Better Rates for Private Linear Regression in the Proportional Regime via Aggressive Clipping

Simone Bombari SIMONE.BOMBARI@IST.AC.AT
Institute of Science and Technology Austria

Inbar Seroussi INBARSER@TAUEX.TAU.AC.IL
Tel Aviv University

Marco Mondelli MARCO.MONDELLI@IST.AC.AT
Institute of Science and Technology Austria

Abstract

Differentially private (DP) linear regression has received significant attention in the recent theoretical literature, with several works aimed at obtaining improved error rates. A common approach is to set the clipping constant much larger than the expected norm of the per-sample gradients. While simplifying the analysis, this is however in sharp contrast with what empirical evidence suggests to optimize performance. Our work bridges this gap between theory and practice: we provide sharper rates for DP stochastic gradient descent (DP-SGD) by crucially operating in a regime where clipping happens frequently. Specifically, we consider the setting where the data is multivariate Gaussian, the number of training samples n is proportional to the input dimension d, and the algorithm guarantees constant-order zero concentrated DP. Our method relies on establishing a deterministic equivalent for the trajectory of DP-SGD in terms of a family of ordinary differential equations (ODEs). As a consequence, the risk of DP-SGD is bounded between two ODEs, with upper and lower bounds matching for isotropic data. By studying these ODEs when n/d is large enough, we demonstrate the optimality of aggressive clipping, and we uncover the benefits of decaying learning rate and private noise scheduling.

1. Introduction

Differential privacy (DP) [22] has consolidated as the standard framework for privacy guarantees and data protection in machine learning. This has motivated an extensive research effort in problems spanning from fundamental questions in optimization [5, 6, 8, 13, 14] to deploying DP in large scale deep learning architectures [19, 36, 42]. When there is no prior information on the Lipschitz constant of the objective, training with DP differentiates from standard stochastic gradient descent (SGD) methods due to additional algorithmic steps: *clipping* the per-sample gradients and adding white *noise* to the parameters updates, depending on the desired privacy requirement and the number of training iterations [1]. Carefully defining the hyperparameters of DP-SGD plays a crucial role in the maximization of its performance, and a principled understanding of their impact in different settings is fundamental to reduce costly grid searches and the related privacy leakage [49].

In particular, the problem of *DP linear regression* presents the challenge described above, in a setting amenable to a precise theoretical analysis [12, 38, 46, 59]. Recent work providing efficient algorithms to improve the theoretical guarantees on the test risk [9, 39, 57] has a common pattern: the *clipping constant* C_{clip} is set to be sufficiently large so that, with high probability, gradient clipping *does not take place* throughout the dynamics of the algorithm. This approach brings the benefit of a simpler analysis, as the optimization becomes more easily comparable to a quadratic problem with noisy gradient updates. However, the practical advantage of avoiding clipping remains

© S. Bombari, I. Seroussi & M. Mondelli.

unclear: it is pointed out in [9] that the lowest error occurs under *significant clipping*, and this last conclusion is in agreement with experimental evidence for DP-SGD in deep learning [19, 35, 36], which supports setting C_{clip} sufficiently small, rescaling appropriately the learning rate.

In this work, we show that, in DP linear regression, DP-SGD achieves better rates in the regime where C_{clip} is of the same order of the typical per-sample gradients and, thus, clipping occurs frequently. In particular, we focus on the setting where the number of training data n is assumed to grow proportionally with the number of input dimensions d, with constant-order guarantees in terms of zero concentrated DP (see Definition 1). Notably, in this regime, the upper bounds provided by prior work on DP linear regression [9, 39, 57] diverge (see Appendix A for details), and a crucial reason behind this is the extra logarithmic factors in C_{clip}, which in turn define a regime where clipping does not happen with high probability. An exception to this is given by the recent work [23], which provides constant-order test risk guarantees for the implicit solution of minimization problems with output and objective perturbation. However, the approach of [23] is restricted to isotropic data covariance, it provides limited insight on DP-SGD (the results are in terms of dynamical mean-field theory equations [26, 28], which are then hard to interpret), and it does not characterize how hyper-parameters affect utility for a fixed privacy budget. Our contributions are summarized below.

1. We consider a one-pass DP-SGD algorithm (Algorithm 1) and provide privacy guarantees in terms of zero concentrated DP (Proposition 2). Our argument is based on privacy amplification by iteration [24, 25], and the privacy guarantees only regard the final output of the algorithm.

2. Following recent progress in high-dimensional optimization [17, 41, 51], we track the test risk of DP-SGD via a stochastic differential equation (SDE), dubbed *homogenized DP-SGD* (Theorem 4). This is turn provides a deterministic equivalent for the DP-SGD trajectory in terms of a family of ordinary differential equations (ODEs), and we exploit the equivalence by bounding the test risk between two ODEs, with upper and lower bounds matching for isotropic data.

3. Finally, we give sharp bounds on the ODEs above for polynomially decaying learning rate schedules, in the setting where $d/n \to \gamma$ is sufficiently small (Theorems 6 and 7). This allows us to *(i)* demonstrate the optimality of *aggressive clipping*, i.e., when C_{clip} is set to be of the same size of (or even much smaller than) the expected norm of the per-sample gradients, and to *(ii)* compare the utility of different schedules, proving the benefits of fast-decaying learning rates.

Our analysis is fueled by a methodology that is both innovative compared to earlier work [9, 39, 57] and rather general, thus laying foundations for the sharp characterization of high-dimensional DP optimization, beyond linear regression. We discuss directions for future work in Appendix F.

2. Preliminaries

Notation. Given a vector v, we denote by $\|v\|_2$ its Euclidean norm. Given a matrix A, we denote by $\mathrm{tr}(A)$ and $\|A\|_{\text{op}}$ its trace and operator (spectral) norm. Given a symmetric matrix A, we denote by $\lambda_{\min}(A)$ ($\lambda_{\max}(A)$) its smallest (largest) eigenvalue. All complexity notations $\Omega(\cdot)$, $O(\cdot)$, $\omega(\cdot)$, $o(\cdot)$ and $\Theta(\cdot)$ are understood for large data size n and input dimension d. We indicate with $C > 0$ a numerical constant independent of n and d, whose value may change from line to line, and we say that an event holds with overwhelming probability if it holds with probability at least $1 - e^{-\omega(\log d)}$.

Algorithm 1: DP-SGD

Input: Training data (X, Y), learning rate schedule $\{\eta_k\}_{k=1}^n$, clipping constant C_{clip}, noise multiplier schedule $\{\sigma_k\}_{k=1}^n$, initialization $\theta_0 = 0$.

for $k \in \{1, \dots, n\}$ **do**

Compute the gradient $g_k = \nabla_\theta \left(x_k^\top \theta_{k-1} - y_k\right)^2 / 2 = x_k \left(x_k^\top \theta_{k-1} - y_k\right)$.

Clip the gradient $\bar{g}_k = g_k \min\left(1, \frac{C_{\text{clip}}}{\|g_k\|_2}\right)$.

Set the learning rate adaptively $\bar{\eta}_k = \min\left(\eta_k, \frac{2}{\|x_k\|_2^2}\right)$.

Sample independent Gaussian noise $b_k \sim \mathcal{N}(0, I)$.

Update the model parameters $\theta_k = \theta_{k-1} - \bar{\eta}_k \bar{g}_k + 2C_{\text{clip}}\sigma_k b_k$.

Output: Model parameters $\theta^p = \theta_n$.

Linear regression. Let (X, Y) be a labeled training dataset, with $X = [x_1, \dots, x_n]^\top \in \mathbb{R}^{n \times d}$ and $Y = [y_1, \dots, y_n]^\top \in \mathbb{R}^n$ s.t. input-label pairs are i.i.d. from a joint distribution P_{XY}. We consider a *linear regression* model: $y_i = x_i^\top \theta^* + z_i$, where $\theta^* \in \mathbb{R}^d$, x_i has mean-0 and covariance Σ, and z_i is independent label noise with mean-0 and variance ζ^2. The goal of DP linear regression is to find θ^p guaranteeing a required privacy budget and minimizing the *test risk*: $\mathcal{P}(\theta^p) = \mathbb{E}_{(x,y)\sim P_{XY}}[\left(x^\top \theta^p - y\right)^2]/2 = \|\Sigma^{1/2}\left(\theta^p - \theta^*\right)\|_2^2/2 + \zeta^2/2$. We also use the notation $\mathcal{R}(\theta^p) = \mathcal{P}(\theta^p) - \zeta^2/2$ to denote the noiseless test risk s.t. $\mathcal{R}(\theta^*) = 0$.

Differential privacy (DP). We recall that a dataset D' is adjacent to a dataset D if they differ by only one sample. In this work, we frame privacy in terms of zero-concentrated DP (zCDP).

Definition 1 (Zero concentrated DP [11]) *Given $\alpha \in (1, +\infty)$ and two random variables X and X' with laws p_X and $p_{X'}$, their α-Rényi Divergence [53] is defined as*

$$D_\alpha\left(X \,\|\, X'\right) = \frac{1}{\alpha - 1} \ln \int \left(\frac{p_X(\theta)}{p_{X'}(\theta)}\right)^\alpha p_{X'}(\theta) \mathrm{d}\theta. \tag{1}$$

Then, a randomized algorithm $\mathcal{A}$ satisfies ρ-zero concentrated DP (ρ-zCDP) if, for any pair of adjacent datasets D, D' and any $\alpha \in (1, +\infty)$, we have $D_\alpha\left(\mathcal{A}(D) \,\|\, \mathcal{A}(D')\right) \leq \alpha\rho$.

Guarantees for zCDP can be translated to other formulations, such as (ε, δ)-DP, see Appendix B.

3. Homogenized DP-SGD and deterministic equivalent

We consider DP-SGD performing a single pass on the n training samples (Algorithm 1).

Proposition 2 *Algorithm 1 satisfies $(\rho^2/2)$-zCDP, where $\rho = \max_{k\in[n]} \eta_k / \sqrt{\sum_{j=k}^n \sigma_j^2}$.*

Proposition 2 (whose proof is deferred to Appendix B) states that each sample x_k is "protected" by the overall noise introduced in the following updates $\sum_{j=k}^n \sigma_j^2$. For an assigned privacy guarantee, we can minimize the noise introduced by the algorithm $\sum_{j=1}^n \sigma_j^2$ (and, therefore, optimize its performance) via the schedule below:

$$\eta_k = \rho \sqrt{\sum_{j=k}^n \sigma_j^2}, \quad \text{or, equivalently,} \quad \rho^2 \sigma_k^2 = \begin{cases} \eta_k^2 - \eta_{k+1}^2, & k \in \{1, \dots, n-1\}, \\ \eta_k^2, & k = n. \end{cases} \tag{2}$$

Homogenized DP-SGD. Our analysis is based on tracking the risk $\mathcal{R}$ of Algorithm 1 via an SDE in parameter space. This approach was developed in a series of works [17, 41, 50, 51] aimed at characterizing high-dimensional optimization problems and the implicit bias of stochastic batching in regression tasks. We now make two assumptions on data distribution and hyper-parameter scaling.

