mathbf{C}$ gives

$$\widetilde{\mathcal{A}}(S) = \mathbf{C}\mathbf{X} + \mathbf{Z}, \qquad \widetilde{\mathcal{A}}(S') = \mathbf{C}\mathbf{X}' + \mathbf{Z}.$$

Since the transformation is invertible, the original and transformed mechanisms have the same privacy profile.

Following the standard reduction used in Choquette-Choo et al. (2024b), the contributions shared by the two neighboring datasets can be removed, so it suffices to distinguish

$$\mathbf{C}\mathbf{G}^{(n)} + \mathbf{Z} \qquad \text{from} \qquad \mathbf{Z}.$$

Given a participation vector $\mathbf{y} \in \{0,1\}^T$, with $y_u = 1$ if the differing record participates at iteration u, the t-th row satisfies

$$\|(\mathbf{C}\mathbf{G}^{(n)})_{t,:}\|_2 \le \sum_{u=1}^{T} C_{t,u} y_u = (\mathbf{C}\mathbf{y})_t,$$

where we used $\|\mathbf{G}^{(n)}_{u,:}\|_2 \le y_u$ and $C_{t,u} \ge 0$. Lemma 38 therefore reduces the adaptive matrix-valued Gaussian mixture to the vector-valued Gaussian mixture stated in the theorem. The result for the reverse neighboring relation follows from the preceding symmetry lemma. □

E.2.2 Step 2: Analytical noise multiplier bound

We now use the dominating pair from Theorem 39 to derive an explicit sufficient noise multiplier κ for a desired privacy budget (ϵ, δ).

The main idea is to derive a high-probability bound on the number of iterations in which the differing record participates. This gives a corresponding high-probability bound on the Gaussian mean $\|\mathbf{C}\mathbf{y}\|_2$. We calibrate the Gaussian noise to this sensitivity while allocating part of the privacy failure probability to the complement of the high-probability event.

Lemma 40. *Consider an event $E \subseteq \{0,1\}^T$ with probability $S(E)$ under the participation distribution $s(\mathbf{y})$. Let P, Q be defined as in Theorem 39. Then, for all $\epsilon \ge 0$,*

$$\max\{H_{e^\epsilon}(P\|Q), H_{e^\epsilon}(Q\|P)\} \le S(E) \max_{\mathbf{y}\in E} H_{e^\epsilon}\left(\mathcal{N}(\mathbf{C}\mathbf{y}, \kappa^2 I_T)\|\mathcal{N}(0, \kappa^2 I_T)\right) + S(\overline{E}).$$

Proof. For the (P, Q) direction, by the law of total probability,

$$P = S(E)P(\cdot \mid E) + S(\overline{E})P(\cdot \mid \overline{E}).$$

Joint convexity of the hockey-stick divergence gives

$$\begin{aligned} H_{e^\epsilon}(P\|Q) &\le S(E) H_{e^\epsilon}(P(\cdot \mid E)\|Q) + S(\overline{E}) H_{e^\epsilon}(P(\cdot \mid \overline{E})\|Q) \\ &\le S(E) \max_{\mathbf{y}\in E} H_{e^\epsilon}\left(\mathcal{N}(\mathbf{C}\mathbf{y}, \kappa^2 I_T)\|\mathcal{N}(0, \kappa^2 I_T)\right) + S(\overline{E}), \end{aligned}$$

where we used that the hockey-stick divergence is at most one and that joint convexity implies quasi-convexity over the mixture on E.

For the reverse direction, write

$$Q = \sum_{\mathbf{y}} s(\mathbf{y}) Q \qquad \text{and} \qquad P = \sum_{\mathbf{y}} s(\mathbf{y}) P_{\mathbf{y}},$$

where

$$P_{\mathbf{y}} = \mathcal{N}(\mathbf{C}\mathbf{y}, \kappa^2 I_T).$$

Joint convexity gives the same bound with $H_{e^\epsilon}(Q\|P_{\mathbf{y}})$. For Gaussians with a common covariance, the two directions have the same hockey-stick divergence by translation and reflection symmetry. Hence the stated bound follows. □

Next, we control the number of participations of the differing record.

Lemma 41. *Given T training steps, dataset size n, and batch size B, define*

$$r = \frac{B}{n}.$$

Choose any constant $c_1 \in (0,1)$ satisfying

$$rT \ge 3 \log\left(\frac{1}{c_1 \delta}\right).$$

Let $\mathbf{y}$ *denote the participation vector of a fixed record over the* T *independently sampled mini-batches. Define*

$$\tau = rT + \sqrt{3rT \log\left(\frac{1}{c_1\delta}\right)}$$

and

$$E = \left\{\mathbf{y} \in \{0,1\}^T : \|\mathbf{y}\|_0 \le \tau\right\}.$$

Then

$$S(E) \ge 1 - c_1\delta, \qquad S(\overline{E}) \le c_1\delta.$$

Proof. For a fixed record, each iteration includes that record with probability $r = B/n$. Since the mini-batches are sampled independently across iterations, the indicators y_t are independent Bernoulli(r) random variables. Hence

$$\|\mathbf{y}\|_0 \sim \text{Binomial}(T, r).$$

The claim follows directly from the multiplicative Chernoff bound. □

We next bound the encoded sensitivity.

Lemma 42. *Let* $\mathbf{C}$ *be the* $T \times T$ *lower-triangular matrix with*

$$C_{i,j} = \lambda^{i-j}, \qquad i \ge j.$$

For any $\mathbf{y} \in \{0,1\}^T$ *satisfying* $\|\mathbf{y}\|_0 \le \tau$*,*

$$\|\mathbf{C}\mathbf{y}\|_2 \le \left(\frac{1-\lambda^T}{1-\lambda}\right)\sqrt{\tau}.$$

Proof. By consistency of the induced matrix 2-norm,

$$\|\mathbf{C}\mathbf{y}\|_2 \le \|\mathbf{C}\|_2 \|\mathbf{y}\|_2.$$

Since $\mathbf{y}$ is binary,

$$\|\mathbf{y}\|_2 = \sqrt{\|\mathbf{y}\|_0} \le \sqrt{\tau}.$$

Moreover,

$$\|\mathbf{C}\|_2 \le \sqrt{\|\mathbf{C}\|_1 \|\mathbf{C}\|_\infty} \le \frac{1-\lambda^T}{1-\lambda}.$$

Combining the two inequalities yields the result. □

We can now complete the proof of Theorem 35. Set $c_1 = 1/2$ and define

$$\tau = rT + \sqrt{3rT \log\left(\frac{2}{\delta}\right)}.$$

By Lemma 41,

$$S(\overline{E}) \le \delta/2.$$

On the event E, Lemma 42 gives

$$\Delta_2 := \max_{\mathbf{y} \in E} \|\mathbf{C}\mathbf{y}\|_2 \le \left(\frac{1-\lambda^T}{1-\lambda}\right)\sqrt{\tau}.$$

The Gaussian mechanism on E is $(\epsilon, \delta/2)$-DP whenever

$$\kappa^2 \ge \Delta_2^2 \frac{2\log(2.5/\delta)}{\epsilon^2}.$$

Therefore, the stronger sufficient condition

$$\kappa^2 \ge 8\left(\frac{1-\lambda^T}{1-\lambda}\right)^2 \left(rT + \sqrt{3rT \log\left(\frac{2}{\delta}\right)}\right) \frac{\log\left(\frac{2.5}{\delta}\right)}{\epsilon^2}$$

implies that

$$\max_{\mathbf{y}\in E} H_{e^\epsilon}\left(\mathcal{N}(\mathbf{C}\mathbf{y},\kappa^2 I_T)\|\mathcal{N}(0,\kappa^2 I_T)\right)\le \delta/2.$$

Lemma 40 then yields

$$\max\left\{H_{e^\epsilon}(P\|Q), H_{e^\epsilon}(Q\|P)\right\}\le S(E)\frac{\delta}{2}+S(\overline{E})\le\delta.$$

Hence Algorithm 1 satisfies (ϵ,δ)-DP, completing the proof of Theorem 35.

Proof of Corollary 36. Since $C_{\text{clip}}\asymp G_\delta$, we have $\rho_{\text{clip}}=\Theta(1)$, $\eta_{\text{eff}}\asymp\eta$, and $c_{\text{eff}}\asymp c_{\text{priv}}$. Hence the additional factors arising from active clipping are absorbed into absolute constants throughout the proof.

Since $\lambda\in(0,1)$ is fixed and bounded away from 1, we have $\frac{1-\lambda^T}{1-\lambda}=\mathcal{O}(1)$ and $(1-\lambda)^2+\lambda^2\eta=\mathcal{O}(1)$. Moreover, $B=\rho n$, so $r=\rho$. Since $rT\ge 3\log(2/\delta)$, the square-root term in Theorem 35 is no larger than a constant multiple of rT. Therefore, choosing κ according to Theorem 35 gives

$$\kappa^2\lesssim\frac{T\log(1/\delta)}{\epsilon^2}.$$

In the remainder of the proof, logarithmic factors are absorbed into $\text{polylog}(n/\delta)$.

We first verify the conditions of Theorem 1. The lower bound on m is satisfied because $m\asymp\gamma^{-6}\text{polylog}(n/\delta)$ is larger than γ^{-4} up to logarithmic factors. For the upper bound on m, using $B=\rho n$, the bound on κ, and $(1-\lambda)^2+\lambda^2\eta=\mathcal{O}(1)$, the right-hand side of the width upper bound is, up to logarithmic factors,

$$\frac{B^2}{\eta^2\kappa^2\gamma^2 d}\cdot\frac{1}{T}\gtrsim\frac{n^2\epsilon^2}{\eta^2\gamma^2 dT^2}.$$

Since $T\lesssim n\epsilon\gamma^2/\sqrt{d}$ and $m\asymp\gamma^{-6}\text{polylog}(n/\delta)$, the width upper bound is satisfied after adjusting constants and logarithmic factors.

We next verify the step-size condition in Theorem 1. Since $\rho_{\text{clip}}=\Theta(1)$ and $\eta\asymp 1$ is chosen as a sufficiently small absolute constant, we have $\eta_{\text{eff}}\asymp\eta$. Moreover, since $B=\rho n$ and $T\lesssim\sqrt{n}/\gamma$, we have

$$\eta\gamma\sqrt{T}\frac{1}{\sqrt{B}}\lesssim 1,\qquad \eta^2\gamma^2\left(\frac{T}{B}+1\right)\lesssim 1.$$

Using $\kappa\lesssim\sqrt{T}\,\text{polylog}^{1/2}(n/\delta)/\epsilon$, we further obtain, up to logarithmic factors,

$$\eta\gamma\sqrt{T}\,\frac{\kappa(1-\lambda)}{B}\lesssim\eta\gamma\frac{T}{n\epsilon}\lesssim 1,$$

where the last inequality follows from $T\lesssim n\epsilon\gamma^2/\sqrt{d}$, together with $\gamma\le 1$ and $d\ge 1$. Hence the step-size condition holds.

We now bound A_{corr}. Since $\rho_{\text{clip}}=\Theta(1)$, $\eta_{\text{eff}}\asymp\eta$, and $c_{\text{eff}}\asymp c_{\text{priv}}$, the active-clipping version of Theorem 1 yields, up to logarithmic factors,

$$A_{\text{corr}}\lesssim\frac{1}{\gamma^2\eta T}+\frac{1}{\gamma\sqrt{nT}}+\eta\left(\frac{1}{n}+\frac{1}{T}\right)+\frac{1}{n\epsilon}\left(\frac{1}{\gamma}+\frac{\eta\sqrt{md}}{\sqrt{n}}\right)+\frac{\eta Tmd}{n^2\epsilon^2}.$$

By the choice

$$T\asymp\min\left\{\frac{\sqrt{n}}{\gamma},\,\frac{n\epsilon\gamma^2}{\sqrt{d}}\right\},$$

the first term satisfies

$$\frac{1}{\gamma^2\eta T}\lesssim\frac{1}{\gamma\sqrt{n}}+\frac{\sqrt{d}}{\gamma^4 n\epsilon}.$$

The second and third terms are of no larger order. For the linear-noise term,

$$\frac{1}{n\epsilon}\left(\frac{1}{\gamma}+\frac{\eta\sqrt{md}}{\sqrt{n}}\right)\lesssim\text{polylog}\left(\frac{n}{\delta}\right)\frac{\sqrt{d}}{\gamma^4 n\epsilon},$$

where we used $d \geq 1$, $\gamma \leq 1$, $\eta \asymp 1$, and $m \asymp \gamma^{-6}\mathrm{polylog}(n/\delta)$. For the quadratic-noise term, using $T \lesssim n\epsilon\gamma^2/\sqrt{d}$, we obtain

$$\frac{\eta T m d}{n^2\epsilon^2} \lesssim \frac{\eta m \gamma^2 \sqrt{d}}{n\epsilon} \lesssim \mathrm{polylog}\left(\frac{n}{\delta}\right)\frac{\sqrt{d}}{\gamma^4 n\epsilon}.$$

Hence

$$A_{\mathrm{corr}} \lesssim \mathrm{polylog}\left(\frac{n}{\delta}\right)\left(\frac{1}{\gamma\sqrt{n}} + \frac{\sqrt{d}}{\gamma^4 n\epsilon}\right).$$

This proves the optimization bound.

It remains to verify the additional condition in Theorem 2. Since $R_* \asymp \gamma^{-1}\mathrm{polylog}(n/\delta)$, the bound above and the choice of T give $TA_{\mathrm{corr}} \lesssim \gamma^{-2}\mathrm{polylog}(n/\delta)$. The additional mini-batch term in the width condition of Theorem 2 is no larger than TA_{corr} under the present choices of B and T. Therefore, the right-hand side of the required width condition is bounded by

$$R_*^2\eta^2(TA_{\mathrm{corr}})^2 \lesssim \gamma^{-6}\mathrm{polylog}(n/\delta).$$

Hence the condition is satisfied by $m \asymp \gamma^{-6}\mathrm{polylog}(n/\delta)$.

Applying Theorem 2, we obtain

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_{S,\mathcal{A}}[\mathcal{L}(\mathbf{W}_t)] \lesssim \left(1 + \frac{\eta T\log(1/\delta)}{n}\right)\left(A_{\mathrm{corr}} + \log\left(\frac{n}{\delta}\right)(\sqrt{m} + R_*)\delta\right).$$

Since $T \lesssim \sqrt{n}/\gamma$, the factor $1 + \eta T\log(1/\delta)/n$ is absorbed into the logarithmic factor. Moreover, $\delta \leq \min\{1/(n\sqrt{m}),\, \gamma/n\}$ implies

$$\log\left(\frac{n}{\delta}\right)(\sqrt{m} + R_*)\delta \lesssim \mathrm{polylog}\left(\frac{n}{\delta}\right)\frac{1}{n},$$

which is dominated by the displayed rate. Therefore,

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_{S,\mathcal{A}}\big[\mathcal{L}(\mathbf{W}_t)\big] \lesssim \mathrm{polylog}\left(\frac{n}{\delta}\right)\left(\frac{1}{\gamma\sqrt{n}} + \frac{\sqrt{d}}{\gamma^4 n\epsilon}\right).$$

This completes the proof. □

F Proofs for DP-SGD with Independent Noise

F.1 Privacy guarantee of DP-SGD with independent noise

In this subsection, we prove the privacy guarantee of Algorithm 1 in the case $\lambda = 0$, i.e., standard DP-SGD with independent Gaussian noise. The proof is standard. Nevertheless, for consistency, we include the details.

Throughout the proof, we use Rényi differential privacy (RDP) that allows for a more refined analysis of privacy loss. In this subsection, for the independent-noise baseline, we use the standard replacement neighboring relation: two datasets S and S' of the same size are neighboring if they differ in one data point.

Definition 8 (Rényi differential privacy Mironov (2017))**.** For $\alpha > 1$ and $\rho > 0$, a randomized algorithm $\mathcal{A}$ is said to satisfy (α, ρ)-RDP if, for every pair of neighboring datasets S, S',

$$D_\alpha(\mathcal{A}(S)\,\|\,\mathcal{A}(S')) := \frac{1}{\alpha - 1}\log\mathbb{E}_{\theta\sim\mathcal{A}(S')}\left[\left(\frac{\mathcal{A}(S)(\theta)}{\mathcal{A}(S')(\theta)}\right)^\alpha\right] \leq \rho.$$

A connection (ϵ, δ)-DP and RDP is established in the following lemma.

Lemma 43 (RDP to (ϵ, δ)-DP Mironov (2017))**.** *If $\mathcal{A}$ satisfies (α, ρ)-RDP for some $\alpha > 1$, then for every $\delta \in (0, 1)$, $\mathcal{A}$ also satisfies*

$$\left(\rho + \frac{\log(1/\delta)}{\alpha - 1},\, \delta\right)\textit{-DP}.$$

To achieve DP, we need the concept of ℓ_2-sensitivity defined as follows.

Definition 9 (ℓ_2-sensitivity)**.** The ℓ_2-sensitivity of a function (mechanism) $\mathcal{M} : \mathcal{Z}^n \to \mathcal{W}$ is defined as $\Delta = \sup_{S,S'} \|\mathcal{M}(S) - \mathcal{M}(S')\|_2$, where S and S' are neighboring datasets.

A basic mechanism to obtain RDP is Gaussian mechanism.

Lemma 44 (Gaussian mechanism Mironov (2017))**.** *Consider a function $\mathcal{M} : \mathcal{Z}^n \to \mathcal{R}^d$ with the ℓ_2-sensitivity parameter Δ, and a dataset $S \subset \mathcal{Z}^n$. The Gaussian mechanism $\mathcal{G}(S, \sigma) = \mathcal{M}(S) + \mathbf{b}$, where $\mathbf{b} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d)$, satisfies $(\lambda, \frac{\lambda \Delta^2}{2\sigma^2})$-RDP.*

The following post-processing property enables flexible use of private data outputs while preserving rigorous privacy guarantees.

Lemma 45 (Post-processing Mironov (2017))**.** *If $\mathcal{A}$ satisfies (α, ρ)-RDP and f is any deterministic function, then $f \circ \mathcal{A}$ also satisfies (α, ρ)-RDP.*

The following RDP composition theorem characterizes the privacy of a composition of parallel or adaptive mechanisms in terms of the privacy guarantees of the individual mechanisms.

Lemma 46 (Composition of RDP Mironov (2017))**.** *Fix an order $\alpha > 1$. For each $i \in [k]$, let $\mathcal{A}_i : \mathcal{Z}^n \to \mathcal{W}_i$ be a randomized mechanism satisfying (α, ρ_i)-RDP. Then the following statements hold.*

(a) Joint (simultaneous) release. Let $\mathcal{A}(S) = (\mathcal{A}_1(S), \ldots, \mathcal{A}_k(S))$. Suppose $\{\mathcal{A}_i\}_{i=1}^k$ are independent. Then $\mathcal{A}$ satisfies $(\alpha, \sum_{i=1}^k \rho_i)$-RDP.

(b) Adaptive composition. Suppose $\mathcal{A}_1, \ldots, \mathcal{A}_k$ are applied sequentially, and for each $i \in [k]$, $\mathcal{A}_i$ may depend on the previous outputs $\mathcal{A}_1(S), \ldots, \mathcal{A}_{i-1}(S)$. If for every fixed realization $w_{<i} := (\mathcal{A}_1(S), \ldots, \mathcal{A}_{i-1}(S))$ of previous outputs, the conditional mechanism $\mathcal{A}_i(\cdot\,; w_{<i})$ satisfies (α, ρ_i)-RDP, then the overall mechanism

$$\mathcal{A}(S) = \big(\mathcal{A}_1(S), \mathcal{A}_2(S; \mathcal{A}_1(S)), \ldots, \mathcal{A}_k(S; \mathcal{A}_1(S), \ldots, \mathcal{A}_{k-1}(S))\big)$$

satisfies $(\alpha, \sum_{i=1}^k \rho_i)$-RDP.

Lemma 47 (One-step privacy of the uniformly subsampled Gaussian mechanism)**.** *Fix an iteration t and condition on the past iterate $\mathbf{W}_{t-1}$. Let $\widetilde{g}_{t,i} := g_{t,i} \cdot \min\big\{1, \frac{C_{\text{clip}}}{\|g_{t,i}\|_2}\big\}$. Consider the mechanism*

$$\mathcal{M}_t(S) = \frac{1}{B} \sum_{i \in \mathcal{B}_t} \widetilde{g}_{t,i} + \frac{C_{\text{clip}}}{B} \kappa Z_t, \qquad Z_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

where $\mathcal{B}_t$ is sampled uniformly without replacement from $[n]$ with $|\mathcal{B}_t| = B$. Then, in the standard low-sampling regime, there exists an absolute constant $c_{\text{sgm}} > 0$ such that M_t satisfies $\big(\alpha, c_{\text{sgm}} \frac{\alpha B^2}{n^2 \kappa^2}\big)$-RDP.

Proof. To ensure the consistency of the paper, we provide proof here. Condition on the past iterate $\mathbf{W}_{t-1}$. Then the clipped gradients $\{\widetilde{g}_{t,i}\}_{i=1}^n$ are deterministic vectors and satisfy $\|\widetilde{g}_{t,i}\|_2 \le C_{\text{clip}}$ for all $i \in [n]$. Define $U = (u_1, \ldots, u_B)$ by

$$M(U) = \frac{1}{B} \sum_{i=1}^B u_i + \frac{C_{\text{clip}}}{B} \kappa Z, \qquad Z \sim \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

If U and U' differ in exactly one entry, then

$$\Big\| \frac{1}{B} \sum_{i=1}^B u_i - \frac{1}{B} \sum_{i=1}^B u_i' \Big\|_2 \le \frac{2C_{\text{clip}}}{B}.$$

Therefore, M is a Gaussian mechanism with ℓ_2-sensitivity $\Delta = \frac{2C_{\text{clip}}}{B}$ and noise standard deviation $\sigma = \frac{C_{\text{clip}}}{B} \kappa$. By Lemma 44, we know M satisfies $(\alpha, \varepsilon_M(\alpha))$-RDP with $\varepsilon_M(\alpha) = \frac{\alpha \Delta^2}{2\sigma^2} = \frac{2\alpha}{\kappa^2}$. Note that $\mathcal{M}_t$ is exactly the mechanism $\mathcal{M}_t = M \circ \text{subsample}$ with subsampling ratio $q = \frac{B}{n}$.