Assumption 1 (Data distribution) *$\{x_i\}_{i=1}^n$ are n i.i.d. samples from the multivariate, mean-0, Gaussian distribution $\mathcal{P}_X$, with covariance $\Sigma := \mathbb{E}\left[xx^\top\right] \in \mathbb{R}^{d\times d}$. Furthermore, the noise z_i is mean-0, Gaussian, with variance $\zeta^2 > 0$, and $\|\theta^*\|_2 = \Theta(1)$. We also assume that $\mathrm{tr}(\Sigma) = d$ and $\kappa = \lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma) = \Theta(1)$, i.e., the data covariance is well-conditioned.*

Assumption 2 (Hyper-parameter scaling) *Let the clipping constant in Algorithm 1 be $C_{\text{clip}} = c\sqrt{d}$, where c is a constant independent of n and d. Furthermore, the learning rate schedule is given by $\eta_k = \frac{\tilde{\eta}(k/n)}{n}$, where $\tilde{\eta} : [0,1] \to \mathbb{R}$ is a function such that both $\tilde{\eta}^2(\cdot)$ and the absolute value of its first and second derivatives are uniformly bounded by a constant independent of n and d.*

Definition 3 (Homogenized DP-SGD) *For any $t \in [0,1)$, we define the* homogenized DP-SGD *(H-DP-SGD) as the solution of the SDE*

$$\mathrm{d}\Theta_t = -\tilde{\eta}(t)\mu_c(\Theta_t)\nabla\mathcal{P}(\Theta_t)\mathrm{d}t + \tilde{\eta}(t)\sqrt{\frac{2\nu_c(\Theta_t)\mathcal{P}(\Theta_t)\Sigma}{n}}\mathrm{d}B_t^s + 2\frac{\sqrt{d}}{n}c\tilde{\sigma}(t)\mathrm{d}B_t^p, \tag{3}$$

where $\Theta_0 = \theta_0 = 0$, B_t^s and B_t^p are two independent standard Brownian motions in $\mathbb{R}^d$, $\tilde{\eta}(t)$ is defined as in Assumption 2, $\tilde{\sigma}(t)$ is such that $\rho^2\tilde{\sigma}^2(t) = -\mathrm{d}\tilde{\eta}^2(t)/\mathrm{d}t$, and $\mu_c(\theta)$, $\nu_c(\theta)$ are the descent *and the* variance reduction factor *respectively, defined in* (17) *in Appendix C.*

Theorem 4 *Let Assumptions 1 and 2 hold. Let $\rho = \Theta(1)$, $n, d \to \infty$ s.t. $d/n \to \gamma \in (0,\infty)$, and $\sup_{t\in[0,1]}\tilde{\eta}(t) < 2/\gamma$. Denote by Θ_t and θ_k independent realizations of H-DP-SGD (as per Definition 3) and Algorithm 1. Then, with overwhelming probability, we have*

$$\sup_{t\in[0,1)}\left|\mathcal{R}(\Theta_t) - \mathcal{R}(\theta_{\lfloor tn\rfloor})\right| = O\left(\frac{\log^2 n}{\sqrt{n}}\right). \tag{4}$$

Theorem 4 formalizes the equivalence in terms of risk between Algorithm 1 and the H-DP-SGD dynamics in (3), and its proof is deferred to Appendix D.1.

Deterministic equivalent. The SDE in (3) can be well approximated in terms of d coupled ODEs (see Lemma 15 in Appendix D.1), and a similar approximation was pursued by earlier work on SGD without private noise [17]. This system of ODEs then provides a deterministic equivalent for the dynamics of DP-SGD, since it approximates sharply its risk without depending on the stochasticity of the private noise and of the data. Given the difficulty of handling d coupled ODEs, for general covariance we opt to give an upper and a lower bounds in terms of two decoupled ODEs, namely

$$\begin{aligned}\mathrm{d}\overline{R}(t) &= -2\lambda_{\min}\tilde{\eta}(t)\mu_c(\overline{R})\overline{R}\mathrm{d}t + \lambda_{\max}\tilde{\eta}^2(t)\nu_c(\overline{R})(\overline{R} + \zeta^2/2)\gamma\mathrm{d}t + 2c^2\tilde{\sigma}^2(t)\gamma^2\mathrm{d}t,\\ \mathrm{d}\underline{R}(t) &= -2\lambda_{\max}\tilde{\eta}(t)\mu_c(\underline{R})\underline{R}\mathrm{d}t + \tilde{\eta}^2(t)\nu_c(\underline{R})(\underline{R} + \zeta^2/2)\gamma\mathrm{d}t + 2c^2\tilde{\sigma}^2(t)\gamma^2\mathrm{d}t,\end{aligned} \tag{5}$$

where, importantly, the upper bound $\overline{R}$ and the lower bound $\underline{R}$ match when the data is isotropic ($\Sigma = I$). A formal statement is in Proposition 13 (deferred to Appendix D) and Figure 2 (also in Appendix D) shows that the convergence is already evident at moderate values of n, d.

Noise in the last iteration. Theorem 4 holds for $\{\theta_k\}_{k=1}^{n-1}$, as the supremum in (4) is taken on the open interval $t \in [0, 1)$. The last iterate $\theta_n = \theta^p$, which corresponds to $t = 1$ and gives the (private) output of the algorithm, is treated separately via the result below.

Proposition 5 *Let Assumptions 1 and 2 hold. Let $\rho = \Theta(1)$ and $n, d \to \infty$ s.t. $d/n \to \gamma \in (0, \infty)$. Then, with overwhelming probability, $\left|\mathcal{R}(\theta^p) - \mathcal{R}(\theta_{n-1}) - 2c^2\tilde{\eta}^2(1)\gamma^2/\rho^2\right| = O\left(\log n/\sqrt{n}\right)$.*

Better rates in the proportional regime. Combining Theorem 4, Eq. (5), and Proposition 5 gives that, when $d/n \to \gamma$ with $n, d \to \infty$, $\mathcal{R}(\theta^p)$ is bounded between $\underline{R}(1) + 2c^2\tilde{\eta}^2(1)\gamma^2/\rho^2$ and $\overline{R}(1) + 2c^2\tilde{\eta}^2(1)\gamma^2/\rho^2$. This is a constant-order upper bound on the risk in the proportional regime, which improves upon prior work on DP linear regression [9, 39, 57], due to the additional log terms implicit in their notation $\tilde{O}(\cdot)$, as discussed in Appendix A.

4. Polynomial schedules and the benefits of aggressive clipping

Our analysis not only provides better rates than earlier work, but it is also sharp enough to *(i)* establish optimal hyper-parameter choices, and to *(ii)* compare different learning rate schedules. To do so, we focus on the case $\gamma = d/n \to 0$. This limit is taken *after* the limit $d, n \to \infty$, which means that d and n are incomparably larger than $1/\gamma$. Thus, our bounds on $\mathcal{R}(\theta^p)$ neglect the smaller terms that vanish as $d, n \to \infty$. We use the asymptotic notation $\Omega_\gamma(\cdot), O_\gamma(\cdot), \Theta_\gamma(\cdot), \omega_\gamma(\cdot), o_\gamma(\cdot)$ intended for small enough γ, and the quantities ρ and α are allowed to depend on γ (but not on n, d). Below, we provide informal statements where we assume $\lambda_{\max} = \Theta_\gamma(1)$ and $\lambda_{\min} = \Theta_\gamma(1)$. Formal statements that also tracks the dependence on $\lambda_{\max}, \lambda_{\min}$, together with the corresponding proofs, are given in Appendix E. We also assume $\|\Sigma^{1/2}\theta^*\|_2 = \Theta_\gamma(1)$, $\zeta^2 = \Theta_\gamma(1)$, i.e., the test risk of the model at initialization and the label noise variance are fixed strictly positive constants.

We study the ODEs in (42) for a family of polynomially decaying learning rate schedules:

$$\tilde{\eta}(t) = \tilde{\eta}(0)(1-t)^\alpha, \tag{6}$$

for $\alpha = 0$, $\alpha = 1/2$, and $\alpha \geq 1$. The case $\alpha = 0$ corresponds to output perturbation: the learning rate is fixed, and the private noise is added only at the end of the algorithm, as (2) implies $\sigma_k = 0$ for $k \in \{1, \dots, n-1\}$ and $\sigma_n = \eta_1/\rho$. The case $\alpha = 1/2$ corresponds to a linearly decaying $\tilde{\eta}^2(t)$, which in turn gives a constant level of noise $\sigma_k = \eta_1/(\sqrt{n}\rho)$ in the iterations of DP-SGD.

Theorem 6 (Informal) *Let θ_0^p and $\theta_{1/2}^p$ be the solutions obtained with Algorithm 1 with $\tilde{\eta}(t)$ given by (6) for $\alpha = 0$ and $\alpha = 1/2$ respectively, in the setting $\gamma = o_\gamma(1)$. Pick $c = O_\gamma(1)$, $\tilde{\eta}(0)c = C\ln(1/\gamma)$, $\tilde{\eta}(0) \leq 2/\gamma$, for a large enough constant C which does not depend on γ, ρ, Σ. Then, under some technical assumptions, we have that, with overwhelming probability,*

$$\mathcal{R}(\theta_0^p) = O_\gamma\left(\gamma\ln(1/\gamma) + \frac{\gamma^2\ln^2(1/\gamma)}{\rho^2}\right), \quad \mathcal{R}(\theta_{1/2}^p) = O_\gamma\left(\gamma\ln^{2/3}(1/\gamma) + \frac{\gamma^2\ln^{4/3}(1/\gamma)}{\rho^2}\right).$$

Furthermore, for any c and $\tilde{\eta}(0)$ s.t. $\tilde{\eta}(0) < 2/\gamma$, a matching lower bound holds (up to a universal multiplicative constant).

The result has two remarkable consequences: *(i)* the proposed hyper-parameters are optimal (assuming $\tilde{\eta}(0) < 2/\gamma$), and *(ii)* DP-SGD outperforms output perturbation. In Appendix E, we comment on the hyper-parameters and support our conclusions via the simulations of Figure 3. Our analysis also

shows the *benefits of aggressive clipping*: the lower bound can be increased by a factor $\max\left(1, c^2\right)$, which demonstrates the sub-optimality of $c = \omega_\gamma(1)$ (corresponding to infrequent clipping), see the end of Appendix E.3 for details.

Finally, as $\alpha = 1/2$ improves over $\alpha = 0$, we reduce the noise at the end of training and pick $\tilde{\sigma}^2(t)$ proportional to $2\alpha(1-t)^{2\alpha-1}$ for $\alpha \geq 1$, corresponding to a decay at least linear.

Theorem 7 (informal) *Let θ_α^p be the solution obtained with Algorithm 1, with $\tilde{\eta}(t)$ given by (6) for $\alpha \geq 1$, in the setting $\gamma = o_\gamma(1)$. Pick $c = O_\gamma(1)$, $\tilde{\eta}(0)c = C\alpha \ln(1/\gamma)$, $\tilde{\eta}(0) \leq 2/\gamma$, for a large enough constant C which does not depend on $\gamma, \rho, \Sigma, \alpha$. Then, under some technical assumptions, we have that, with overwhelming probability,*

$$\mathcal{R}(\theta_\alpha^p) = O_\gamma\left(\alpha\gamma \ln^{\frac{1}{1+\alpha}}(1/\gamma) + \frac{\alpha^2\gamma^2 \ln^{\frac{2}{1+\alpha}}(1/\gamma)}{\rho^2}\right).$$

Thus, for small γ, it is convenient to increase α and decay the noise faster during training, up to a level $\alpha = \Theta_\gamma(\ln\ln(1/\gamma))$, which gives a bound of order $\gamma \ln\ln(1/\gamma) + \gamma^2(\ln\ln(1/\gamma))^2/\rho^2$. Different learning rate schedules are also compared numerically in Figure 4 in Appendix E.

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *ACM SIGSAC Conference on Computer and Communications Security*, 2016.

[2] Kareem Amin, Alex Kulesza, Andres Munoz, and Sergei Vassilvtiskii. Bounding user contributions: A bias-variance trade-off in differential privacy. In *International Conference on Machine Learning*, 2019.

[3] Galen Andrew, Om Thakkar, Hugh Brendan McMahan, and Swaroop Ramaswamy. Differentially private learning with adaptive clipping. In *Advances in Neural Information Processing Systems*, 2021.

[4] Peter L. Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *Acta Numerica*, 30:87–201, 2021. doi: 10.1017/S0962492921000027.

[5] Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, 2014.

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

[7] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[8] Simone Bombari and Marco Mondelli. Privacy for free in the overparameterized regime. *Proceedings of the National Academy of Sciences*, 122(15):e2423072122, 2025. doi: 10.1073/pnas.2423072122.

[9] Gavin R Brown, Krishnamurthy Dj Dvijotham, Georgina Evans, Daogao Liu, Adam Smith, and Abhradeep Guha Thakurta. Private gradient descent for linear regression: Tighter error bounds and instance-specific uncertainty estimation. In *International Conference on Machine Learning*, 2024.

[10] Zhiqi Bu, Hua Wang, Zongyu Dai, and Qi Long. On the convergence and calibration of deep learning with differential privacy. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

[11] Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of cryptography conference*, pages 635–658. Springer, 2016.

[12] T. Tony Cai, Yichen Wang, and Linjun Zhang. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *The Annals of Statistics*, 49(5):2825 – 2850, 2021.

[13] Kamalika Chaudhuri and Claire Monteleoni. Privacy-preserving logistic regression. In *Advances in Neural Information Processing Systems*, 2008.

[14] Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(29):1069–1109, 2011.

[15] Xiangyi Chen, Steven Z. Wu, and Mingyi Hong. Understanding gradient clipping in private sgd: A geometric perspective. In *Advances in Neural Information Processing Systems*, 2020.