Applying Theorem 9 of Wang et al. (2019) to $\mathcal{M}_t$, for every integer $\alpha \geq 2$,

$$\varepsilon_t(\alpha) \leq \frac{1}{\alpha-1}\log\Big(1+q^2\binom{\alpha}{2}\min\Big\{4\big(e^{\varepsilon_M(2)}-1\big),\ e^{\varepsilon_M(2)}\min\{2,(e^{\varepsilon_M(\infty)}-1)^2\}\Big\} + 2\sum_{j=3}^{\alpha} q^j\binom{\alpha}{j}e^{(j-1)\varepsilon_M(j)}\min\{2,(e^{\varepsilon_M(\infty)}-1)^j\}\Big). \tag{20}$$

Since $\varepsilon_M(\infty)=\infty$ and $\varepsilon_M(j)=\frac{2j}{\kappa^2}$, then

$$\varepsilon_t(\alpha) \leq \frac{1}{\alpha-1}\log\Big(1+q^2\binom{\alpha}{2}\min\Big\{4\big(e^{4/\kappa^2}-1\big),\ 2e^{4/\kappa^2}\Big\} + 2\sum_{j=3}^{\alpha} q^j\binom{\alpha}{j}e^{2j(j-1)/\kappa^2}\Big).$$

In the standard low-sampling/high-noise regime, the $j=2$ term dominates the above expression. Accordingly, as noted in Wang et al. (2019), the RDP of the subsampled Gaussian mechanism simplifies to

$$\varepsilon_t(\alpha) = \mathcal{O}\Big(\frac{\alpha q^2}{\kappa^2}\Big) = \mathcal{O}\Big(\frac{\alpha B^2}{n^2\kappa^2}\Big).$$

Hence there exists an absolute constant $c_{\mathrm{sgm}} > 0$ such that $\varepsilon_t(\alpha) \leq c_{\mathrm{sgm}}\frac{\alpha B^2}{n^2\kappa^2}$. The proof is completed. □

Based on the above lemmas, we now prove Theorem 48. We note the argument below is intended for the subsampled mini-batch regime $B < n$, where privacy amplification by subsampling is relevant. In the full-batch case $B = n$, amplification is not needed, and the privacy guarantee follows directly from the standard Gaussian mechanism (equivalently, from the full-batch DP-GD result in Wang et al. (2026)).

Theorem 48 (Privacy guarantee for DP-SGD with independent noise)**.** *Assume $\delta > 0$, $\lambda = 0$, and that the subsampled Gaussian mechanism operates in the regime of Lemma 47. If*

$$\kappa^2 \geq \frac{2c_{\mathrm{sgm}}B^2T}{n^2\epsilon} + \frac{4c_{\mathrm{sgm}}B^2T\log(1/\delta)}{n^2\epsilon^2},$$

then Algorithm 1 satisfies (ϵ,δ)-DP. For simplicity, we set $\kappa^2 = \frac{2c_{\mathrm{sgm}}B^2T}{n^2\epsilon} + \frac{4c_{\mathrm{sgm}}B^2T\log(1/\delta)}{n^2\epsilon^2}$.

Proof. At iteration t, the released noisy mini-batch gradient is

$$\hat{v}_t = \frac{1}{B}\sum_{i\in\mathcal{B}_t}\widetilde{g}_{t,i} + \frac{C_{\mathrm{clip}}}{B}\kappa Z_t, \qquad Z_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{mdp}),$$

since $\lambda = 0$ implies $\xi_t = \kappa Z_t$.

Condition on the past iterate $\mathbf{W}_{t-1}$. Then all clipped gradients $\{\widetilde{g}_{t,i}\}_{i=1}^n$ are deterministic vectors satisfying $\|\widetilde{g}_{t,i}\|_2 \leq C_{\mathrm{clip}}$. Therefore, by Lemma 47, the conditional mechanism that outputs $\hat{v}_t$ satisfies

$$\Big(\alpha, \frac{c_{\mathrm{sgm}}\alpha B^2}{n^2\kappa^2}\Big)\text{-RDP}.$$

Note the update $\mathbf{W}_t = \Pi_{\mathcal{K}}\big(\mathbf{W}_{t-1} - \eta\hat{v}_t\big)$ is a deterministic function of $\widehat{G}_t$ and $\mathbf{W}_{t-1}$. Hence, by Lemma 45, the conditional mechanism that outputs $\mathbf{W}_t$ also satisfies $\big(\alpha,\ \frac{c_{\mathrm{sgm}}\alpha B^2}{n^2\kappa^2}\big)$-RDP.

Applying Lemma 46 over $t = 0,\dots,T-1$, the full transcript $(\mathbf{W}_0,\dots,\mathbf{W}_{T-1})$ satisfies $\big(\alpha,\ \frac{c_{\mathrm{sgm}}\alpha TB^2}{n^2\kappa^2}\big)$-RDP. By further using post-processing property (Lemma 45), the final iterate $\mathbf{W}_{T-1}$ also satisfies $\big(\alpha,\ \frac{c_{\mathrm{sgm}}\alpha TB^2}{n^2\kappa^2}\big)$-RDP.

Choosing $\alpha = 1 + \frac{2\log(1/\delta)}{\epsilon}$. Then Lemma 43 gives that $\mathbf{W}_{T-1}$ satisfies $\big(c_{\mathrm{sgm}}\frac{\alpha TB^2}{n^2\kappa^2} + \frac{\epsilon}{2},\ \delta\big)$-DP. Note that $\kappa^2 \geq \frac{2c_{\mathrm{sgm}}\alpha TB^2}{n^2\epsilon}$ ensures

$$c_{\mathrm{sgm}}\frac{\alpha TB^2}{n^2\kappa^2} \leq \frac{\epsilon}{2}.$$

Hence, $\mathbf{W}_{T-1}$ satisfies (ϵ, δ)-DP. Substituting $\alpha = 1 + \frac{2\log(1/\delta)}{\epsilon}$ yields

$$\kappa^2 \geq \frac{2c_{\mathrm{sgm}}B^2T}{n^2\epsilon} + \frac{4c_{\mathrm{sgm}}B^2T\log(1/\delta)}{n^2\epsilon^2},$$

which completes the proof. □

F.2 Proofs for optimization of DP-SGD with independent noise

Recall

$$\mathcal{E}_\delta = \Big\{ \|\mathbf{c}\|_2 \leq 4\sqrt{pm} + 2\sqrt{\log(2/\delta)}, \quad \max_{j\in[m]} \|\mathbf{c}_j\|_2 \leq 4\sqrt{p} + 2\sqrt{\log(2m/\delta)} \Big\}$$

and

$$\mathcal{F}_{t-1} = \sigma\big(\mathbf{W}_0, \mathcal{B}_1, \ldots, \mathcal{B}_{t-1}, Z_1, \ldots, Z_{t-1}\big).$$

By Lemma 9, it holds that $\mathbb{P}(\mathcal{E}_\delta) \geq 1 - \delta$. For any $t \in [T]$, define

$$\Delta_t = \frac{1}{B}\sum_{i\in\mathcal{B}_t} g_{t,i} - \nabla\mathcal{L}_S(\mathbf{W}_{t-1}) \quad \text{with} \quad g_{t,i} = \nabla\ell\left(y_i f_{\mathbf{W}_{t-1}}(\mathbf{x}_i)\right). \tag{21}$$

Lemma 49 (Unbiasedness and variance when $\lambda = 0$)**.** *Suppose that $\lambda = 0$ and $C_{\mathrm{clip}} \geq G_\delta$, where G_δ is defined in equation 5. Let $\{\mathbf{W}_t\}_{t=0}^T$ be generated by Algorithm 1. Under the event $\mathcal{E}_\delta$, for all $t \in [T]$,*

$$\mathbb{E}_{\mathcal{B}_t}[\Delta_t \mid \mathcal{F}_{t-1}] = 0 \quad \textit{and} \quad \mathbb{E}_{\mathcal{B}_t}\big[\|\Delta_t\|_2^2 \mid \mathcal{F}_{t-1}\big] \leq \frac{G_\delta^2}{B}.$$

Here the expectation is taken with respect to the uniform sampling of the mini-batch $\mathcal{B}_t$.

Proof. Assume the event $\mathcal{E}_\delta$ holds. For any fixed $t \in [T]$, the uniform gradient bound equation 5 implies

$$\|g_{t,i}\|_2 = \big\|\nabla\ell\big(y_i f_{\mathbf{W}_{t-1}}(\mathbf{x}_i)\big)\big\|_2 \leq G_\delta \leq C_{\mathrm{clip}}, \qquad \forall i \in [n].$$

Hence clipping is inactive, and since $\lambda = 0$, the update becomes

$$\mathbf{W}_t = \Pi_{\mathcal{K}}\Big(\mathbf{W}_{t-1} - \eta\Big(\frac{1}{B}\sum_{i\in\mathcal{B}_t} g_{t,i} + \frac{C_{\mathrm{clip}}\kappa}{B}Z_t\Big)\Big).$$

Conditional on $\mathcal{F}_{t-1}$, the iterate $\mathbf{W}_{t-1}$ is fixed, and thus the vectors $g_{t,1}, \ldots, g_{t,n}$ are deterministic. Since $\mathcal{B}_t$ is sampled uniformly without replacement from $[n]$, the mini-batch average is an unbiased estimator of the full empirical gradient:

$$\mathbb{E}_{\mathcal{B}_t}\Big[\frac{1}{B}\sum_{i\in\mathcal{B}_t} g_{t,i}\,\Big|\,\mathcal{F}_{t-1}\Big] = \frac{1}{n}\sum_{i=1}^n g_{t,i} = \nabla\mathcal{L}_S(\mathbf{W}_{t-1}).$$

This proves the first claim.

Moreover, since $\|g_{t,i}\|_2 \leq G_\delta$ for all i, Lemma 4 gives

$$\mathbb{E}_{\mathcal{B}_t}\big[\|\Delta_t\|_2^2 \mid \mathcal{F}_{t-1}\big] \leq \frac{G_\delta^2}{B},$$

which completes the proof. □

Lemma 50. *Let $\delta \in (0,1)$. Suppose Assumption 1 holds, equation 1 holds, $\lambda = 0$ and $C_{\mathrm{clip}} \geq G_\delta$, where G_δ is defined in equation 5. Let $\{\mathbf{W}_t\}_{t=0}^{T-1}$ be generated by Algorithm 1 with $\eta \leq \frac{1}{12C_{\sigma,b}p^3}$. Assume*

$$m \geq 64C_{\sigma,b}^2 p^3\big(\sqrt{\log(m/\delta)} + \sqrt{p}\big)^2 R_*^4. \tag{22}$$

Then, for any comparator $\mathbf{W}^* \in \mathcal{K}$ and all $t \in [T]$, it holds on $\mathcal{E}_\delta$ that

$$\begin{aligned}\mathcal{L}_S(\mathbf{W}_{t-1}) \leq 4\mathcal{L}_S(\mathbf{W}^*) &+ \frac{3}{2\eta}\Big(\|\mathbf{W}_{t-1} - \mathbf{W}^*\|_2^2 - \|\mathbf{W}_t - \mathbf{W}^*\|_2^2\Big) \\ &+ 6\eta\|\Delta_t\|_2^2 + 6\eta\Big(\frac{C_{\mathrm{clip}}\kappa}{B}\Big)^2\|Z_t\|_2^2 - 3\Big\langle \Delta_t + \frac{C_{\mathrm{clip}}\kappa}{B}Z_t,\, \mathbf{W}_{t-1} - \mathbf{W}^*\Big\rangle.\end{aligned}$$

Proof. Assume $\mathcal{E}_\delta$ holds. Let $g_t := \nabla \mathcal{L}_S(\mathbf{W}_{t-1})$. Since clipping is inactive and $\lambda = 0$, the update can be written as

$$\mathbf{W}_t = \Pi_{\mathcal{K}}\Big(\mathbf{W}_{t-1} - \eta\big(g_t + \Delta_t + c_{\mathrm{priv}} Z_t\big)\Big) \quad \text{with} \quad c_{\mathrm{priv}} = \frac{C_{\mathrm{clip}}\kappa}{B}.$$

Applying Lemma 7 with $g = g_t + \Delta_t + c_{\mathrm{priv}} Z_t$, $\mathbf{W} = \mathbf{W}_{t-1}$, and $\mathbf{W}^+ = \mathbf{W}_t$, we obtain

$$\langle g_t + \Delta_t + c_{\mathrm{priv}} Z_t,\, \mathbf{W}_{t-1} - \mathbf{W}^* \rangle \le \frac{1}{2\eta}\Big(\|\mathbf{W}_{t-1} - \mathbf{W}^*\|_2^2 - \|\mathbf{W}_t - \mathbf{W}^*\|_2^2\Big) + \frac{\eta}{2}\|g_t + \Delta_t + c_{\mathrm{priv}} Z_t\|_2^2.$$

Hence,

$$\begin{aligned}\langle g_t, \mathbf{W}_{t-1} - \mathbf{W}^* \rangle \le \frac{1}{2\eta}\Big(\|\mathbf{W}_{t-1} - \mathbf{W}^*\|_2^2 - \|\mathbf{W}_t - \mathbf{W}^*\|_2^2\Big) + \frac{\eta}{2}\|g_t + \Delta_t + c_{\mathrm{priv}} Z_t\|_2^2 \\ - \langle \Delta_t + c_{\mathrm{priv}} Z_t, \mathbf{W}_{t-1} - \mathbf{W}^* \rangle. \end{aligned} \tag{23}$$

On the other hand, by Lemma 11, we know

$$\lambda_{\min}\big(\nabla^2 \mathcal{L}_S(\mathbf{W})\big) \ge -\frac{C_{\sigma,b}\, p^{3/2}\big(\sqrt{\log(m/\delta)} + \sqrt{p}\big)}{\sqrt{m}}\, \mathcal{L}_S(\mathbf{W}).$$

Since $\mathbf{W}_{t-1}, \mathbf{W}^* \in \mathcal{K} = \mathcal{B}(\mathbf{W}_0, R_*)$, it holds that $\|\mathbf{W}_{t-1} - \mathbf{W}^*\|_2 \le 2R_*$. Combining this with the condition equation 22 ensures that

$$\frac{C_{\sigma,b}\, p^{3/2}\big(\sqrt{\log(m/\delta)} + \sqrt{p}\big)}{\sqrt{m}}(2R_*)^2 \le \frac{1}{2}.$$

Therefore, by Lemma 8 and the same argument as in the proof of the correlated case (see proof of Lemma 13), we have

$$\mathcal{L}_S(\mathbf{W}_{t-1}) \le \mathcal{L}_S(\mathbf{W}^*) + \langle g_t, \mathbf{W}_{t-1} - \mathbf{W}^* \rangle + \frac{1}{3}\Big(\mathcal{L}_S(\mathbf{W}_{t-1}) + \mathcal{L}_S(\mathbf{W}^*)\Big).$$

Rearranging implies

$$\frac{2}{3}\,\mathcal{L}_S(\mathbf{W}_{t-1}) \le \frac{4}{3}\,\mathcal{L}_S(\mathbf{W}^*) + \langle g_t, \mathbf{W}_{t-1} - \mathbf{W}^* \rangle.$$

Substituting equation 23 in to the above observation yields

$$\begin{aligned}\frac{2}{3}\,\mathcal{L}_S(\mathbf{W}_{t-1}) \le \frac{4}{3}\,\mathcal{L}_S(\mathbf{W}^*) + \frac{1}{2\eta}\Big(\|\mathbf{W}_{t-1} - \mathbf{W}^*\|_2^2 - \|\mathbf{W}_t - \mathbf{W}^*\|_2^2\Big) + \frac{\eta}{2}\|g_t + \Delta_t + c_{\mathrm{priv}} Z_t\|_2^2 \\ - \langle \Delta_t + c_{\mathrm{priv}} Z_t, \mathbf{W}_{t-1} - \mathbf{W}^* \rangle. \end{aligned}$$

Note that

$$\|g_t + \Delta_t + c_{\mathrm{priv}} Z_t\|_2^2 \le 2\|g_t\|_2^2 + 4\|\Delta_t\|_2^2 + 4c_{\mathrm{priv}}^2\|Z_t\|_2^2 \le 8C_{\sigma,b}p^3\,\mathcal{L}_S(\mathbf{W}_{t-1}) + 4\|\Delta_t\|_2^2 + 4c_{\mathrm{priv}}^2\|Z_t\|_2^2,$$

where we have used Lemma 11. By further using $4C_{\sigma,b}p^3\eta \le \frac{1}{3}$ implied by $\eta \le (12C_{\sigma,b}p^3)^{-1}$, it holds

$$\frac{\eta}{2}\|g_t + \Delta_t + c_{\mathrm{priv}} Z_t\|_2^2 \le \frac{1}{3}\mathcal{L}_S(\mathbf{W}_{t-1}) + 2\eta\|\Delta_t\|_2^2 + 2\eta c_{\mathrm{priv}}^2\|Z_t\|_2^2.$$

Therefore,

$$\begin{aligned}\frac{1}{3}\,\mathcal{L}_S(\mathbf{W}_{t-1}) \le \frac{4}{3}\,\mathcal{L}_S(\mathbf{W}^*) + \frac{1}{2\eta}\Big(\|\mathbf{W}_{t-1} - \mathbf{W}^*\|_2^2 - \|\mathbf{W}_t - \mathbf{W}^*\|_2^2\Big) + 2\eta\|\Delta_t\|_2^2 + 2\eta c_{\mathrm{priv}}^2\|Z_t\|_2^2 \\ - \langle \Delta_t + c_{\mathrm{priv}} Z_t, \mathbf{W}_{t-1} - \mathbf{W}^* \rangle. \end{aligned}$$

Multiplying both sides by 3 and plugging c_{priv} back complete the proof. □

We now estimate the perturbation terms in Lemma 50.

Lemma 51. *Suppose the assumptions of Lemma 49 hold. Let $C > 0$ be an absolute constant. Then, conditional on the dataset and the initialization satisfying $\mathcal{E}_\delta$ (see equation 1), the following statements hold true.*

(a) For any $\delta_{\Delta^2} \in (0,1)$,

$$\mathbb{P}_{\mathcal{A}}\Big(\sum_{t=1}^{T}\|\Delta_t\|_2^2 \le \frac{2TG_\delta^2}{B} + 8G_\delta^2\log\Big(\frac{1}{\delta_{\Delta^2}}\Big)\Big) \ge 1-\delta_{\Delta^2}.$$

(b) For any $\delta_{Z^2} \in (0,1)$,

$$\mathbb{P}\Big(\sum_{t=1}^{T}\|Z_t\|_2^2 \le Tmdp + 2\sqrt{Tmdp\log(1/\delta_{Z^2})} + 2\log\big(\frac{1}{\delta_{Z^2}}\big)\Big) \ge 1-\delta_{Z^2}.$$

(c) Let $\mathbf{W}^* \in \mathcal{K}$, and define $\mathfrak{M}_T^{(\Delta)} := -3\sum_{t=1}^{T}\langle\Delta_t, \mathbf{W}_{t-1}-\mathbf{W}^*\rangle$. Then, for any $\delta_{\Delta,\mathrm{lin}} \in (0,1)$,

$$\mathbb{P}_{\mathcal{A}}\Big(|\mathfrak{M}_T^{(\Delta)}| \le C\,G_\delta R_*\sqrt{\frac{T\log(1/\delta_{\Delta,\mathrm{lin}})}{B}}\Big) \ge 1-\delta_{\Delta,\mathrm{lin}}.$$

(d) Let $\mathbf{W}^* \in \mathcal{K}$. Define $\mathfrak{M}_T^{(Z)} := -3\frac{C_{\mathrm{clip}}\kappa}{B}\sum_{t=1}^{T}\langle Z_t, \mathbf{W}_{t-1}-\mathbf{W}^*\rangle$. Then, for any $\delta_{Z,\mathrm{lin}} \in (0,1)$,

$$\mathbb{P}_{\mathcal{A}}\Big(|\mathfrak{M}_T^{(Z)}| \le C\,\frac{C_{\mathrm{clip}}\kappa}{B}\,R_*\sqrt{T\log(1/\delta_{Z,\mathrm{lin}})}\Big) \ge 1-\delta_{Z,\mathrm{lin}}.$$

Proof. Part (a) is proved exactly as Lemma 17, but without the stopping indicator. And part (b) directly from Lemma 15.

For the part (c), conditional on the pre-sampling filtration, $\mathbf{W}_{t-1}-\mathbf{W}^*$ is fixed and has norm at most $2R_*$, while Δ_t is centered. The proof is the same as that of Lemma 18, with $\bar{R}$ replaced by $2R_*$ and without the stopping indicator.

Finally, conditional on the filtration up to time $t-1$, the vector $\mathbf{W}_{t-1}-\mathbf{W}^*$ is fixed and has norm at most $2R_*$, while $Z_t \sim \mathcal{N}(0, \mathbf{I}_{mdp})$ is independent of the past. Hence each summand is conditionally Gaussian with variance bounded by $C(C_{\mathrm{clip}}\kappa R_*/B)^2$. The claim then follows from the standard maximal inequality for Gaussian martingale differences, exactly as in Lemma 19. The proof is complete. □

Theorem 52 (General optimization risk bound for standard DP-SGD)**.** *Let $\delta \in (0,1)$. Suppose Assumption 1 holds. Let $\{\mathbf{W}_t\}_{t=0}^{T}$ be generated by Algorithm 1 with $\lambda = 0$, $\eta \le \frac{1}{12C_{\sigma,b}p^3}$ and $C_{\mathrm{clip}} \ge G_\delta$. Assume $m \ge 64C_{\sigma,b}^2p^3\big(\sqrt{\log(m/\delta)}+\sqrt{p}\big)^2R_*^4$. Then, conditioned on the dataset and the initialization satisfying $\mathcal{E}_\delta$, with probability at least $1-\delta$ over the algorithmic randomness,*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathcal{L}_S(\mathbf{W}_t) \lesssim \mathcal{L}_S(\mathbf{W}^*) + \frac{\|\mathbf{W}_0-\mathbf{W}^*\|_2^2}{\eta T} + G_\delta R_*\sqrt{\frac{\log(1/\delta)}{BT}} + \frac{C_{\mathrm{clip}}\kappa}{B}\,R_*\sqrt{\frac{\log(1/\delta)}{T}} + \eta G_\delta^2\Big(\frac{1}{B}+\frac{\log(1/\delta)}{T}\Big) + \eta\frac{C_{\mathrm{clip}}^2\kappa^2}{B^2}\Big(md+\frac{\log(1/\delta)}{T}\Big).$$

Proof. Fix auxiliary failure probabilities $\delta_{\Delta^2} = \delta_{\Delta,\mathrm{lin}} = \delta_{Z^2} = \delta_{Z,\mathrm{lin}} = \frac{\delta}{4}$. Since replacing δ by a constant fraction only affects logarithmic terms by absolute constants, we suppress this distinction below.