[16] Chen Cheng and Andrea Montanari. Dimension free ridge regression. *The Annals of Statistics*, 52(6):2879 – 2912, 2024. doi: 10.1214/24-AOS2449.

[17] Elizabeth Collins-Woodfin, Courtney Paquette, Elliot Paquette, and Inbar Seroussi. Hitting the high-dimensional notes: an ode for sgd learning dynamics on glms and multi-index models. *Information and Inference: A Journal of the IMA*, 13(4):iaae028, 2024. ISSN 2049-8772. doi: 10.1093/imaiai/iaae028.

[18] Rudrajit Das, Satyen Kale, Zheng Xu, Tong Zhang, and Sujay Sanghavi. Beyond uniform lipschitz condition in differentially private optimization. In *International Conference on Machine Learning*, 2023.

[19] Soham De, Leonard Berrada, Jamie Hayes, Samuel L. Smith, and Borja Balle. Unlocking high-accuracy differentially private image classification through scale. *arXiv preprint arXiv:2204.13650*, 2022.

[20] Zeyu Deng, Abla Kammoun, and Christos Thrampoulidis. A model of double descent for high-dimensional binary linear classification. *Information and Inference: A Journal of the IMA*, 11(2):435–495, 2022.

[21] Meng Ding, Mingxi Lei, Liyang Zhu, Shaowei Wang, Di Wang, and Jinhui Xu. Revisiting differentially private relu regression. In *Advances in Neural Information Processing Systems*, 2024.

[22] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006*, 2006.

[23] Cynthia Dwork, Pranay Tankala, and Linjun Zhang. Differentially private learning beyond the classical dimensionality regime. *arXiv preprint arXiv:2411.13682*, 2025.

[24] Vitaly Feldman, Ilya Mironov, Kunal Talwar, and Abhradeep Thakurta. Privacy Amplification by Iteration. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 521–532, 2018.

[25] Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2020, 2020. ISBN 9781450369794.

[26] Cedric Gerbelot, Emanuele Troiani, Francesca Mignacco, Florent Krzakala, and Lenka Zdeborova. Rigorous dynamical mean-field theory for stochastic gradient descent methods. *SIAM Journal on Mathematics of Data Science*, 6(2):400–427, 2024.

[27] Aditya Golatkar, Alessandro Achille, Yu-Xiang Wang, Aaron Roth, Michael Kearns, and Stefano Soatto. Mixed differential privacy in computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8376–8386, June 2022.

[28] Qiyang Han. Entrywise dynamics and universality of general first order methods. *arXiv preprint arXiv:2406.19061*, 2024.

[29] Qiyang Han and Xiaocong Xu. The distribution of ridgeless least squares interpolators. *arXiv preprint arXiv:2307.02044*, 2023.

[30] Trevor J. Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of statistics*, 50 2:949–986, 2019.

[31] Hong Hu, Yue M. Lu, and Theodor Misiakiewicz. Asymptotics of random feature regression beyond the linear scaling regime. *arXiv preprint arXiv:2403.08160*, 2024.

[32] Prateek Jain and Abhradeep Guha Thakurta. (near) dimension independent risk bounds for differentially private learning. In *International Conference on Machine Learning*, 2014.

[33] Hui Jin, Pradeep Kr Banerjee, and Guido Montufar. Learning curves for gaussian process regression with power-law priors and targets. In *International Conference on Learning Representations*, 2022.

[34] Daniel Kifer, Adam Smith, and Abhradeep Thakurta. Private convex empirical risk minimization and high-dimensional regression. In *Conference on Learning Theory*, 2012.

[35] Alexey Kurakin, Shuang Song, Steve Chien, Roxana Geambasu, Andreas Terzis, and Abhradeep Thakurta. Toward training at imagenet scale with differential privacy. *arXiv preprint arXiv:2201.12328*, 2022.

[36] Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. In *International Conference on Learning Representations*, 2022.

[37] Licong Lin, Jingfeng Wu, Sham M Kakade, Peter Bartlett, and Jason D Lee. Scaling laws in linear regression: Compute, parameters, and data. In *Advances in Neural Information Processing Systems*, 2024.

[38] Xiyang Liu, Weihao Kong, and Sewoong Oh. Differential privacy and robust statistics in high dimensions. In *Conference on Learning Theory*, 2022.

[39] Xiyang Liu, Prateek Jain, Weihao Kong, Sewoong Oh, and Arun Suggala. Label robust and differentially private linear regression: Computational and statistical efficiency. In *Advances in Neural Information Processing Systems*, 2023.

[40] Neil Mallinar, James B Simon, Amirhesam Abedsoltan, Parthe Pandit, Mikhail Belkin, and Preetum Nakkiran. Benign, tempered, or catastrophic: a taxonomy of overfitting. In *Advances in Neural Information Processing Systems*, 2022.

[41] Noah Marshall, Ke Liang Xiao, Atish Agarwala, and Elliot Paquette. To clip or not to clip: the dynamics of SGD with gradient clipping in high-dimensions. In *The Thirteenth International Conference on Learning Representations*, 2025.

[42] Ryan McKenna, Yangsibo Huang, Amer Sinha, Borja Balle, Zachary Charles, Christopher A. Choquette-Choo, Badih Ghazi, George Kaissis, Ravi Kumar, Ruibo Liu, Da Yu, and Chiyuan Zhang. Scaling laws for differentially private language models. *arXiv preprint arXiv:2501.18914*, 2025.

[43] H. Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. In *International Conference on Learning Representations*, 2018.

[44] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022.

[45] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Generalization error of random feature and kernel methods: Hypercontractivity and kernel matrix concentration. *Applied and Computational Harmonic Analysis*, 59:3–84, 2022. ISSN 1063-5203.

[46] Jason Milionis, Alkis Kalavasis, Dimitris Fotakis, and Stratis Ioannidis. Differentially private regression with unbounded covariates. In *International Conference on Artificial Intelligence and Statistics*, 2022.

[47] Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275, 2017.

[48] Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. The generalization error of max-margin linear classifiers: Benign overfitting and high dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*, 2019.

[49] Nicolas Papernot and Thomas Steinke. Hyperparameter tuning with renyi differential privacy. In *International Conference on Learning Representations*, 2022.

[50] Courtney Paquette, Elliot Paquette, Ben Adlam, and Jeffrey Pennington. Implicit regularization or implicit conditioning? exact risk trajectories of sgd in high dimensions. In *Advances in Neural Information Processing Systems*, 2022.

[51] Courtney Paquette, Elliot Paquette, Ben Adlam, and Jeffrey Pennington. Homogenization of sgd in high-dimensions: exact dynamics and generalization properties. *Mathematical Programming*, 2024. doi: 10.1007/s10107-024-02171-3.

[52] Venkatadheeraj Pichapati, Ananda Theertha Suresh, Felix X. Yu, Sashank J. Reddi, and Sanjiv Kumar. Adaclip: Adaptive clipping for private sgd. *arXiv preprint arXiv:1908.07643*, 2019.

[53] Alfred Rényi. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 547–561, 1961.

[54] 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*, 2013.

[55] Shuang Song, Thomas Steinke, Om Thakkar, and Abhradeep Thakurta. Evading the curse of dimensionality in unconstrained private glms. In *International Conference on Artificial Intelligence and Statistics*, 2021.

[56] Gerald Teschl. *Ordinary differential equations and dynamical systems*. American Mathematical Society, 2012.

[57] Prateek Varshney, Abhradeep Thakurta, and Prateek Jain. (nearly) optimal private linear regression for sub-gaussian data via adaptive clipping. In *Conference on Learning Theory*, 2022.

[58] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Cambridge university press, 2018.

[59] Yu-Xiang Wang. Revisiting differentially private linear regression: optimal and adaptive prediction & estimation in unbounded domain. *arXiv preprint arXiv:1803.02596*, 2018.

Appendix A. Related work

DP optimization. Since its introduction in [22], DP has provided the golden standard in the field of private data analysis, and different methods have been proposed with the purpose of learning with DP, such as objective and output perturbation [13, 14, 34] or different variants of DP-SGD [1, 5, 54]. In the last decade, a popular line of work has theoretically investigated private learning in various settings, such as Lipschitz and bounded optimization [5, 6, 34] or generalized linear models with Lipschitz loss [32, 55]. In DP linear regression, the objective is non-Lipschitz and, hence, the straightforward adaptation of previous results fails to provide tight guarantees, with improvements being achieved via alternative approaches. Specifically, assumptions on the covariates are used in [59]. These hypotheses are lifted in [46], at the expense of requiring $n = \tilde{\Omega}(d^{3/2})$ samples, where the $\sim$ hides logarithmic terms and the privacy budget is assumed of constant order. DP-SGD with adaptive clipping is shown to achieve a sample complexity of $n = \tilde{\Omega}(d)$ in [57]. This is "nearly optimal", in the sense that it matches, up for logarithmic factors, the min-max lower bound in [12]. Similar nearly optimal rates were previously obtained by [38] (however with a computationally inefficient method), and more recently by [39] and [9]. Notably, in the proportional regime $n = \Theta(d)$, if the privacy budget is of constant order ($\varepsilon/\sqrt{\ln(1/\delta)} = \Theta(1)$ in [39, 57], or $\rho = \Theta(1)$ in [9]), then prior bounds on the test risk diverge logarithmically either in n [57] or in the failure probability [9, 39][1]. This barrier has been recently broken by [23], with the limitations mentioned in Section 1.

Gradient clipping. In the context of private optimization with a non-Lipschitz loss, the role of clipping and the magnitude of the corresponding clipping constant C_{clip} has attracted attention due to its nuanced implications: while a small C_{clip} significantly affects the gradients, larger values force the addition of more private noise, suggesting that the choice of C_{clip} induces a *bias-variance trade-off* [2, 3, 9, 18, 43]. Prior work has argued that the bias induced by small clipping constants can prevent convergence [2, 15, 55], which motivates an adaptive selection of C_{clip} based on (private) statistics of the magnitude of the gradients [1, 3, 27, 52]. Recent experimental studies have given evidence that the best performance is achieved with a sufficiently small C_{clip} [19, 35, 36], but it has also been shown that overly-aggressive clipping can be damaging in the context of model calibration [9, 10]. Theoretical insights on the benefits of small clipping constants have been provided in [15, 18], with [15] proving optimization guarantees when the gradients distribution is sufficiently symmetric and [18] considering the setting where the Lipschitz constant of the loss is sample-dependent. Recent work on DP linear regression [9, 39, 57] shares the common feature of setting the (possibly adaptive) C_{clip} a poly-logarithmic factor larger than the expected norm of the per-sample gradient. This ensures that, with high probability, at most a few gradients are clipped during training, and we note that these logarithmic factors are strongly tied to the consequent logarithmic divergence of the test risk guarantees discussed in the previous paragraph. Finally, the recent work [41] provides a precise analysis of a version of clipped-SGD, although it does not focus on privacy.

Learning in high dimensions. The statistical setting where the input dimension (or model size) d scales with the number of training samples n gained popularity due to its power in explaining many empirical phenomena occurring in practice [4, 7, 30]. In this direction, a line of work has characterized the interplay between over-parameterization and generalization for linear models [16, 29], logistic models [20, 48] and random features [31, 44, 45]. Random features have also been

1. More precisely, in [39] the number of samples would not be sufficient to achieve Eq. (4) with high probability.

considered recently to explain the benefits of scale in private optimization [8]. Another line of work has analyzed the behavior of one-pass SGD algorithms in terms of high-dimensional SDEs [17, 41, 50, 51]. By passing from the SDE to a family of ODEs, this strategy gives a deterministic equivalent for the gradient dynamics, which in turn leads to remarkable insights on optimization stability and the role of stochastic batching.

Appendix B. Proofs on differential privacy

The notion of zCDP can be converted in (ε, δ)-DP, which in turn is defined below.

Definition 8 ((ε, δ)-DP [22]) *A randomized algorithm $\mathcal{A}$ satisfies (ε, δ)-differential privacy if for any pair of adjacent datasets D, D', and for any subset of the parameters space $S \subseteq \mathbb{R}^d$, we have*

$$\mathbb{P}\left(\mathcal{A}(D) \in S\right) \leq e^{\varepsilon}\mathbb{P}\left(\mathcal{A}(D') \in S\right) + \delta. \tag{7}$$

For completeness, we also provide the following definition.

Definition 9 (Rényi DP [47]) *Given $\alpha \in (1, +\infty)$ and $\varepsilon \geq 0$, an algorithm $\mathcal{A}$ satisfies (α, ε) Rényi DP if for any pair of adjacent datasets D, D' we have $D_\alpha\left(\mathcal{A}(D) \,\|\, \mathcal{A}(D')\right) \leq \varepsilon$, where $D_\alpha\left(\mathcal{A}(D) \,\|\, \mathcal{A}(D')\right)$ is the Rényi Divergence [53] between the probability distributions induced by the randomness of $\mathcal{A}$,* i.e.,

$$D_\alpha\left(\mathcal{A}(D) \,\|\, \mathcal{A}(D')\right) = \frac{1}{\alpha - 1} \ln \int \left(\frac{p_{\mathcal{A}(D)}(\theta)}{p_{\mathcal{A}(D')}(\theta)}\right)^{\alpha} p_{\mathcal{A}(D')}(\theta) \mathrm{d}\theta. \tag{8}$$

Note that Definitions 1 and 9 imply that an algorithm is ρ-zCDP if, for any $\alpha > 1$, it is also $(\alpha, \rho\alpha)$ Rényi DP. The following proposition allows to translate Rényi DP and zCDP to (ε, δ)-DP.

Proposition 10 (Proposition 1.3 in [11]) *If $\mathcal{A}$ satisfies $\rho^2/2$-zCDP, it also satisfies $\left(\rho^2/2 + \rho\sqrt{2\ln(1/\delta)}, \delta\right)$-DP, for any $\delta \in (0, 1)$.*

Then, if we consider δ such that $\rho \leq \sqrt{\ln(1/\delta)}$, we achieve (ε, δ)-DP if we have

$$\rho^2/2 + \rho\sqrt{2\ln(1/\delta)} \leq 2\rho\sqrt{\ln(1/\delta)} \leq \varepsilon, \tag{9}$$

which means that for algorithms respecting $\rho^2/2$-zCDP, we can replace 2ρ by $\varepsilon/\sqrt{\ln(1/\delta)}$ in the error bounds to evaluate the cost of privacy in terms of (ε, δ)-DP.

B.1. Proof of Proposition 2

Let us define the family of functions $\ell_{k,C_{\text{clip}}}(\cdot) : \mathbb{R} \to \mathbb{R}$, for all $k \in [n]$, such that $\ell_{k,C_{\text{clip}}}(0) = 0$, and

$$\ell'_{k,C_{\text{clip}}}(z) = z \min\left(1, \frac{C_{\text{clip}}}{|z| \, \|x_k\|_2}\right). \tag{10}$$

In words, $\ell_{k,C_{\text{clip}}}(z)$ is a quadratic function, but linearized for sufficiently large values of $|z|$, such that it is $C_{\text{clip}}/\left\|x_k\right\|_2$-Lipschitz. Then, we have

$$\begin{aligned}\bar{g}_k &= g_k \min\left(1, \frac{C_{\text{clip}}}{\|g_k\|_2}\right)\\ &= x_k\left(x_k^\top\theta_{k-1} - y_k\right)\min\left(1, \frac{C_{\text{clip}}}{\|x_k\|_2\left|x_k^\top\theta_{k-1} - y_k\right|}\right)\\ &= x_k \ell'_{k,C_{\text{clip}}}\left(x_k^\top\theta_{k-1} - y_k\right)\\ &= \nabla_\theta \ell_{k,C_{\text{clip}}}\left(x_k^\top\theta_{k-1} - y_k\right),\end{aligned} \tag{11}$$

where the first step follows from the definition of $\bar{g}_k$ in Algorithm 1. Furthermore, we have

$$\begin{aligned}&\left\|\nabla_\theta \ell_{k,C_{\text{clip}}}(x_k^\top\theta - y_k) - \nabla_\theta \ell_{k,C_{\text{clip}}}(x_k^\top\theta' - y_k)\right\|_2\\ &\quad= \|x_k\|_2\left|\ell'_{k,C_{\text{clip}}}(x_k^\top\theta - y_k) - \ell'_{k,C_{\text{clip}}}(x_k^\top\theta' - y_k)\right|\\ &\quad\le \|x_k\|_2\left|x_k^\top\left(\theta - \theta'\right)\right|\\ &\quad\le \|x_k\|_2^2\left\|\theta - \theta'\right\|_2,\end{aligned} \tag{12}$$

where the second step follows from the fact that $\ell'_{k,C_{\text{clip}}}(z)$ is a 1-Lipschitz function. Let us now define

$$\bar{\ell}_{k,C_{\text{clip}}}(z) = \min\left(1, \frac{2}{\|x_k\|_2^2\,\eta_k}\right)\ell_{k,C_{\text{clip}}}(z). \tag{13}$$

Then, we have that every iteration of Algorithm 1 takes the form

$$\theta_k = \theta_{k-1} - \eta_k \nabla_\theta \bar{\ell}_{k,C_{\text{clip}}}(x_k^\top\theta_{k-1} - y_k) + 2C_{\text{clip}}\sigma_k b_k, \tag{14}$$

where $\bar{\ell}_{k,C_{\text{clip}}}(x_k^\top\theta - y_k)$ is C_{clip}-Lipschitz with respect to θ, and it is $2/\eta_k$-smooth, *i.e.*,

$$\left\|\nabla_\theta \bar{\ell}_{k,C_{\text{clip}}}(x_k^\top\theta - y_k) - \nabla_\theta \bar{\ell}_{k,C_{\text{clip}}}(x_k^\top\theta' - y_k)\right\|_2 \le \frac{2}{\eta_k}\left\|\theta - \theta'\right\|_2, \tag{15}$$

due to (12) and (13). Thus, the desired result follows from Theorem 3.1 in [25], after setting their batch sizes $\{B_k\}$ identically equal to 1, and their projection set $\mathcal{K}$ equal to all $\mathbb{R}^d$. ■

Appendix C. The auxiliary functions $\mu_c(\theta)$ and $\nu_c(\theta)$

Let us introduce

$$r(\theta, x, y) = x^\top\theta - y, \qquad r_c(\theta, x, y) = r(\theta, x, y)\min\left(1, \frac{c}{|r(\theta, x, y)|}\right), \tag{16}$$

where $r(\theta, x, y)$ represents the residual in θ, and $r_c(\theta, x, y)$ is a clipped version of it. Then, as done in [41], we define the *descent reduction factor* and the *variance reduction factor*

$$\mu_c(\theta) = \frac{\left\|\mathbb{E}_{(x,y)\sim P_{XY}}\left[r_c(\theta, x, y)\,x\right]\right\|_2}{\left\|\mathbb{E}_{(x,y)\sim P_{XY}}\left[r(\theta, x, y)\,x\right]\right\|_2}, \qquad \nu_c(\theta) = \frac{\mathbb{E}_{(x,y)\sim P_{XY}}\left[r_c(\theta, x, y)^2\right]}{\mathbb{E}_{(x,y)\sim P_{XY}}\left[r(\theta, x, y)^2\right]}. \tag{17}$$

Lemma 11 below provides a closed-form expression for $\mu_c(\theta)$ and $\nu_c(\theta)$ in terms of $\mathcal{P}(\theta)$ and c, and the subsequent Lemma 12 gives bounds that will be useful in the rest of the analysis.

Lemma 11 *Let Assumption 1 hold, and let $\mu_c(\theta)$ and $\nu_c(\theta)$ be defined according to (17). Then, we have*

$$\mu_c(\theta) = \operatorname{erf}\left(\frac{c}{2\sqrt{\mathcal{P}(\theta)}}\right), \tag{18}$$

$$\nu_c(\theta) = \frac{c^2}{2\mathcal{P}(\theta)}\left(1 - \operatorname{erf}\left(\frac{c}{2\sqrt{\mathcal{P}(\theta)}}\right)\right) + F\left(\frac{c}{\sqrt{2\mathcal{P}(\theta)}}\right), \tag{19}$$

where

$$\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2}\mathrm{d}t, \qquad F(z) = \frac{1}{\sqrt{2\pi}}\int_{-z}^z t^2 e^{-t^2/2}\mathrm{d}t = \operatorname{erf}\left(\frac{z}{\sqrt{2}}\right) - \sqrt{\frac{2}{\pi}} z e^{-z^2/2}. \tag{20}$$

In particular, $\mu_c(\theta)$ and $\nu_c(\theta)$ depend only on c and the test risk $\mathcal{P}(\theta)$ via the ratio $c/\sqrt{2\mathcal{P}(\theta)}$.

Proof Recall that

$$r(\theta, x, y) = x^\top\theta - y, \qquad r_c(\theta, x, y) = r(\theta, x, y)\min\left(1, \frac{c}{|r(\theta, x, y)|}\right), \tag{21}$$

$$\mu_c(\theta) = \frac{\|\mathbb{E}_{x,y}\left[r_c(\theta, x, y)\, x\right]\|_2}{\|\mathbb{E}_{x,y}\left[r(\theta, x, y)\, x\right]\|_2}, \qquad \nu_c(\theta) = \frac{\mathbb{E}_{x,y}\left[r_c(\theta, x, y)^2\right]}{\mathbb{E}_{x,y}\left[r(\theta, x, y)^2\right]}. \tag{22}$$

Until the end of the proof, we will use the notation $\operatorname{clip}_c(\cdot) : \mathbb{R} \to \mathbb{R}$ to denote the function such that

$$\operatorname{clip}_c(a) = a\min\left(1, \frac{c}{|a|}\right). \tag{23}$$

In particular, $r_c(\theta, x, y) = \operatorname{clip}_c\left(r(\theta, x, y)\right)$.

Let us look at the first entry of the vector $\mathbb{E}_{x,y}\left[r_c(\theta, x, y)\, x\right]$,

$$\mathbb{E}_{x,y}\left[r_c(\theta, x, y)x^\top e_1\right] = \mathbb{E}_{\rho_1,\rho_2}\left[\operatorname{clip}_c(\rho_1)\rho_2\right], \tag{24}$$

where the second step introduced ρ_1 and ρ_2, defined as two mean-0 Gaussian random variables, such that

$$\operatorname{Var}(\rho_1) = \left\|\Sigma^{1/2}(\theta - \theta^*)\right\|_2^2 + \zeta^2, \qquad \operatorname{Var}(\rho_2) = \Sigma_{11}, \qquad \operatorname{Cov}(\rho_1, \rho_2) = e_1^\top\Sigma(\theta - \theta^*). \tag{25}$$

Then, we have

$$\begin{aligned}
\mathbb{E}_{\rho_1,\rho_2}\left[\operatorname{clip}_c(\rho_1)\rho_2\right] &= \frac{\operatorname{Cov}(\rho_1, \rho_2)}{\operatorname{Var}(\rho_1)}\mathbb{E}_{\rho_1}\left[\operatorname{clip}_c(\rho_1)\rho_1\right] \\
&= \frac{\operatorname{Cov}(\rho_1, \rho_2)}{\sqrt{\operatorname{Var}(\rho_1)}}\mathbb{E}_{\hat\rho}\left[\operatorname{clip}_c(\sqrt{\operatorname{Var}(\rho_1)}\hat\rho)\hat\rho\right] \\
&= \frac{e_1^\top\Sigma(\theta - \theta^*)}{\sqrt{2\mathcal{P}(\theta)}}\mathbb{E}_{\hat\rho}\left[\operatorname{clip}_c(\sqrt{2\mathcal{P}(\theta)}\hat\rho)\hat\rho\right],
\end{aligned} \tag{26}$$

where we used $\mathcal{P}(\theta) = \mathcal{R}(\theta)+\zeta^2/2 = \mathrm{Var}(\rho_1)/2$ and we introduced the standard Gaussian random variable $\hat{\rho}$. As this argument holds for any component of the vector $\mathbb{E}_{x,y}\left[r_c(\theta,x,y)\,x\right]$, plugging the equation above in (24) gives

$$\begin{aligned}\left\|\mathbb{E}_{x,y}\left[r_c(\theta,x,y)\,x\right]\right\|_2 &= \frac{\|\Sigma(\theta-\theta^*)\|_2}{\sqrt{2\mathcal{P}(\theta)}}\mathbb{E}_{\hat{\rho}}\left[\mathrm{clip}_c(\sqrt{2\mathcal{P}(\theta)}\hat{\rho})\hat{\rho}\right] \\ &= \frac{\left\|\mathbb{E}_{x,y}\left[r(\theta,x,y)\,x\right]\right\|_2}{\sqrt{2\mathcal{P}(\theta)}}\mathbb{E}_{\hat{\rho}}\left[\mathrm{clip}_c(\sqrt{2\mathcal{P}(\theta)}\hat{\rho})\hat{\rho}\right],\end{aligned} \tag{27}$$

where in the second step we used that $\mathbb{E}_{x,y}\left[r(\theta,x,y)\,x\right] = \Sigma(\theta-\theta^*)$. Then, we also have

$$\mu_c(\theta) = \frac{\mathbb{E}_{\hat{\rho}}\left[\mathrm{clip}_c(\sqrt{2\mathcal{P}(\theta)}\hat{\rho})\hat{\rho}\right]}{\sqrt{2\mathcal{P}(\theta)}}. \tag{28}$$