Assume the event $\mathcal{E}_\delta$ defined in equation 1 holds. Summing the inequality in Lemma 50 over $t = 1, \dots, T$ yields

$$\sum_{t=0}^{T-1}\mathcal{L}_S(\mathbf{W}_t) \le 4T\mathcal{L}_S(\mathbf{W}^*) + \frac{3}{2\eta}\|\mathbf{W}_0-\mathbf{W}^*\|_2^2 + 6\eta\sum_{t=1}^{T}\|\Delta_t\|_2^2 + 6\eta\Big(\frac{C_{\mathrm{clip}}\kappa}{B}\Big)^2\sum_{t=1}^{T}\|Z_t\|_2^2 + \mathfrak{M}_T^{(\Delta)} + \mathfrak{M}_T^{(Z)}.$$

Plugging the results of Lemma 51 into the above inequality and dividing by T gives

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathcal{L}_S(\mathbf{W}_t) \lesssim \mathcal{L}_S(\mathbf{W}^*) + \frac{\|\mathbf{W}_0-\mathbf{W}^*\|_2^2}{\eta T} + G_\delta R_*\sqrt{\frac{\log(1/\delta)}{BT}} + \frac{C_{\text{clip}}\kappa}{B}\,R_*\sqrt{\frac{\log(1/\delta)}{T}} + \eta G_\delta^2\Big(\frac{1}{B}+\frac{\log(1/\delta)}{T}\Big) + \eta\frac{C_{\text{clip}}^2\kappa^2}{B^2}\Big(md+\frac{\log(1/\delta)}{T}\Big).$$

The stated probability follows from a union bound over the four good events. □

Now, we present our optimization risk bound under the NTK separability assumption.

Theorem 53 (Optimization Risk Bound)**.** *Let $\delta \in (0,1)$. Suppose Assumptions 1 and 2 hold. Let $\{\mathbf{W}_t\}_{t=0}^T$ be generated by Algorithm 1 with $\eta \le \frac{1}{12C_{\sigma,b}p^3}$ and $C_{\text{clip}} = G_\delta$. Assume $m \gtrsim \max\big\{\frac{\log(\frac{m}{\delta})(\log^4(T)+\log^2(\frac{n}{\delta}))}{\gamma^4}, \frac{\log(nT/\delta)}{Td}\big\}$ and $R_* \asymp \frac{1}{\gamma}(\log(T)+\sqrt{\log(n/\delta)})$. Then, with probability at least $1-\delta$ over the randomness of the initialization and the algorithmic randomness,*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathcal{L}_S(\mathbf{W}_t) \lesssim \frac{\log^2(T)+\log(n/\delta)}{\gamma^2\eta T} + \frac{\log(1/\delta)\big(\log(T)+\sqrt{\log(n/\delta)}\big)}{\gamma\sqrt{BT}} + \frac{\log(1/\delta)\big(\log(T)+\sqrt{\log(n/\delta)}\big)\kappa}{\gamma B\sqrt{T}} + \eta\log(1/\delta)\Big(\frac{1}{B}+\frac{\log(1/\delta)}{T}\Big) + \frac{\eta md\log^2(1/\delta)\kappa^2}{B^2} =: A_{\text{van}}.$$

Proof. Fix $\delta_{\text{init}} = \delta_{\text{ntk}} = \delta_{\text{alg}} = \frac{\delta}{3}$. Take

$$R_* \asymp \frac{1}{\gamma}\big(\log(T)+\sqrt{\log(n/\delta)}\big).$$

Under the corresponding width conditions, Lemma 23 implies that, with probability at least $1-\delta_{\text{ntk}}$ over the initialization, there exists a comparator $\mathbf{W}^* \in \mathcal{K}$ such that

$$\mathcal{L}_S(\mathbf{W}^*) \le \frac{1}{T}, \qquad \|\mathbf{W}^*-\mathbf{W}_0\|_2^2 \lesssim \frac{\log^2(T)+\log(n/\delta)}{\gamma^2}.$$

Intersecting this event with $\mathcal{E}_{\delta_{\text{init}}}$ and then applying Theorem 52 with failure probability δ_{alg}, we obtain, with probability at least $1-\delta$,

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathcal{L}_S(\mathbf{W}_t) \lesssim \frac{1}{T} + \frac{\log^2(T)+\log(n/\delta)}{\gamma^2\eta T} + G_\delta R_*\sqrt{\frac{\log(1/\delta)}{BT}} + \frac{C_{\text{clip}}\kappa}{B}\,R_*\sqrt{\frac{\log(1/\delta)}{T}} + \eta G_\delta^2\Big(\frac{1}{B}+\frac{\log(1/\delta)}{T}\Big) + \eta\frac{C_{\text{clip}}^2\kappa^2}{B^2}\Big(md+\frac{\log(1/\delta)}{T}\Big).$$

Recall that $C_{\text{clip}} = G_\delta \lesssim \sqrt{\log(1/\delta)}$, $\kappa^2 \asymp \frac{B^2T\log(1/\delta)}{n^2\epsilon^2}$ and $R_* \asymp \frac{1}{\gamma}\big(\log(T)+\sqrt{\log(n/\delta)}\big)$.

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathcal{L}_S(\mathbf{W}_t) \lesssim \frac{\log^2(T)+\log(n/\delta)}{\gamma^2\eta T} + \frac{\log(1/\delta)\big(\log(T)+\sqrt{\log(n/\delta)}\big)}{\gamma\sqrt{BT}} + \frac{\log(1/\delta)\big(\log(T)+\sqrt{\log(n/\delta)}\big)\kappa}{\gamma B\sqrt{T}} + \eta\log(1/\delta)\Big(\frac{1}{B}+\frac{\log(1/\delta)}{T}\Big) + \frac{\eta md\log^2(1/\delta)\kappa^2}{B^2}.$$

This proves the theorem. □

F.3 Proofs for population of DP-SGD with independent noise

The stability and generalization analysis developed for the correlated-noise case is noise-agnostic once the corresponding optimization good event is available. Therefore, when $\lambda = 0$, the same argument applies verbatim after replacing A_{corr} by A_{van}. Recall that $U_{\delta,\delta_f} \lesssim \sqrt{\log(n/\delta)}$ and $b_{\delta,m} \lesssim \sqrt{\log(1/\delta)}$.

Theorem 54 (On-average argument stability for standard DP-SGD)**.** *Let $\delta \in (0,1)$. Suppose Assumption 1 holds and the parameter conditions of Theorem 53 are satisfied. Assume further that, conditional on the initialization, an i.i.d. sample $S \sim \mathcal{P}^n$ satisfies Assumption 2 with margin γ with probability at least $1-\delta_{\mathrm{sep}}$, where $\delta_{\mathrm{sep}} \le \delta/8$. Assume*

$$m \gtrsim \log(\frac{m}{\delta})\eta^2\Big(T^2A_{\mathrm{van}}^2 + \log(n/\delta)\log^2(1/\delta)\Big(\frac{T}{B}+1\Big)\Big).$$

Then, with probability at least $1-\delta$ over initialization,

$$\mathbb{E}_{S,\widetilde{S},\mathcal{A}}\Big[\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{W}_T - \mathbf{W}_T^{(i)}\|_2\Big] \lesssim \frac{\eta\sqrt{\log(1/\delta)}}{n}\sum_{t=0}^{T-1}\mathbb{E}_{S,\mathcal{A}}\big[\mathcal{L}_S(\mathbf{W}_t)\big] + \frac{(\log(T)+\sqrt{\log(n/\delta)})\delta}{\gamma}.$$

Proof. The proof follows the stability argument in the proof of Theorem 2, including the conditionalization and sample-separability arguments for S and $S^{(i)}$. The only difference is that the optimization good event $\mathcal{E}_{\mathrm{opt}}^{(i)}$ is now supplied by Theorem 53 instead of Theorem 1. Under the stated lower bound on m, Lemma 30 applies with A_{corr} replaced by A_{van}, and the rest of the argument is unchanged. □

Theorem 55 (Population risk bound)**.** *Suppose Assumption 1 holds and the parameter conditions of Theorem 53 are satisfied. Assume further that, conditional on the initialization, an i.i.d. sample $S \sim \mathcal{P}^n$ satisfies Assumption 2 with margin γ with probability at least $1-\delta_{\mathrm{sep}}$, where $\delta_{\mathrm{sep}} \le \frac{\delta}{8(1+\log(\frac{n}{\delta})(\sqrt{m}+R_*))}$. Assume*

$$m \gtrsim \log(\frac{m}{\delta})\eta^2\Big(T^2A_{\mathrm{van}}^2 + \log(n/\delta)\log^2(1/\delta)\Big(\frac{T}{B}+1\Big)\Big).$$

Then, with probability at least $1-\delta$ over initialization,

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_{S,\mathcal{A}}\big[\mathcal{L}(\mathbf{W}_t)\big] \lesssim \Big(1+\frac{\eta T\log(\frac{1}{\delta})}{n}\Big)\big(A_{\mathrm{van}} + \delta\sqrt{\log(\frac{n}{\delta})}\big) + \sqrt{\log(\frac{1}{\delta})}\frac{(\log(T)+\sqrt{\log(\frac{n}{\delta})})\delta}{\gamma}.$$

Proof. By Lemma 10 and $|\ell'(u)| \le 1$ for the logistic loss, for every $\mathbf{W} \in \mathcal{K}$ and every $(\mathbf{x},y) \in \mathcal{Z}$,

$$\|\nabla_{\mathbf{W}}\ell(yf_{\mathbf{W}}(\mathbf{x}))\|_2 = |\ell'(yf_{\mathbf{W}}(\mathbf{x}))|\,\|\nabla f_{\mathbf{W}}(\mathbf{x})\|_2 \le b_{\delta_0,m}.$$

Hence the loss is $b_{\delta_0,m}$-Lipschitz with respect to $\mathbf{W}$. Applying Lemma 25 together with Theorem 54 yields

$$\mathbb{E}_{S,\mathcal{A}}\big[\mathcal{L}(\mathbf{W}_T) - \mathcal{L}_S(\mathbf{W}_T)\big] \lesssim \frac{\eta\log(1/\delta)}{n}\sum_{t=0}^{T-1}\mathbb{E}_{S,\mathcal{A}}\big[\mathcal{L}_S(\mathbf{W}_t)\big] + \sqrt{\log(\frac{1}{\delta})}\frac{(\log(T)+\sqrt{\log(n/\delta)})\delta}{\gamma}. \tag{24}$$

Now, we consider the population risk bound. Since replacing δ by a smaller auxiliary failure probability only changes logarithmic factors, we suppress this distinction below.