Defining the shorthand $c'(\theta) = c/\sqrt{2\mathcal{P}(\theta)}$, the numerator of the expression above yields

$$\begin{aligned}\mathbb{E}_{\hat{\rho}}\left[\mathrm{clip}_c(\sqrt{2\mathcal{P}(\theta)}\hat{\rho})\hat{\rho}\right] &= \frac{\sqrt{2\mathcal{P}(\theta)}}{\sqrt{2\pi}}\int_{-c'(\theta)}^{c'(\theta)}\hat{\rho}^2 e^{-\hat{\rho}^2/2}\mathrm{d}\hat{\rho} + \frac{2c}{\sqrt{2\pi}}\int_{c'(\theta)}^{+\infty}\hat{\rho}e^{-\hat{\rho}^2/2}\mathrm{d}\hat{\rho} \\ &= \frac{\sqrt{2\mathcal{P}(\theta)}}{\sqrt{2\pi}}\left(-\hat{\rho}e^{-\hat{\rho}^2/2}\Big|_{-c'(\theta)}^{c'(\theta)} + \int_{-c'(\theta)}^{c'(\theta)} e^{-\hat{\rho}^2/2}\mathrm{d}\hat{\rho}\right) - \frac{2c}{\sqrt{2\pi}}\,e^{-\hat{\rho}^2/2}\Big|_{c'(\theta)}^{+\infty} \\ &= \frac{\sqrt{2\mathcal{P}(\theta)}}{\sqrt{2\pi}}\left(-2c'(\theta)e^{-c'(\theta)^2/2} + \int_{-c'(\theta)}^{c'(\theta)} e^{-\hat{\rho}^2/2}\mathrm{d}\hat{\rho}\right) + \frac{2c}{\sqrt{2\pi}}e^{-c'(\theta)^2/2} \\ &= \frac{\sqrt{\mathcal{P}(\theta)}}{\sqrt{\pi}}\int_{-c'(\theta)}^{c'(\theta)} e^{-\hat{\rho}^2/2}\mathrm{d}\hat{\rho} \\ &= \frac{2\sqrt{2}\sqrt{\mathcal{P}(\theta)}}{\sqrt{\pi}}\int_{0}^{c'(\theta)/\sqrt{2}} e^{-\hat{\rho}^2}\mathrm{d}\hat{\rho} \\ &= \sqrt{2\mathcal{P}(\theta)}\,\mathrm{erf}\left(\frac{c}{\sqrt{4\mathcal{P}(\theta)}}\right),\end{aligned} \tag{29}$$

which, plugged in (28), gives the first part of the thesis.

For the second part of the thesis, following the same argument we used to write (28), we have

$$\nu_c(\theta) = \frac{\mathbb{E}_{\hat{\rho}}\left[\mathrm{clip}_c(\sqrt{2\mathcal{P}(\theta)}\hat{\rho})^2\right]}{2\mathcal{P}(\theta)}, \tag{30}$$

where, as before, $\hat{\rho}$ denotes a standard Gaussian random variable. Then, we have

$$\begin{aligned}\nu_c(\theta) &= \frac{1}{\sqrt{2\pi}}\int_{-c'(\theta)}^{c'(\theta)}\hat{\rho}^2 e^{-\hat{\rho}^2/2}\mathrm{d}\hat{\rho} + \frac{2c'(\theta)^2}{\sqrt{2\pi}}\int_{c'(\theta)}^{+\infty} e^{-\hat{\rho}^2/2}\mathrm{d}\hat{\rho} \\ &= \frac{1}{\sqrt{2\pi}}\int_{-c'(\theta)}^{c'(\theta)}\hat{\rho}^2 e^{-\hat{\rho}^2/2}\mathrm{d}\hat{\rho} + c'(\theta)^2\left(1 - \frac{2}{\sqrt{2\pi}}\int_{0}^{c'(\theta)} e^{-\hat{\rho}^2/2}\mathrm{d}\hat{\rho}\right) \\ &= \frac{c^2}{2\mathcal{P}(\theta)}\left(1 - \mathrm{erf}\left(\frac{c}{2\sqrt{\mathcal{P}(\theta)}}\right)\right) + F(c'(\theta)),\end{aligned} \tag{31}$$

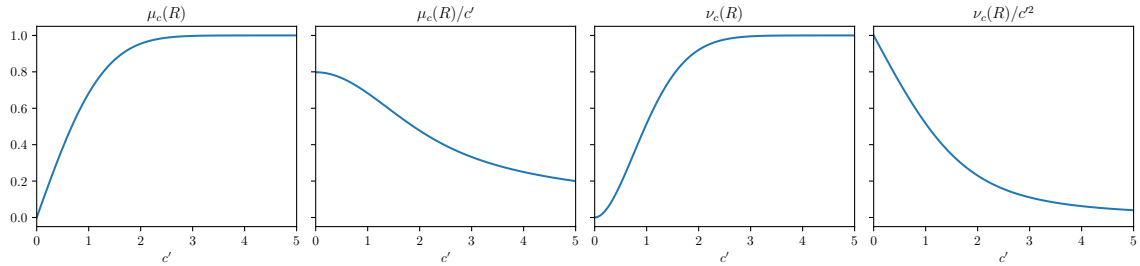

Figure 1: The functions $\mu_c(R)$, $\mu_c(R)/c'$, $\nu_c(R)$, $\nu_c(R)/c'$, plotted as a function of $c' = c/\sqrt{2R+\zeta^2}$.

which concludes the proof. ■

Lemma 12 *Let Assumption 1 hold, and let $\mu_c(R)$ and $\nu_c(R)$ be defined according to (17), where R denotes a generic value of the test risk, since $\mu_c(\theta)$ and $\nu_c(\theta)$ depend only on c and the test risk $\mathcal{P}(\theta)$ due to Lemma 11. Then, for any $c > 0$, we have that*

$$\underline{c}_\mu(c,\zeta) < \frac{\mu_c(R)\sqrt{2R+\zeta^2}}{c} < \sqrt{\frac{2}{\pi}}, \qquad \underline{c}_\nu(c,\zeta) < \frac{\nu_c(R)(2R+\zeta^2)}{c^2} < 1, \tag{32}$$

where $\underline{c}_\mu(c,\zeta)$ and $\underline{c}_\nu(c,\zeta)$ denote two positive constants which depend on the values of c and ζ and are monotonously decreasing in c.

We also have that, as $c/\zeta \to 0$,

$$\left|\frac{\mu_c(R)\sqrt{2R+\zeta^2}}{c} - \sqrt{\frac{2}{\pi}}\right| = o(1), \qquad \left|\frac{\nu_c(R)\left(2R+\zeta^2\right)}{c^2} - 1\right| = o(1). \tag{33}$$

Furthermore, we have

$$\begin{aligned} &\frac{\nu_c(R)(2R+\zeta^2)}{c^2} > \frac{1}{2}, && \text{if } \frac{c}{\sqrt{2R+\zeta^2}} \le 1, \\ &\nu_c(R) > \frac{1}{2}, && \text{if } \frac{c}{\sqrt{2R+\zeta^2}} > 1. \end{aligned} \tag{34}$$

Proof Note that, introducing the notation

$$c' = \frac{c}{\sqrt{2R+\zeta^2}} \le \frac{c}{\zeta}, \tag{35}$$

we have that

$$\mu_c(R) = \operatorname{erf}\left(c'/\sqrt{2}\right) < 1, \tag{36}$$

and

$$\nu_c(R) = (c')^2\left(1 - \operatorname{erf}\left(c'/\sqrt{2}\right)\right) + F(c') < 1, \tag{37}$$

where the last inequalities can be verified directly via the definitions in (17). Furthermore, we have that both $\mu_c(R)$ and $\nu_c(R)$ are increasing functions of c', equal to 0 when $c' = 0$. This follows from the definition for $\mu_c(R)$, and can be promptly verified for $\nu_c(R)$ via derivation.

We further have that, for $c' > 0$,

$$\frac{\mu_c(R)\sqrt{2R+\zeta^2}}{c} = \frac{1}{c'}\mathrm{erf}\left(c'/\sqrt{2}\right) < \sqrt{\frac{2}{\pi}}, \tag{38}$$

and

$$\frac{\nu_c(R)\left(2R+\zeta^2\right)}{c^2} = \left(1-\mathrm{erf}\left(c'/\sqrt{2}\right)\right) + \frac{F(c')}{(c')^2} < 1, \tag{39}$$

where both the LHSs are decreasing functions of c', going to 0 for $c' \to +\infty$ (this can be seen via explicit derivation with respect to c', and via the identity $0 \geq \int_0^z -2t^2 e^{-t^2}\mathrm{d}t = ze^{-z^2} - \int_0^z e^{-t^2}\mathrm{d}t$), and where the last inequalities can be verified computing the limit for $c' \to 0^+$ via l'Hôpital rule, which gives

$$\lim_{c'\to 0^+} \frac{\mu_c(R)\sqrt{2R+\zeta^2}}{c} = \sqrt{\frac{2}{\pi}}, \qquad \lim_{c'\to 0^+} \frac{\nu_c(R)\left(2R+\zeta^2\right)}{c^2} = 1. \tag{40}$$

Note that, as $R \geq 0$, the above limit is achieved when $c/\zeta \to 0$. Then, due to the inequality in (35), the first and second part of the thesis follow.

Note that, for $c' = 1$, we have

$$\nu_c(R) = 1 - \sqrt{\frac{2}{\pi e}} \approx 0.516. \tag{41}$$

Thus, the third and fourth part of the thesis follow from the monotonicity of $\nu_c(R)/c'$ and $\nu_c(R)$. ■

Appendix D. Deterministic equivalent

As $\mu_c(\theta)$ and $\nu_c(\theta)$ depend only on c and the test risk $\mathcal{P}(\theta)$ via the ratio $c/\sqrt{2\mathcal{P}(\theta)}$ (see Lemma 11 and Figure 1 in Appendix C), we use the notation $\mu_c(R)$ and $\nu_c(R)$, where R is the noiseless test risk. We also use the shorthand $\lambda_{\max}$ ($\lambda_{\min}$) to denote the largest (smallest) eigenvalue of the covariance matrix Σ.

Proposition 13 *Let Assumptions 1 and 2 hold. Let $\rho = \Theta(1)$ and $n, d \to \infty$ s.t. $d/n \to \gamma \in (0,\infty)$. Define $\overline{R}(t), \underline{R}(t) : [0,1] \to \mathbb{R}$ as the unique solutions of the following ODEs*

$$\begin{aligned} \mathrm{d}\overline{R}(t) &= -2\lambda_{\min}\tilde{\eta}(t)\mu_c(\overline{R})\overline{R}\mathrm{d}t + \lambda_{\max}\tilde{\eta}^2(t)\nu_c(\overline{R})(\overline{R}+\zeta^2/2)\gamma\mathrm{d}t + 2c^2\tilde{\sigma}^2(t)\gamma^2\mathrm{d}t, \\ \mathrm{d}\underline{R}(t) &= -2\lambda_{\max}\tilde{\eta}(t)\mu_c(\underline{R})\underline{R}\mathrm{d}t + \tilde{\eta}^2(t)\nu_c(\underline{R})(\underline{R}+\zeta^2/2)\gamma\mathrm{d}t + 2c^2\tilde{\sigma}^2(t)\gamma^2\mathrm{d}t, \end{aligned} \tag{42}$$

where $\overline{R}(0) = \underline{R}(0) = \|\Sigma^{1/2}\theta^*\|_2^2/2$. Then, with overwhelming probability, we have

$$\sup_{t\in[0,1)} \left(\mathcal{R}(\theta_{\lfloor tn\rfloor}) - \overline{R}(t)\right) = O\left(\frac{\log^2 n}{\sqrt{n}}\right), \qquad \sup_{t\in[0,1)} \left(\underline{R}(t) - \mathcal{R}(\theta_{\lfloor tn\rfloor})\right) = O\left(\frac{\log^2 n}{\sqrt{n}}\right). \tag{43}$$

Proposition 13 gives that $\overline{R}(t)$ and $\underline{R}(t)$ are asymptotically an upper bound and a lower bound for the test risk $\mathcal{R}(\theta_{\lfloor tn\rfloor})$. In the isotropic case $\Sigma = I$, the upper and lower bounds coincide, i.e., $\overline{R}(t) = \underline{R}(t)$ for all $t \in [0,1]$. Intuitively, compared to the isotropic case, the upper bound $\overline{R}(t)$

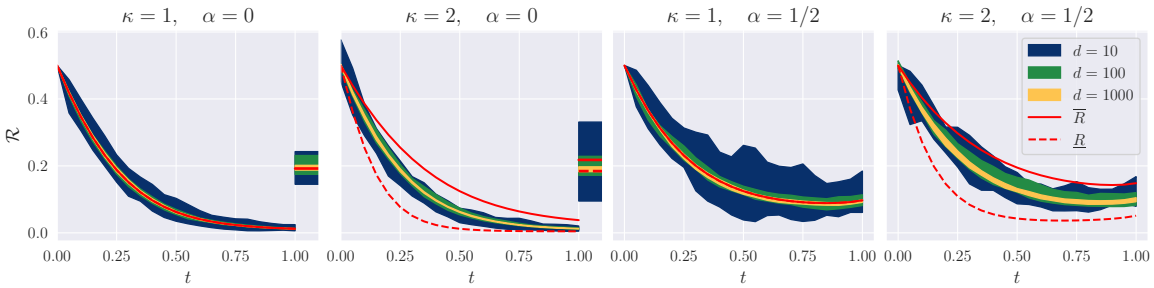

Figure 2: Numerical simulations for Algorithm 1 ($d = 10, 100, 1000$) and the ODEs in (42). We consider two schedules of the form in (6): output perturbation ($\alpha = 0$, first and second panel) and DP-SGD with constant noise ($\alpha = 1/2$, third and fourth panel). We fix $\gamma = 0.1$, $\rho = 1$, $\zeta = 0.3$, $\tilde{\eta}(0) = 3$, $c = 1$, and consider both isotropic data ($\kappa = 1$, first and third panel) and data covariance with condition number $k = 2$ (second and fourth panel). For $\alpha = 0$, we also report for $t \geq 1$ the risk $\mathcal{R}(\theta^p)$, and the values of $\overline{R}(1) + 2c^2\tilde{\eta}^2(1)\gamma^2/\rho^2$ and $\underline{R}(1) + 2c^2\tilde{\eta}^2(1)\gamma^2/\rho^2$ with a red continuous and dashed line respectively. θ^* is sampled uniformly on the unit sphere and the spectrum of Σ follows a power law with appropriate exponent to achieve the specified value of κ. For each value of d, we report bands corresponding to 1 standard deviation around the mean over 10 independent trials of Algorithm 1. In the first and third panel, we have $\overline{R}(t) = \underline{R}(t)$ as the ODEs in (42) match.

reduces the descent term by a factor $\lambda_{\min} \leq 1$ and increases the SGD noise diffusion term by a factor $\lambda_{\max} \geq 1$. Instead, the lower bound $\underline{R}(t)$ just increases the descent term by a factor $\lambda_{\max} \geq 1$. Tighter bounds are possible by making additional assumptions on the covariance spectrum, e.g., a power-law decay (as considered in the theoretical literature [33, 37, 40] also in the context of DP algorithms [21]). The proof of Proposition 13 is tied to the one of Theorem 4, and it relies on the fact that the predictable part from the Doob's decomposition of DP-SGD can be expressed via a family of coupled ODEs. To give the bounds in (42), we decouple this system relying on ODE comparison arguments. The complete proof is in the later Appendix D.1.

The convergence of Proposition 13 is already evident at moderate values of n, d, as showcased by Figure 2 for different schedules ($\tilde{\eta}(t) = 1$ and $\tilde{\eta}(t) = \sqrt{1-t}$) and different data covariances (having condition numbers $\kappa = 1, 2$): the upper and lower bounds on $\mathcal{R}(\theta_{\lfloor tn \rfloor})$ coming from the ODEs in (42) become more accurate as d, n increase and, for isotropic data, they match.

Both Theorem 4 and Proposition 13 require that $\sup_{t\in[0,1]} \tilde{\eta}(t) < 2/\gamma$, which guarantees, due to Lemma 16, that the adaptive learning rate step does not take place with overwhelming probability, *i.e.* $\bar{\eta}_k = \eta_k$ for every $k \in [n]$. Notice that this corresponds to the stability conditions for SGD from [17], with the difference in scaling motivated by $\mathrm{tr}(\Sigma) = d$ and Assumption 2.

D.1. Proof of Theorem 4 and Propositions 13 and 5

We will use the notation $\|Z\|_{\psi_p} = \inf\{t > 0 : \mathbb{E}\exp(|Z|^p/t^p) \leq 2\}$ to denote the Orlicz norm of a random variable Z for any $p \geq 1$. We denote the inner product between two vectors a and b as $\langle a, b\rangle = a^\top b$. Given two real valued quantities a, b, we denote by $a \wedge b = \min(a, b)$. Furthermore, we will denote with γ_n the ratio d/n for a fixed value of n.

It is useful to provide a statement of Theorem 4 for a more general class of functions other than $\mathcal{R}$. As in [41], we will work with a set of quadratic functions of the form

$$Q := \{v \mapsto v^\top R(z;\Sigma)v, \forall z \in \Omega\}, \tag{44}$$

where $\Omega := \{w \in \mathbb{C} : |w| = \max(1, 2\|\Sigma\|_{\text{op}})\}$ and $R(z;\Sigma) = (\Sigma - zI)^{-1}$ is the resolvent matrix of Σ. We further introduce a norm $\|\cdot\|_{C^2}$ on quadratic functions $q : \mathbb{R}^d \to \mathbb{C}$ such that

$$\|q\|_{C^2} := \|\nabla^2 q\|_{\text{op}} + \|\nabla q(0)\|_2 + |q(0)|. \tag{45}$$

The proof relies on few separate technical lemmas, whose statements and proofs are deferred to the later Appendix D.2.

Theorem 14 *Let Assumptions 1 and 2 hold. Let $\rho = \Theta(1)$, $n, d \to \infty$ s.t. $d/n \to \gamma \in (0,\infty)$, and $\sup_{t\in[0,1]} \tilde{\eta}(t) < 2/\gamma$. Denote by Θ_t and θ_k independent realizations of H-DP-SGD (as per Definition 3) and Algorithm 1. Let $D_0 = \|\theta_0 - \theta^*\|^2$. Then there exists a constant $C = C(c, \gamma, \rho, D_0, \zeta)$ such that, for any function $q \in Q$*

$$\sup_{t\in[0,1)} \left|q(\Theta_t) - q(\theta_{\lfloor tn \rfloor})\right| = O\left(\frac{\mathcal{E}\log^2 n}{\sqrt{n}}\right), \tag{46}$$

with overwhelming probability, where $\mathcal{E} = \exp\left(C\int_0^1 \frac{1}{\sqrt{\mathcal{R}(\Theta_s)+\mathcal{R}(\theta_{\lfloor sn \rfloor})}}\mathrm{d}s\right)$.

Proof Let us first focus on Algorithm 1 where no adaptive learning rate step is taken (this will be shown to happen with overwhelming probability in (59)). Then, introducing the notation $u_k = \theta_k - \theta^*$, we have the following update rule

$$u_{k+1} = u_k - \eta_k \bar{g}_k + 2c\sqrt{d}\sigma_k b_{k+1}, \tag{47}$$

where we recall that

$$\bar{g}_k = g_k \min\left(1, \frac{c\sqrt{d}}{\|g_k\|_2}\right), \qquad g_k = \langle x_{k+1}, u_k\rangle x_{k+1}. \tag{48}$$

Let $q : \mathbb{R}^d \to \mathbb{R}$ be any quadratic function. Then, using a Taylor expansion, the update rule for $q(u_k)$ reads

$$\begin{aligned} q(u_{k+1}) = q(u_k) &- \frac{\tilde{\eta}(k/n)}{n}\langle \bar{g}_k, \nabla q(u_k)\rangle + 2c\sqrt{d}\sigma_k\langle b_{k+1}, \nabla q(u_k)\rangle \\ &+ \frac{1}{2}\langle (2c\sqrt{d}\sigma_k b_{k+1} - \frac{1}{n}\tilde{\eta}(k/n)\bar{g}_k)^{\otimes 2}, \nabla^2 q(u_k)\rangle. \end{aligned} \tag{49}$$

Defining the σ-algebra $\mathcal{F}_k := \sigma(\{u_i\}_{i=0}^k)$ generated by the iterates of DP-SGD in (49), it then follows via Doob's decomposition that the above process can be decomposed into its predictable martingales and errors parts (see also Eq. (49) in [41]):

$$\begin{aligned} q(u_{k+1}) - q(u_k) = &-\frac{\tilde{\eta}(k/d)}{n}\mu_c(u_k)\langle \Sigma u_k, \nabla q(u_k)\rangle \\ &+ \frac{\tilde{\eta}^2(k/n)}{2n}\nu_c(u_k)\mathcal{P}(u_k)\frac{1}{n}\mathrm{tr}(\Sigma\nabla^2 q(u_k)) + \frac{1}{n}\langle \frac{2d}{n^2}c^2\tilde{\sigma}(k/n)^2 I_d, \nabla^2 q(u_k)\rangle \\ &+ \Delta\mathcal{M}_k^{\text{Grad}}(q) + \Delta\mathcal{M}_k^{\text{Hess}}(q) + \mathbb{E}[\Delta\mathcal{E}_k^{\text{Hess}}(q) \mid \mathcal{F}_k] \\ &+ \mathcal{M}_k^{\text{Noise}}(q) + \mathbb{E}[\Delta\mathcal{E}_k^{\text{Noise}}(q) \mid \mathcal{F}_k], \end{aligned} \tag{50}$$

where we introduced the shorthands

$$\begin{aligned}
\Delta\mathcal{M}_k^{\text{Grad}}(q) &:= -\frac{\tilde{\eta}(k/n)}{n}\langle \bar{g}_k, \nabla q(u_k)\rangle + \frac{\tilde{\eta}(k/d)}{n}\mu_c(u_k)\langle \Sigma u_k, \nabla q(u_k)\rangle \qquad (51)\\
\Delta\mathcal{M}_k^{\text{Hess}}(q) &:= \frac{1}{2n^2}\langle (\tilde{\eta}(k/n)\bar{g}_k)^{\otimes 2}, \nabla^2 q(u_k)\rangle - \frac{1}{2n^2}\langle \mathbb{E}[(\tilde{\eta}(k/n)\bar{g}_k)^{\otimes 2} \mid \mathcal{F}_k], \nabla^2 q(u_k)\rangle\\
\mathbb{E}[\Delta\mathcal{E}_k^{\text{Hess}}(q) \mid \mathcal{F}_k] &:= \frac{1}{2n^2}\langle \mathbb{E}[(\tilde{\eta}(k/n)\bar{g}_k)^{\otimes 2} \mid \mathcal{F}_k], \nabla^2 q(u_k)\rangle\\
&\quad - \frac{\tilde{\eta}(k/n)^2}{2n}\nu_c(u_k)\mathcal{P}(u_k)\frac{1}{n}\text{tr}(\Sigma\nabla^2 q(u_k))\\
\Delta\mathcal{M}_k^{\text{Noise}}(q) &:= 2c\sqrt{d}\sigma_k\langle b_{k+1}, \nabla q(u_k)\rangle + \frac{1}{2\sqrt{n}}\langle 2c\sqrt{\frac{d}{n}}\sigma_k b_{k+1}\tilde{\eta}(k/n)\bar{g}_k^\top, \nabla^2 q(u_k)\rangle\\
&\quad + \frac{d}{2}\langle (2c\sigma_k b_{k+1})^{\otimes 2}, \nabla^2 q(u_k)\rangle - 2dc^2\langle \sigma_k^2 I_d, \nabla^2 q(u_k)\rangle\\
\mathbb{E}[\Delta\mathcal{E}_k^{\text{Noise}}(q) \mid \mathcal{F}_k] &:= 2dc^2\langle \left(\sigma_k^2 - \frac{1}{n^3}\tilde{\sigma}(k/n)^2\right) I_d, \nabla^2 q(u_k)\rangle,
\end{aligned}$$

where the first three terms are in common with the analysis in [41], while the last two are the result of the private noise in Algorithm 1.