On the initialization event $\mathcal{E}_{\delta_0}$, for any $\mathbf{x} \in \mathcal{X}$, any $y \in \{-1,+1\}$, and any $\mathbf{W} \in \mathcal{K} = B(\mathbf{W}_0, R_*)$, it holds that

$$|f_{\mathbf{W}}(\mathbf{x})| \le |f_{\mathbf{W}_0}(\mathbf{x})| + \|\nabla f_{\widetilde{\mathbf{W}}}(\mathbf{x})\|_2\,\|\mathbf{W}-\mathbf{W}_0\|_2 \le B_b\sqrt{p}\,\|c\|_2 + b_{\delta_0,m}R_* \lesssim \sqrt{m} + \log(\tfrac{n}{\delta})R_*,$$

where we used $\|c\|_2 \lesssim \log(1/\delta_0)\sqrt{m}$ on $\mathcal{E}_{\delta_0}$, the bound $\|\nabla f_{\widetilde{\mathbf{W}}}(\mathbf{x})\|_2 \le b_{\delta_0,m}$ from Lemma 10, and $\|\mathbf{W}-\mathbf{W}_0\|_2 \le R_*$. Consequently,

$$\ell(yf_{\mathbf{W}}(\mathbf{x})) = \log\big(1+e^{-yf_{\mathbf{W}}(\mathbf{x})}\big) \le \log 2 + |f_{\mathbf{W}}(\mathbf{x})| \lesssim \sqrt{m} + \log(\tfrac{n}{\delta})R_*.$$

Since $\mathbf{W}_t \in \mathcal{K}$ for all t by construction of Algorithm 1, we obtain the uniform bound

$$\mathcal{L}_S(\mathbf{W}_t) \lesssim \log(\tfrac{n}{\delta})(\sqrt{m} + R_*) \qquad \forall t.$$

Now choose an auxiliary optimization failure probability $\delta'_{\mathrm{opt}} = \frac{\delta}{1+\log(\frac{n}{\delta})(\sqrt{m}+R_*)}$. Applying Theorem 53 with a sufficiently smaller auxiliary failure probability and Lemma 33, there exists an initialization event $\mathcal{E}'_{\mathrm{cond}}$ with probability at least $1 - \delta/8$ such that, on $\mathcal{E}'_{\mathrm{cond}}$, the conditional optimization failure probability is at most δ'_{opt}. Here the smaller auxiliary failure probability only changes logarithmic factors. Let $\mathcal{E}_{\mathrm{opt}}$ be the event on which S satisfies Assumption 2 with margin γ and

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathcal{L}_S(\mathbf{W}_t) \le A_{\mathrm{van}}.$$

By the sample-separability assumption, on $\mathcal{E}'_{\mathrm{cond}}$, $\mathbb{P}(\mathcal{E}^c_{\mathrm{opt}}) \le \delta_{\mathrm{sep}} + \delta'_{\mathrm{opt}}$. Therefore, conditioning on $\mathcal{E}_{\delta_0} \cap \mathcal{E}'_{\mathrm{cond}}$, we have

$$\begin{aligned}
\mathbb{E}_{S,\mathcal{A}}\Big[\frac{1}{T}\sum_{t=0}^{T-1} \mathcal{L}_S(\mathbf{W}_t)\Big] &= \mathbb{E}_{S,\mathcal{A}}\Big[\frac{1}{T}\sum_{t=0}^{T-1} \mathcal{L}_S(\mathbf{W}_t)\mathbf{1}_{\mathcal{E}_{\mathrm{opt}}}\Big] + \mathbb{E}_{S,\mathcal{A}}\Big[\frac{1}{T}\sum_{t=0}^{T-1} \mathcal{L}_S(\mathbf{W}_t)\mathbf{1}_{\mathcal{E}^c_{\mathrm{opt}}}\Big] \\
&\lesssim A_{\mathrm{van}} + \log(\tfrac{n}{\delta})(\sqrt{m} + R_*)\,\mathbb{P}(\mathcal{E}^c_{\mathrm{opt}}) \\
&\lesssim A_{\mathrm{van}} + \delta \lesssim A_{\mathrm{van}} + \delta\sqrt{\log(\tfrac{n}{\delta})}.
\end{aligned}$$

Combining the above observation with the first part of the theorem gives

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}_{S,\mathcal{A}}[\mathcal{L}(\mathbf{W}_t)] \lesssim A_{\mathrm{van}} + \delta\sqrt{\log(\tfrac{n}{\delta})} + \frac{\eta\log(\frac{1}{\delta})}{n}\sum_{t=0}^{T-1} \mathbb{E}_{S,\mathcal{A}}[\mathcal{L}_S(\mathbf{W}_t)] \\
+ \sqrt{\log(\tfrac{1}{\delta})}\frac{(\log(T) + \sqrt{\log(\frac{n}{\delta})})\delta}{\gamma}.
\end{aligned}$$

Substituting

$$\sum_{t=0}^{T-1} \mathbb{E}_{S,\mathcal{A}}[\mathcal{L}_S(\mathbf{W}_t)] \lesssim T\big(A_{\mathrm{van}} + \delta\sqrt{\log(\tfrac{n}{\delta})}\big)$$

yields the claimed bound. □

The result below shows that independent noise DP-SGD attains the rate $\mathcal{O}(n^{-1/2} + \sqrt{d}/(n\epsilon))$.

Corollary 56 (Risk rates under representative parameter regime)**.** *Suppose Assumptions 1 and 2 hold. Assume $\delta \le \frac{\gamma}{n}$. Let $\{\mathbf{W}_t\}_{t=0}^{T-1}$ be generated by Algorithm 1 with $\eta \le \frac{1}{12C_{\sigma,b}p^3}$ and $C_{\mathrm{clip}} = G_\delta$. If $m \asymp \frac{\mathrm{polylog}(n/\delta)}{\gamma^4}$, set $B \asymp \gamma\sqrt{n}$, $\eta \asymp \min\big\{1, \frac{\gamma^2\sqrt{n}\,\epsilon}{\sqrt{d}}\big\}$ and $T \asymp \frac{\sqrt{n}}{\gamma}$. With probability at least $1-\delta$ over the initialization and the algorithmic randomness,*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathcal{L}_S(\mathbf{W}_t) \lesssim \frac{1}{\gamma\sqrt{n}} + \frac{\sqrt{d}}{\gamma^3 n\epsilon}.$$

For the population risk statement, assume further the sample-separability condition of Theorem 55. Moreover, with probability at least $1-\delta$ over the initialization,

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}_{S,\mathcal{A}}\big[\mathcal{L}(\mathbf{W}_t)\big] \lesssim \frac{1}{\gamma\sqrt{n}} + \frac{\sqrt{d}}{\gamma^3 n\epsilon}.$$

Proof. We first check the optimization bound in Theorem 53. Under the stated choice of parameters, $\frac{1}{\gamma\sqrt{BT}} \asymp \frac{1}{\gamma\sqrt{\gamma\sqrt{n}\cdot\sqrt{n}/\gamma}} \asymp \frac{1}{\gamma\sqrt{n}}$. Moreover, $\frac{1}{\gamma^2\eta T} \asymp \frac{1}{\gamma\eta\sqrt{n}}$. If $\eta \asymp 1$, then $\frac{1}{\gamma^2\eta T} \asymp \frac{1}{\gamma\sqrt{n}}$. If $\eta \asymp \frac{\gamma^2\sqrt{n}\,\epsilon}{\sqrt{d}}$, then $\frac{1}{\gamma^2\eta T} \asymp \frac{\sqrt{d}}{\gamma^3 n\epsilon}$. Hence, in all cases, $\frac{1}{\gamma^2\eta T} \lesssim \frac{1}{\gamma\sqrt{n}} + \frac{\sqrt{d}}{\gamma^3 n\epsilon}$.

For the term η/B, if $\eta \asymp 1$, then $\frac{\eta}{B} \asymp \frac{1}{\gamma\sqrt{n}}$. If $\eta \asymp \frac{\gamma^2\sqrt{n}\,\epsilon}{\sqrt{d}}$, then $\frac{\eta}{B} \asymp \frac{\gamma\epsilon}{\sqrt{d}} \lesssim \frac{\sqrt{d}}{\gamma^3 n\epsilon}$, where the last inequality follows from the branch condition $\gamma^2\sqrt{n}\,\epsilon \le \sqrt{d}$. Therefore, in all cases, $\frac{\eta}{B} \lesssim \frac{1}{\gamma\sqrt{n}} + \frac{\sqrt{d}}{\gamma^3 n\epsilon}$.

By Theorem 48, $\kappa^2 \asymp \frac{B^2 T}{n^2\epsilon^2}$. Since $B \asymp \gamma\sqrt{n}$ and $T \asymp \sqrt{n}/\gamma$, this gives $\kappa^2 \lesssim \frac{\gamma}{\sqrt{n}\,\epsilon^2}$, and $\frac{\kappa}{\gamma B\sqrt{T}} \asymp \frac{1}{\gamma n\epsilon} \lesssim \frac{\sqrt{d}}{\gamma^3 n\epsilon}$.

For the quadratic noise term, using $\kappa^2/B^2 \asymp T/(n^2\epsilon^2)$, we have $\frac{\eta m d\kappa^2}{B^2} \asymp \frac{\eta m d T}{n^2\epsilon^2}$. If $\eta \asymp 1$, then $\frac{\eta m d\kappa^2}{B^2} \asymp \frac{md}{\gamma n^{3/2}\epsilon^2} \lesssim \frac{1}{\gamma\sqrt{n}}$, where we used the branch condition $\epsilon \ge \sqrt{d}/(\gamma^2\sqrt{n})$ and $m \asymp \mathrm{polylog}(n/\delta)/\gamma^4$. If $\eta \asymp \frac{\gamma^2\sqrt{n}\,\epsilon}{\sqrt{d}}$, then $\frac{\eta m d\kappa^2}{B^2} \asymp \frac{m\gamma\sqrt{d}}{n\epsilon} \lesssim \frac{\sqrt{d}}{\gamma^3 n\epsilon}$. Hence, $\frac{\eta m d\kappa^2}{B^2} \lesssim \frac{1}{\gamma\sqrt{n}} + \frac{\sqrt{d}}{\gamma^3 n\epsilon}$.

Combining the above estimates, Theorem 53 yields

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathcal{L}_S(\mathbf{W}_t) \lesssim \frac{1}{\gamma\sqrt{n}} + \frac{\sqrt{d}}{\gamma^3 n\epsilon}$$

with probability at least $1-\delta$.

Now we show the generalization and population risk rates. By equation 24 and the above optimization bound,

$$\mathbb{E}_{S,\mathcal{A}}\big[\mathcal{L}(\mathbf{W}_T) - \mathcal{L}_S(\mathbf{W}_T)\big] \lesssim \frac{\eta\log(1/\delta)}{n}\sum_{t=0}^{T-1}\mathbb{E}_{S,\mathcal{A}}\big[\mathcal{L}_S(\mathbf{W}_t)\big] + \frac{\delta}{\gamma},$$

where we suppress logarithmic factors in the first term. Therefore,

$$\mathbb{E}_{S,\mathcal{A}}\big[\mathcal{L}(\mathbf{W}_T) - \mathcal{L}_S(\mathbf{W}_T)\big] \lesssim \frac{\eta T}{n}\Big(\frac{1}{\gamma\sqrt{n}} + \frac{\sqrt{d}}{\gamma^3 n\epsilon}\Big) + \frac{\delta}{\gamma}.$$

Since $\frac{\eta T}{n} \asymp \min\big\{\frac{1}{\gamma\sqrt{n}}, \frac{\gamma\epsilon}{\sqrt{d}}\big\}$, we have $\frac{\eta T}{n}\cdot\frac{1}{\gamma\sqrt{n}} \lesssim \frac{1}{\gamma^2 n}$ and $\frac{\eta T}{n}\cdot\frac{\sqrt{d}}{\gamma^3 n\epsilon} \lesssim \frac{1}{\gamma^2 n}$. Because $\delta \lesssim \gamma/n$, the residual term δ/γ is also of order at most $1/n$, and hence

$$\mathbb{E}_{S,\mathcal{A}}\big[\mathcal{L}(\mathbf{W}_T) - \mathcal{L}_S(\mathbf{W}_T)\big] \lesssim \frac{1}{\gamma^2 n}.$$

Further, Theorem 55 implies

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_{S,\mathcal{A}}\big[\mathcal{L}(\mathbf{W}_t)\big] \lesssim \Big(1 + \frac{\eta T}{n}\Big)\big(A_{\mathrm{van}} + \delta\big) + \frac{\delta}{\gamma},$$

where logarithmic factors are suppressed. Since $\eta T/n \lesssim 1$, $\delta \lesssim \gamma/n$, and $A_{\mathrm{van}} \lesssim \frac{1}{\gamma\sqrt{n}} + \frac{\sqrt{d}}{\gamma^3 n\epsilon}$, we obtain

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_{S,\mathcal{A}}\big[\mathcal{L}(\mathbf{W}_t)\big] \lesssim \frac{1}{\gamma\sqrt{n}} + \frac{\sqrt{d}}{\gamma^3 n\epsilon}.$$

Finally, we verify that the width conditions in Theorems 53 and 55 are satisfied. Since $m \asymp \frac{\mathrm{polylog}(n/\delta)}{\gamma^4}$, it suffices to check that $\eta^2\Big(T^2 A_{\mathrm{van}}^2 + \frac{T}{B} + 1\Big) \lesssim \frac{1}{\gamma^4}$. Since $T/B \asymp 1/\gamma^2$, then $\eta^2(T/B+1) \lesssim 1/\gamma^4$. Also, $A_{\mathrm{van}} \lesssim \frac{1}{\gamma\sqrt{n}} + \frac{\sqrt{d}}{\gamma^3 n\epsilon}$, and $\frac{\eta T}{\gamma\sqrt{n}} \lesssim \frac{1}{\gamma^2}$, $\frac{\eta T\sqrt{d}}{\gamma^3 n\epsilon} \lesssim \frac{1}{\gamma^2}$ in both cases $\eta \asymp 1$ and $\eta \asymp \frac{\gamma^2\sqrt{n}\,\epsilon}{\sqrt{d}}$. Therefore $(\eta T)^2 A_{\mathrm{van}}^2 \lesssim \frac{1}{\gamma^4}$, and Theorems 53 and 55 apply. This completes the proof. □

Recovery of Full-batch DP-GD. Since $B = n$, the mini-batch gradient coincides with the full empirical gradient, so $\Delta_t = 0$ for all t (see equation 21). The sampling fluctuation term disappears and the recursion reduces to full-batch DP-GD. We therefore specialize the bounds in Theorems 53 and 55 to this regime. For the generalization and population risk statements below, assume further the sample-separability condition of Theorem 55.