In a similar way, we introduce the shorthand $V_t = \Theta_t - \theta^*$ such that $V_0 = u_0$. Using Itô's formula on (3), for any quadratic function q, we have that

$$\begin{aligned}
\mathrm{d}q(V_t) = &- \tilde{\eta}(t)\mu_c(V_t)\langle \Sigma V_t, \nabla q(V_t)\rangle \mathrm{d}t + \tilde{\eta}^2(t)\nu_c(V_t)\mathcal{P}(V_t)\frac{1}{n}\text{tr}\left(\Sigma\nabla^2 q(V_t)\right)\mathrm{d}t\\
&+ 2\frac{d}{n^2}c^2\tilde{\sigma}^2(t)\text{tr}\left(\nabla^2 q(V_t)\right)\mathrm{d}t + \mathrm{d}\mathcal{M}_t^{\text{H-DP-SGD}},
\end{aligned} \qquad (52)$$

where we introduced the shorthand

$$\mathrm{d}\mathcal{M}_t^{\text{H-DP-SGD}} := \langle \nabla q(V_t), \sqrt{\frac{2\tilde{\eta}(t)^2\nu_c(\Theta_t)\mathcal{P}(\Theta_t)\Sigma}{n} + 4\frac{d}{n^2}c^2\tilde{\sigma}(t)^2 I_d}, \mathrm{d}B_t\rangle, \qquad (53)$$

with B_t being a d-dimensional standard Brownian motion.

Let M be a positive constant that will be fixed later. The dynamic is first controlled up to the stopping time

$$\tau := \inf\{k : \|u_k\|_2 \geq M \cup \lfloor tn \rfloor : \|V_t\|_2 \geq M\}. \qquad (54)$$

Then, we will denote the stopped processes $u_k^\tau = u_{k\wedge\tau}$ and $V_t^\tau = V_{t\wedge(\tau/n)}$, which will be the objects we will compare in the following arguments. This stopping time is introduced for technical reasons (see, *e.g.*, (86)), and we will later show that $\tau \geq n$.

Denoting with $\bar{\eta} = \sup_{t\in[0,1]}\tilde{\eta}(t)$, taking the difference between (50) and (52), and following the same argument in Lemma 1 in [41], we get that there are two absolute constants $C_1 =$

$C_1(\|\Sigma\|_{\mathrm{op}}, c, \bar{\eta})$, $C_2 = C_2(\zeta, c, \bar{\eta}) > 0$ such that, almost surely,

$$\begin{aligned}\sup_{0\le t<1}\left|q(u^{\tau}_{\lfloor tn\rfloor}) - q(V^{\tau}_t)\right| &\le \int_0^1 (C_1 + \frac{C_2}{m_s}) \sup_{q\in Q}\left|q(u^{\tau}_{\lfloor sn\rfloor}) - q(V^{\tau}_s)\right| \mathrm{d}s \\ &+ \sup_{0\le t<(1\wedge(\tau/n))}\left(|\mathcal{M}^{\text{Grad}}_{\lfloor tn\rfloor}(q)| + |\mathcal{M}^{\text{Hess}}_{\lfloor tn\rfloor}(q)| + |\mathcal{M}^{\text{Noise}}_{\lfloor tn\rfloor}(q)| + |\mathcal{M}^{\text{H-DP-SGD}}_t(q)|\right) \\ &+ \sup_{0\le t<(1\wedge(\tau/n))} |\sum_{k=1}^{\lfloor tn\rfloor}\mathbb{E}[\Delta\mathcal{E}^{\text{Hess}}_k(q) \mid \mathcal{F}_k]| + |\sum_{k=1}^{\lfloor tn\rfloor}\mathbb{E}[\Delta\mathcal{E}^{\text{Noise}}_k(q) \mid \mathcal{F}_k]| + O(d^{-1}), \quad (55)\end{aligned}$$

where $m_s = \sqrt{\mathcal{R}(\Theta_s) + \mathcal{R}(\theta_{\lfloor sn\rfloor})}$. Here, we are also using the notation $\mathcal{M}_k(q) = \sum_{j=1}^k \Delta\mathcal{M}_j(q)$, and $\mathcal{M}^{\text{H-DP-SGD}}_t(q) = \int_0^1 \mathrm{d}\mathcal{M}^{\text{H-DP-SGD}}_t(q)$. The last term follows from transitioning (50) to the continuum limit, which involves an additional discretization error of $O(d^{-1})$ (see also Section A.3 in [17] for more details).

Denoting with $\mathcal{M}$ the sum of the last two lines in (55), by Lemma 2 in [41] (for $|\mathcal{M}^{\text{Grad}}_{\lfloor tn\rfloor}(q)|$, $|\mathcal{M}^{\text{Hess}}_{\lfloor tn\rfloor}(q)|$ and $|\sum_{k=1}^{\lfloor tn\rfloor}\mathbb{E}[\Delta\mathcal{E}^{\text{Hess}}_k(q) \mid \mathcal{F}_k]|$), Lemma 17 (for $|\mathcal{M}^{\text{Noise}}_{\lfloor tn\rfloor}(q)|$ and $|\sum_{k=1}^{\lfloor tn\rfloor}\mathbb{E}[\Delta\mathcal{E}^{\text{Noise}}_k(q) \mid \mathcal{F}_k]|$), and Lemma 18 (for $|\mathcal{M}^{\text{H-DP-SGD}}_t(q)|$), there are two constants $C_3(\|\Sigma\|_{\mathrm{op}}, \bar{\eta}, M, c, \gamma)$, and $C_4(\|\Sigma\|_{\mathrm{op}}, \bar{\eta}, M, c) > 0$ such that, for any $u \ge 1$,

$$\mathcal{M} \le C_3 n^{-1/2}(u + C_4), \quad (56)$$

with probability at least $1 - e^{-u}$.

Then, by Lemma 3 in [41] or Lemma 2.2 in [17], we can define a set $\bar{Q} \subseteq Q$ with $|\bar{Q}| \le C(\|\Sigma\|_{\mathrm{op}})d^4$, such that for all $q \in Q$, there exists a $\bar{q} \in \bar{Q}$ that satisfies $\|q - \bar{q}\|_{C^2} \le d^{-2}$. Then, taking the union bound over this set yields

$$\sup_{q\in Q}\sup_{0\le t<1}\left|q(u^{\tau}_{\lfloor tn\rfloor}) - q(V^{\tau}_t)\right| \le \int_0^1 (C_1 + \frac{C_2}{m_s}) \sup_{q\in Q}\left|q(u^{\tau}_{\lfloor sn\rfloor}) - q(V^{\tau}_s)\right| \mathrm{d}s + \mathcal{M}, \quad (57)$$

with probability at least $1 - C(\|\Sigma\|_{\mathrm{op}})d^4 e^{-u}$. Thus, with this same probability, the application of Gronwall's inequality gives

$$\sup_{q\in Q}\sup_{0\le t<1}\left|q(u^{\tau}_{\lfloor tn\rfloor}) - q(V^{\tau}_t)\right| \le \mathcal{M}\exp\left(C_1 + C_2\int_0^1 \frac{1}{m_s}\mathrm{d}s\right). \quad (58)$$

Then, we are left to show that $\tau \ge n$. This follows the same approach as in Lemma 4 in [41], which is here formalized in Lemma 19. In particular, we show that there exists a constant $C(\|\Sigma\|_{\mathrm{op}}, c, v) > 0$, such that for any $r \ge 0$, with probability at least $1 - 2e^{-r^2/2}$, it holds that $\sup_{t\in[0,1]}\|V_t\|^2 \le \|V_0\|^2 e^{Cd^{-1/2}r}$. Then, it can be shown that M (see (54)) can be chosen as a constant independent from d and n, such that $\tau > n$ with overwhelming probability. This is shown before on V_t (via Lemma 19), and later on $u_{\lfloor tn\rfloor}$ via the argument in Eq. (81) in [41].

Lemma 16 guarantees that, if $\sup_{t\in[0,1]}\tilde{\eta}(t) < 2/\gamma$, we have, for all $k \in [n]$,

$$\bar{\eta}_k = \eta_k, \quad (59)$$

with overwhelming probability. Then, the event in (58) intersected with $\tau \geq n$, after setting $u = \log^2 n$, holds with overwhelming probability also on the original algorithm, and the thesis readily follows.

■

Lemma 15 *Let Assumptions 1 and 2 hold. Let $\rho = \Theta(1)$ and $n, d \to \infty$ s.t. $d/n \to \gamma \in (0, \infty)$. Define $\overline{R}(t), \underline{R}(t) : [0,1] \to \mathbb{R}$ as the unique solutions of the following ODEs*

$$\begin{aligned} \mathrm{d}\overline{R}(t) &= -2\lambda_{\min}\tilde{\eta}(t)\mu_c(\overline{R})\overline{R}\mathrm{d}t + \lambda_{\max}\tilde{\eta}^2(t)\nu_c(\overline{R})(\overline{R} + \zeta^2/2)\gamma\mathrm{d}t + 2c^2\tilde{\sigma}^2(t)\gamma^2\mathrm{d}t, \\ \mathrm{d}\underline{R}(t) &= -2\lambda_{\max}\tilde{\eta}(t)\mu_c(\underline{R})\underline{R}\mathrm{d}t + \tilde{\eta}^2(t)\nu_c(\underline{R})(\underline{R} + \zeta^2/2)\gamma\mathrm{d}t + 2c^2\tilde{\sigma}^2(t)\gamma^2\mathrm{d}t, \end{aligned} \tag{60}$$

where $\overline{R}(0) = \underline{R}(0) = \|\Sigma^{1/2}\theta^*\|_2^2/2$. Then denoting with Θ_t a realization of H-DP-SGD (as per Definition 3), with overwhelming probability, we have

$$\sup_{t\in[0,1)} \left(\mathcal{R}(\Theta_t) - \overline{R}(t)\right) = O\left(\frac{\log^2 n}{\sqrt{n}}\right), \qquad \sup_{t\in[0,1)} \left(\underline{R}(t) - \mathcal{R}(\Theta_t)\right) = O\left(\frac{\log^2 n}{\sqrt{n}}\right). \tag{61}$$

Proof Let (λ_i, ω_i) be the eigenvalues and eigenvectors of Σ, and consider the shorthand $\mathcal{D}_i(t) := d\langle V_t, \omega_i\rangle^2/2$. Set $q(V_t) = \frac{1}{2d}\sum_{i=1}^d \frac{1}{\lambda_i - z}\langle V_t, \omega_i\rangle^2$, the argument used in the proof of Theorem 14 can be extended to the set of contours that enclose only the i-th eigenvalue. Then, integrating over both sides of (52) via the Cauchy integral formula, we have that there exists a set of coupled ODEs

$$\mathrm{d}D_i = -2\lambda_i\tilde{\eta}(t)\mu_c(R(t))D_i\mathrm{d}t + \lambda_i\tilde{\eta}^2(t)\nu_c(R(t))(R(t) + \zeta^2/2)\gamma\mathrm{d}t + 2c^2\tilde{\sigma}^2(t)\gamma^2\mathrm{d}t, \tag{62}$$

with $R(t) = \frac{1}{d}\sum_{i=1}^d \lambda_i D_i(t)$ and $D_i(0) = d\lambda_i\langle\omega_i, \theta^*\rangle^2/2$, such that

$$\sup_{t\in[0,1)} |\mathcal{R}(\Theta_t) - R(t))| = O\left(\frac{\mathcal{E}'\log^2 n}{\sqrt{n}}\right), \tag{63}$$

with overwhelming probability, where $\mathcal{E}' = \exp\left(C\int_0^1 \frac{1}{\sqrt{\mathcal{R}(\Theta_s) + R(t)}}\mathrm{d}s\right)$ and C is a positive constant (see pages 9-10 in [17] and Appendix G in [41] for details).