Assume $m \asymp \frac{\mathrm{polylog}(n/\delta)}{\gamma^4}$, $T \asymp n$, $\eta \asymp \min\Big\{1, \frac{\epsilon}{\gamma\sqrt{md}}\Big\}$, $C_{\mathrm{clip}} = G_\delta$, and $\delta \lesssim \gamma/n$. For $B = n$, the privacy guarantee follows directly from the standard Gaussian mechanism. Up to logarithmic factors, we have $\kappa^2 \asymp \frac{T}{\epsilon^2} \asymp \frac{n}{\epsilon^2}$ and $\kappa \asymp \frac{\sqrt{n}}{\epsilon}$.

Plugging these parameter choices into the optimization bound in Theorem 53, we first note that $\frac{1}{\gamma n} \lesssim \frac{1}{\gamma^2 n}$, $\frac{\eta}{n} \lesssim \frac{1}{\gamma^2 n}$ and $\frac{\kappa}{\gamma n\sqrt{n}} \asymp \frac{1}{\gamma n\epsilon} \lesssim \frac{\sqrt{d}}{\gamma^3 n\epsilon}$. Moreover, $\frac{1}{\gamma^2\eta n} \lesssim \frac{1}{\gamma^2 n} + \frac{\sqrt{md}}{\gamma n\epsilon} \lesssim \frac{1}{\gamma^2 n} + \frac{\sqrt{d}}{\gamma^3 n\epsilon}$, where we used $m \asymp \gamma^{-4}\mathrm{polylog}(n/\delta)$. Similarly, $\frac{\eta m d\kappa^2}{n^2} \asymp \frac{\eta m d}{n\epsilon^2} \lesssim \frac{1}{\gamma^2 n} + \frac{\sqrt{d}}{\gamma^3 n\epsilon}$. Therefore,

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathcal{L}_S(\mathbf{W}_t) \lesssim \frac{1}{\gamma^2 n} + \frac{\sqrt{d}}{\gamma^3 n\epsilon}$$

with probability at least $1 - \delta$.

For the generalization risk, Theorem 55 gives

$$\mathbb{E}_{S,\mathcal{A}}\big[\mathcal{L}(\mathbf{W}_T) - \mathcal{L}_S(\mathbf{W}_T)\big] \lesssim \frac{\eta}{n}\sum_{t=0}^{T-1} \mathbb{E}_{S,\mathcal{A}}\big[\mathcal{L}_S(\mathbf{W}_t)\big] + \frac{\delta}{\gamma}.$$

. Using $T \asymp n$ and the optimization bound above,

$$\mathbb{E}_{S,\mathcal{A}}\big[\mathcal{L}(\mathbf{W}_T) - \mathcal{L}_S(\mathbf{W}_T)\big] \lesssim \eta\Big(\frac{1}{\gamma^2 n} + \frac{\sqrt{d}}{\gamma^3 n\epsilon}\Big) + \frac{\delta}{\gamma}.$$

By the choice $\eta \asymp \min\{1, \epsilon/(\gamma\sqrt{md})\}$ and $m \asymp \gamma^{-4}\mathrm{polylog}(n/\delta)$, $\eta\frac{\sqrt{d}}{\gamma^3 n\epsilon} \lesssim \frac{1}{\gamma^2 n}$. Together with $\delta \lesssim \gamma/n$, this yields

$$\mathbb{E}_{S,\mathcal{A}}\big[\mathcal{L}(\mathbf{W}_T) - \mathcal{L}_S(\mathbf{W}_T)\big] \lesssim \frac{1}{\gamma^2 n}.$$

Furthermore, Theorem 55 shows

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}_{S,\mathcal{A}}\big[\mathcal{L}(\mathbf{W}_t)\big] \lesssim \Big(1 + \frac{\eta T}{n}\Big)\big(A_{\mathrm{van}} + \delta\big) + \frac{\delta}{\gamma}.$$

Since $T \asymp n$, $\eta \lesssim 1$, $\delta \lesssim \gamma/n$, and $A_{\mathrm{van}} \lesssim \frac{1}{\gamma^2 n} + \frac{\sqrt{d}}{\gamma^3 n\epsilon}$, we obtain

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}_{S,\mathcal{A}}\big[\mathcal{L}(\mathbf{W}_t)\big] \lesssim \frac{1}{\gamma^2 n} + \frac{\sqrt{d}}{\gamma^3 n\epsilon}.$$

Finally, we verify the width condition in Theorem 55. Since $T/B = 1$ and $\eta T A_{\mathrm{van}} \lesssim \frac{\eta}{\gamma^2} + \frac{\eta\sqrt{d}}{\gamma^3\epsilon} \lesssim \frac{1}{\gamma^2}$, we have $\eta^2\big(T^2 A_{\mathrm{van}}^2 + \frac{T}{B} + 1\big) \lesssim \frac{1}{\gamma^4}$. Hence the required width condition is satisfied by $m \asymp \gamma^{-4}\mathrm{polylog}(n/\delta)$. This completes the proof.

G Proofs for Mini-batch SGD

We now specialize the preceding analysis to non-private mini-batch SGD by setting the noise level $\kappa = 0$. For the representative rates below, we take $C_{\mathrm{clip}} = G_\delta$, so that clipping is inactive.

Theorem 57 (Optimization, generalization and population risk bounds)**.** *Let $\delta \in (0,1)$. Suppose Assumptions 1 and 2 hold. Let $\{\mathbf{W}_t\}_{t=0}^T$ be generated by mini-batch SGD with $\eta \le 1/(12C_{\sigma,b}p^3)$ and $C_{\mathrm{clip}} = G_\delta$. Assume $m \gtrsim \frac{\log(m/\delta)(\log^4(T)+\log^2(n/\delta))}{\gamma^4}$. Then, with probability at least $1-\delta$ over the initialization and the algorithmic randomness,*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathcal{L}_S(\mathbf{W}_t) \lesssim \frac{1}{\gamma^2\eta T} + \frac{1}{\gamma\sqrt{BT}} + \frac{\eta}{B} =: A_{\mathrm{non}}.$$

For the generalization and population risk statements below, assume further the sample-separability condition of Theorem 55. Further assume $m \gtrsim \log(m/\delta)\eta^2\big(T^2A_{\mathrm{non}}^2+\log(n/\delta)\log^2(1/\delta)\big(\frac{T}{B}+1\big)\big)$. *Then, with probability at least* $1-\delta$ *over the initialization,*

$$\mathbb{E}_{S,\mathcal{A}}\big[\mathcal{L}(\mathbf{W}_T)-\mathcal{L}_S(\mathbf{W}_T)\big] \lesssim \frac{\eta}{n}\sum_{t=0}^{T-1}\mathbb{E}_{S,\mathcal{A}}\big[\mathcal{L}_S(\mathbf{W}_t)\big]+\frac{\delta}{\gamma},$$

and

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_{S,\mathcal{A}}\big[\mathcal{L}(\mathbf{W}_t)\big] \lesssim \Big(1+\frac{\eta T}{n}\Big)\big(A_{\mathrm{non}}+\delta\big)+\frac{\delta}{\gamma}.$$

Proof. The non-private mini-batch SGD bounds are obtained as direct special cases of the standard DP-SGD analysis by setting $\kappa = 0$, including the sample-separability condition required for the generalization and population risk statements. All privacy-noise terms then disappear, and the corresponding optimization, generalization, and population risk bounds follow immediately. □

Corollary 58 (Risk rates under representative parameter regime)**.** *Suppose Assumptions 1 and 2 hold. Let* $\{\mathbf{W}_t\}_{t=0}^{T-1}$ *be generated by mini-batch SGD with* $C_{\mathrm{clip}} = G_\delta$. *Assume* $\delta \le 1/(\gamma n)$ *and* $m \gtrsim \frac{\mathrm{polylog}(n/\delta)}{\gamma^4}$. *Set* $B \lesssim \gamma\sqrt{n}$, $\eta \asymp \frac{B}{n}$, *and* $T \gtrsim \frac{n^2}{\gamma^2 B}$. *Then, with probability at least* $1-\delta$ *over the initialization and the algorithmic randomness,*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathcal{L}_S(\mathbf{W}_t) \lesssim \frac{1}{n}.$$

For the population risk statement, assume further that the sample-separability condition in Theorem 55 holds. If $T \asymp n^2/(\gamma^2 B)$, *then with probability at least* $1-\delta$ *over the initialization,*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_{S,\mathcal{A}}\big[\mathcal{L}(\mathbf{W}_t)\big] \lesssim \frac{1}{\gamma^2 n}.$$

Proof. Set $B \lesssim \gamma\sqrt{n}$, $\eta \asymp B/n$, and $T \gtrsim n^2/(\gamma^2 B)$. Under these choices, $\frac{\eta}{B} \asymp \frac{1}{n}$. Moreover, $\frac{1}{\gamma^2\eta T} \asymp \frac{n}{\gamma^2 BT} \lesssim \frac{1}{n}$, and $\frac{1}{\gamma\sqrt{BT}} \lesssim \frac{1}{\gamma\sqrt{B\cdot n^2/(\gamma^2 B)}} = \frac{1}{n}$. Hence Theorem 57 gives $A_{\mathrm{non}} \lesssim \frac{1}{n}$ and therefore

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathcal{L}_S(\mathbf{W}_t) \lesssim \frac{1}{n}$$

with probability at least $1-\delta$ over the initialization and the algorithmic randomness.

Now further take $T \asymp n^2/(\gamma^2 B)$. Then

$$\sum_{t=0}^{T-1}\mathbb{E}_{S,\mathcal{A}}\big[\mathcal{L}_S(\mathbf{W}_t)\big] \lesssim \frac{T}{n}.$$

By Theorem 57,

$$\mathbb{E}_{S,\mathcal{A}}\big[\mathcal{L}(\mathbf{W}_T)-\mathcal{L}_S(\mathbf{W}_T)\big] \lesssim \frac{\eta}{n}\sum_{t=0}^{T-1}\mathbb{E}_{S,\mathcal{A}}\big[\mathcal{L}_S(\mathbf{W}_t)\big]+\frac{\delta}{\gamma}.$$

Since $\eta T \asymp \frac{B}{n}\cdot\frac{n^2}{\gamma^2 B} = \frac{n}{\gamma^2}$, and $\delta \le 1/(\gamma n)$, we obtain

$$\mathbb{E}_{S,\mathcal{A}}\big[\mathcal{L}(\mathbf{W}_T)-\mathcal{L}_S(\mathbf{W}_T)\big] \lesssim \frac{1}{\gamma^2 n}.$$

For the averaged population risk, Theorem 57 gives

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_{S,\mathcal{A}}\big[\mathcal{L}(\mathbf{W}_t)\big] \lesssim \Big(1+\frac{\eta T}{n}\Big)\big(A_{\mathrm{non}}+\delta\big)+\frac{\delta}{\gamma}.$$

Using $A_{\mathrm{non}} \lesssim \frac{1}{n}$, $\frac{\eta T}{n} \asymp \frac{1}{\gamma^2}$, and $\delta \le \frac{1}{\gamma n}$, we obtain

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}_{S,\mathcal{A}}\big[\mathcal{L}(\mathbf{W}_t)\big] \lesssim \frac{1}{\gamma^2 n}.$$

Finally, we verify the additional width condition in Theorem 57. Since $A_{\mathrm{non}} \lesssim 1/n$ and $T \asymp n^2/(\gamma^2 B)$, $\eta^2 T^2 A_{\mathrm{non}}^2 \lesssim \frac{1}{\gamma^4}$, while $\eta^2\Big(\frac{T}{B}+1\Big) \lesssim \frac{1}{\gamma^4}$. Hence $m \gtrsim \gamma^{-4}\mathrm{polylog}(n/\delta)$ is sufficient. This completes the proof. □

Recovery of full-batch GD. Take $m \gtrsim \frac{\mathrm{polylog}(n/\delta)}{\gamma^4}$, $B = n$, $\eta \asymp 1$, and $C_{\mathrm{clip}} = G_\delta$. Since $B = n$, the mini-batch gradient coincides with the full empirical gradient, so $\Delta_t = 0$ for all t (see equation 21). Hence the sampling fluctuation terms disappear from the underlying recursion, and the bound reduces to

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathcal{L}_S(\mathbf{W}_t) \lesssim \frac{1}{\gamma^2 T}$$

with probability at least $1-\delta$ over the initialization.

Now set $T \asymp \frac{n}{\gamma^2}$. Then $\frac{1}{T}\sum_{t=0}^{T-1} \mathcal{L}_S(\mathbf{W}_t) \lesssim \frac{1}{n}$. For the generalization and population risk statements below, assume further the sample-separability condition of Theorem 57. Then

$$\sum_{t=0}^{T-1} \mathbb{E}_{S,\mathcal{A}}\big[\mathcal{L}_S(\mathbf{W}_t)\big] \lesssim \frac{T}{n} \asymp \frac{1}{\gamma^2}.$$

Therefore, by the generalization bound in Theorem 57,

$$\mathbb{E}_{S,\mathcal{A}}\big[\mathcal{L}(\mathbf{W}_T) - \mathcal{L}_S(\mathbf{W}_T)\big] \lesssim \frac{1}{n}\sum_{t=0}^{T-1} \mathbb{E}_{S,\mathcal{A}}\big[\mathcal{L}_S(\mathbf{W}_t)\big] + \frac{\delta}{\gamma} \lesssim \frac{1}{\gamma^2 n},$$

where we used $\delta \le 1/(\gamma n)$.

Finally, the averaged population-risk bound in Theorem 57 yields

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}_{S,\mathcal{A}}\big[\mathcal{L}(\mathbf{W}_t)\big] \lesssim \Big(1+\frac{T}{n}\Big)\Big(\frac{1}{n}+\delta\Big) + \frac{\delta}{\gamma} \lesssim \frac{1}{\gamma^2 n}.$$

The corresponding width condition is satisfied since $T^2(\frac{1}{n})^2 \asymp \frac{1}{\gamma^4}$, and therefore $m \gtrsim \gamma^{-4}\mathrm{polylog}(n/\delta)$ is sufficient. This completes the proof.

H Further Experiments

H.1 Fixed-Privacy Experiments

Figure 3 (Left) follows the experimental setup of Kalinin et al. (2026). The main modification is that we replace the CNN used in their experiments with the two-layer KAN considered in this paper, while studying the effect of temporal correlation under a fixed privacy budget. We additionally conduct the same experiment on synthetic logistic data generated following the setup of Wang et al. (2026). Unlike the fixed-noise experiments in Section 4, these experiments evaluate held-out predictive performance under a common privacy budget, and thus provide an empirical population-level comparison across different correlation levels.

H.2 Further Results for Section 4

Held-out accuracy for $\eta = 1$. We first report the held-out accuracy corresponding to the $\eta = 1$ experiments in Figure 2. Although Theorem 1 directly controls the empirical loss rather than the prediction accuracy, Figure 4 shows that the reduction in training loss is also accompanied by

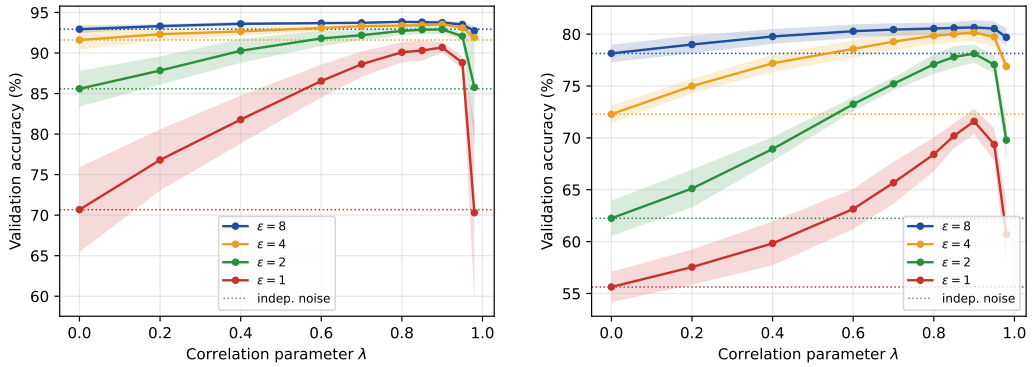
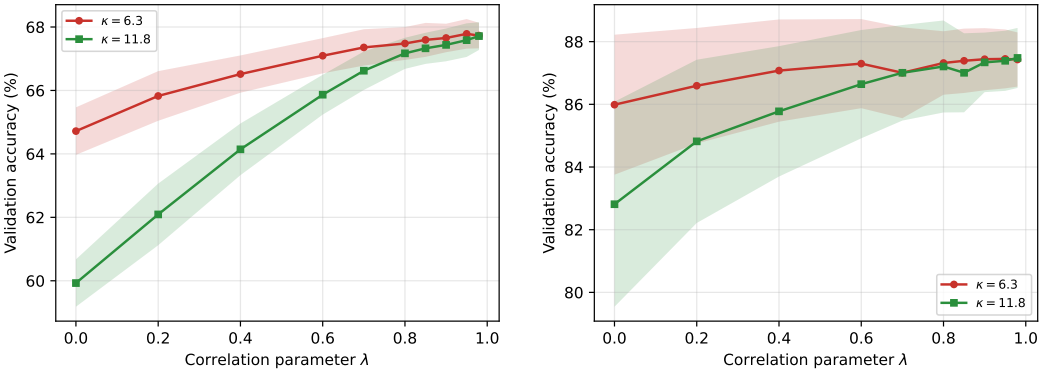


Figure 3: Effect of temporal correlation on two-layer KANs trained under a fixed privacy budget on MNIST and synthetic logistic data. The noise multiplier is recalibrated for each value of the correlation parameter λ to maintain the same privacy budget.



(a) Synthetic logistic data.

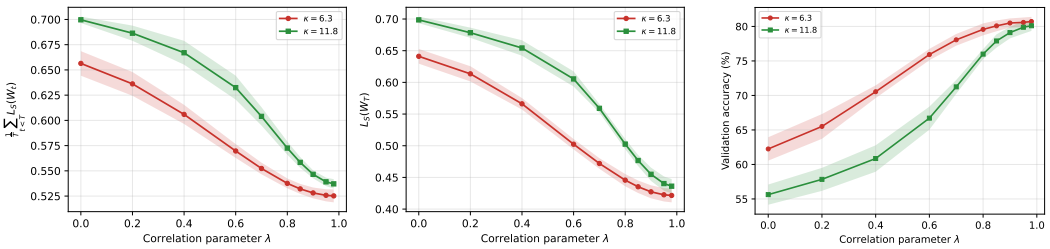
(b) MNIST.

Figure 4: Held-out accuracy for the fixed-noise-multiplier experiments with $\eta = 1$. The noise multiplier κ is held fixed while the correlation parameter λ varies. Curves show the mean over five random seeds, with shaded regions indicating one standard deviation.

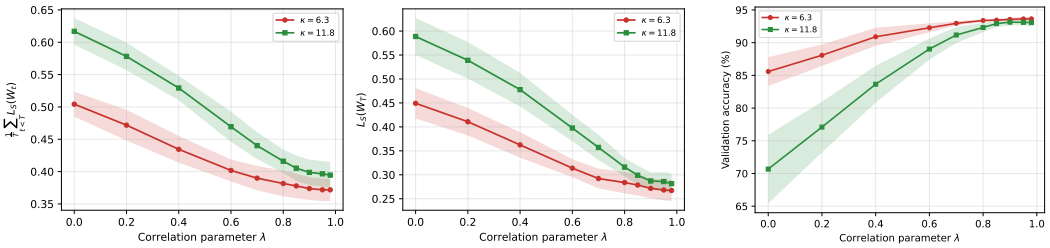
improved predictive performance. On both synthetic logistic data and MNIST 3 vs. 8, the held-out accuracy generally increases as λ increases, with the improvement being more pronounced at the larger noise multiplier $\kappa = 11.8$. Small fluctuations under strong correlation are within the variation across random seeds. Thus, the improvement observed in the training objective is also reflected in held-out predictive performance.

Results with a larger nominal step size. We further repeat the fixed-noise-multiplier experiments with the larger nominal step size $\eta = 5$. We emphasize that Theorem 1 constrains the effective step size $\eta_{\text{eff}} = \rho_{\text{clip}}\eta$, rather than the nominal step size η itself. Hence, choosing $\eta = 5$ is compatible with the theorem whenever the corresponding η_{eff} satisfies the stated condition.

Figure 5 shows that the same qualitative behavior persists. For both datasets and both noise multipliers, increasing temporal correlation reduces the trajectory-averaged empirical loss, and the final-iterate loss follows the same trend. Compared with $\eta = 1$, the dependence on λ is more pronounced, particularly at the larger noise multiplier. The corresponding held-out accuracies exhibit a consistent improvement. These results provide additional evidence that the correlation-dependent reduction observed in the main text persists across different nominal step sizes.



(a) Synthetic logistic data.



(b) MNIST.

Figure 5: Additional fixed-noise-multiplier experiments with the larger nominal step size $\eta = 5$. For each dataset, the first two panels report the trajectory-averaged empirical loss and the empirical loss of the final iterate, respectively, while the third panel reports the held-out accuracy. The noise multiplier κ is held fixed while the correlation parameter λ varies. Curves show the mean over five random seeds, with shaded regions indicating one standard deviation.