Until the end of the proof, we will define more auxiliary ODEs, such that the RHS is uniformly Lipschitz in the dependent variable at all times $t \in [0,1]$. Then, by the extension of the Picard–Lindelöf theorem (see Corollary 2.6 in [56]), we have that their solutions exist and are unique, and therefore also have continuous derivatives. Then, defining

$$\begin{aligned} \mathrm{d}\overline{D}_i &= -2\lambda_{\min}\tilde{\eta}(t)\mu_c(R(t))\overline{D}_i\mathrm{d}t + \lambda_i\tilde{\eta}^2(t)\nu_c(R(t))(R(t) + \zeta^2/2)\gamma\mathrm{d}t + 2c^2\tilde{\sigma}^2(t)\gamma^2\mathrm{d}t, \\ \mathrm{d}\underline{D}_i &= -2\lambda_{\max}\tilde{\eta}(t)\mu_c(R(t))\underline{D}_i\mathrm{d}t + \lambda_i\tilde{\eta}^2(t)\nu_c(R(t))(R(t) + \zeta^2/2)\gamma\mathrm{d}t + 2c^2\tilde{\sigma}^2(t)\gamma^2\mathrm{d}t, \end{aligned} \tag{64}$$

by standard ODE comparison arguments (see Theorem 1.3 in [56]) we have that $\overline{D}_i(t) \geq D_i(t) \geq \underline{D}_i(t)$ for all $t \in [0,1]$. Then, averaging over i (weighting by λ_i) the equations in (64) and (62), and defining $\overline{R}'(t) = \frac{1}{d}\sum_{i=1}^d \lambda_i\overline{D}_i(t)$ and $\underline{R}'(t) = \frac{1}{d}\sum_{i=1}^d \lambda_i\underline{D}_i(t)$, we get $\overline{R}'(t) \geq R(t) \geq \underline{R}'(t)$ for all $t \in [0,1]$, with

$$\begin{aligned} \mathrm{d}\overline{R}' &= -2\lambda_{\min}\tilde{\eta}(t)\mu_c(R(t))\overline{R}'\mathrm{d}t + \frac{\mathrm{tr}(\Sigma^2)}{d}\tilde{\eta}^2(t)\nu_c(R(t))(R(t) + \zeta^2/2)\gamma\mathrm{d}t + 2c^2\tilde{\sigma}^2(t)\gamma^2\mathrm{d}t, \\ \mathrm{d}\underline{R}' &= -2\lambda_{\max}\tilde{\eta}(t)\mu_c(R(t))\underline{R}'\mathrm{d}t + \frac{\mathrm{tr}(\Sigma^2)}{d}\tilde{\eta}^2(t)\nu_c(R(t))(R(t) + \zeta^2/2)\gamma\mathrm{d}t + 2c^2\tilde{\sigma}^2(t)\gamma^2\mathrm{d}t. \end{aligned} \tag{65}$$

Furthermore, for a fixed value of c, we have that the functions $\mu_c(R)$ and $\nu_c(R)(R+\zeta^2/2)$ are respectively monotonically decreasing and increasing with respect to R (see Lemma 11). Then, since by Jensen inequality we have that $\mathrm{tr}(\Sigma^2) \geq \mathrm{tr}(\Sigma) = d$ (where the last step holds due to Assumption 1), and since we also have $\mathrm{tr}(\Sigma^2) \geq \lambda_{\max}\mathrm{tr}(\Sigma) = \lambda_{\max} d$, by defining

$$\begin{aligned} \mathrm{d}\overline{R} &= -2\lambda_{\min}\tilde{\eta}(t)\mu_c(\overline{R})\overline{R}\mathrm{d}t + \lambda_{\max}\tilde{\eta}^2(t)\nu_c(\overline{R})(\overline{R}+\zeta^2/2)\gamma\mathrm{d}t + 2c^2\tilde{\sigma}^2(t)\gamma^2\mathrm{d}t, \\ \mathrm{d}\underline{R} &= -2\lambda_{\max}\tilde{\eta}(t)\mu_c(\underline{R})\underline{R}\mathrm{d}t + \tilde{\eta}^2(t)\nu_c(\underline{R})(\underline{R}+\zeta^2/2)\gamma\mathrm{d}t + 2c^2\tilde{\sigma}^2(t)\gamma^2\mathrm{d}t, \end{aligned} \tag{66}$$

we get that $\overline{R}(t) \geq \overline{R}'(t) \geq R(t) \geq \underline{R}'(t) \geq \underline{R}(t)$ for all $t \in [0,1]$. Thus, to obtain the thesis, it is sufficient to prove that the term $\mathcal{E}'$ in (63) is of constant order, which is in turn implied by showing that $\underline{R}(t) = \Omega(1)$ for all $t \in [0,1]$. This is readily implied by the fact that

$$\mathrm{d}\underline{R}^{\mathrm{low}} = -2\lambda_{\max}\bar{\eta}\underline{R}^{\mathrm{low}}\mathrm{d}t, \tag{67}$$

$$\mathrm{d}\overline{R}^{\mathrm{up}} = \lambda_{\max}\bar{\eta}^2c^2\gamma\mathrm{d}t + 2c^2\bar{\sigma}^2\gamma^2\mathrm{d}t, \tag{68}$$

are respectively a lower and upper bound of $\underline{R}(t)$ and $\overline{R}(t)$, which have a closed form solution and guarantee that $\underline{R}(t) \geq C_1 > 0$ for every $t \in [0,1]$, giving the desired result. ■

Proof of Theorem 4. The result follows from Theorem 14, after setting $q(\theta-\theta^*) = \left\|\Sigma^{1/2}(\theta-\theta^*)\right\|_2^2$ and proving that $\mathcal{E} = O(1)$. This is due to Lemma 15, which guarantees a lower bound on $\mathcal{R}(\Theta_t)$ via $\underline{R}(t)$ and the latter is shown in the argument after (68) to be $\Omega(1)$ for all $t \in [0,1]$. Note that the upper bound in (68) and the following argument also guarantee $\mathcal{R}(\Theta_t) = O(1)$, for all $t \in [0,1]$, which then yields, with overwhelming probability,

$$\mathcal{R}(\theta_k) = \Theta(1), \tag{69}$$

for any iterate $k \in [n-1]$. ■

Proof of Proposition 13. The result follows from Theorem 4 and Lemma 15, after an application of the triangle inequality.

Proof of Proposition 5. Due to the update rule in Algorithm 1, we have

$$\begin{aligned} 2\mathcal{R}(\theta_n) &= \left\|\Sigma^{1/2}\left(\theta_n-\theta^*\right)\right\|_2^2 \\ &= \left\|\Sigma^{1/2}\left(\theta_{n-1}-\bar{\eta}_n\bar{g}_n+2C_{\mathrm{clip}}\sigma_n b_n-\theta^*\right)\right\|_2^2 \\ &= \left\|\Sigma^{1/2}\left(\theta_{n-1}-\theta^*+2C_{\mathrm{clip}}\sigma_n b_n\right)\right\|_2^2 + \left\|\Sigma^{1/2}\bar{\eta}_n\bar{g}_n\right\|_2^2 \\ &\quad - 2\bar{\eta}_n\bar{g}_n^\top\Sigma\left(\theta_{n-1}-\theta^*\right) - 4\bar{\eta}_n\bar{g}_n^\top\Sigma C_{\mathrm{clip}}\sigma_n b_n. \end{aligned} \tag{70}$$

By Assumptions 1 and 2, and due to the definition of $\bar{g}_n$, we have

$$\left\|\Sigma^{1/2}\bar{\eta}_n\bar{g}_n\right\|_2 \leq \|\Sigma\|_{\mathrm{op}}^{1/2}\left|\bar{\eta}_n\right|\|\bar{g}_n\|_2 = O\left(\frac{\sqrt{d}}{n}\right) = O\left(\frac{1}{\sqrt{d}}\right). \tag{71}$$

Theorem 3.1.1 in [58] also guarantees that $\|b_n\|_2 = O(\sqrt{d})$ with probability at least $1-2\exp(-c_1 d)$, for some absolute constant $c_1 > 0$. Thus,

$$\left|4\bar{\eta}_n \bar{g}_n^\top \Sigma C_{\text{clip}} \sigma_n b_n\right| = \left|4\bar{\eta}_n \bar{g}_n^\top \Sigma C_{\text{clip}} \frac{\eta_n}{\rho} b_n\right| = O\left(\frac{1}{n}\sqrt{d}\sqrt{d}\frac{1}{n}\sqrt{d}\right) = O\left(\frac{1}{\sqrt{d}}\right), \tag{72}$$

which gives

$$\left|2\mathcal{R}(\theta_n) - \left\|\Sigma^{1/2}\left(\theta_{n-1} - \theta^* + 2C_{\text{clip}}\sigma_n \Sigma^{1/2} b_n\right)\right\|_2^2\right| = O\left(\frac{1+\sqrt{\mathcal{R}(\theta_{n-1})}}{\sqrt{d}}\right). \tag{73}$$

Similarly, we have

$$\begin{aligned}\left\|\Sigma^{1/2}\left(\theta_{n-1} - \theta^* + 2C_{\text{clip}}\sigma_n b_n\right)\right\|_2^2 &= \left\|\Sigma^{1/2}\left(\theta_{n-1} - \theta^*\right)\right\|_2^2 + \left\|2C_{\text{clip}}\sigma_n \Sigma^{1/2} b_n\right\|_2^2 \\ &+4C_{\text{clip}}\sigma_n b_n^\top \Sigma\left(\theta_{n-1} - \theta^*\right).\end{aligned} \tag{74}$$

However, since b_n is a standard Gaussian vector, we have that, with probability at least $1-2\exp\left(-c_2\log^2 d\right)$,

$$\begin{aligned}\left|4C_{\text{clip}}\sigma_n b_n^\top \Sigma\left(\theta_{n-1} - \theta^*\right)\right| &\le 4C_{\text{clip}}\sigma_n \left\|\Sigma^{1/2}\right\|_{\text{op}} \left\|\Sigma^{1/2}\left(\theta_{n-1} - \theta^*\right)\right\|_2 \log d \\ &= 4c\sqrt{d}\frac{\tilde{\eta}(1)}{\rho n}\left\|\Sigma^{1/2}\right\|_{\text{op}} \left\|\Sigma^{1/2}\left(\theta_{n-1} - \theta^*\right)\right\|_2 \log d \\ &= O\left(\frac{\sqrt{d}\log d}{n}\right)\left\|\Sigma^{1/2}\left(\theta_{n-1} - \theta^*\right)\right\|_2 \\ &= O\left(\frac{\sqrt{\mathcal{R}(\theta_{n-1})}\log d}{\sqrt{d}}\right).\end{aligned} \tag{75}$$

Then, an application of the Hanson-Wright inequality (see Theorem 6.2.1 in [58]) yields

$$\left|\left\|2C_{\text{clip}}\sigma_n \Sigma^{1/2} b_n\right\|_2^2 - 4c^2 d\frac{\tilde{\eta}^2(1)}{\rho^2 n^2}\text{tr}(\Sigma)\right| \le 4c^2 d\frac{\tilde{\eta}^2(1)}{\rho^2 n^2}\|\Sigma\|\log d = O\left(\frac{\log d}{\sqrt{d}}\right), \tag{76}$$

with probability at least $1 - 2\exp\left(-c_3\log^2 d\right)$. Putting everything together gives

$$\left|\mathcal{R}(\theta_n) - \mathcal{R}(\theta_{n-1}) - 2c^2\tilde{\eta}^2(1)\gamma^2/\rho^2\right| = O\left(\frac{\log d}{\sqrt{d}}\right), \tag{77}$$

with overwhelming probability, where we used $\text{tr}(\Sigma) = d$ and (69). ■

D.2. Technical lemmas

In this section we provide the statements and proofs for the technical lemmas used for the proof of Theorem 4. The notation is defined accordingly, and Assumptions 1 and 2 will always be assumed to hold. In this section, we will use the shorthand $\|\cdot\|$ to denote both the Euclidean norm of vectors and Frobenius norm of matrices.

Lemma 16 *If* $\sup_{t\in[0,1]} \tilde{\eta}(t) < 2/\gamma_n$*, we have that, for all* $k \in [n]$*,*

$$\bar{\eta}_k = \eta_k, \tag{78}$$

with overwhelming probability.

Proof Since we have $\operatorname{tr}(\Sigma) = d$ and $\|\Sigma\|_{\text{op}} = O(1)$ by Assumption 1, due to Theorem 6.3.2 in [58], for every $k \in [n]$, we have that $\left\| \|x_k\|_2 - \sqrt{d} \right\|_{\psi_2} = O(1)$. This implies

$$\mathbb{P}\left(\|x_k\|_2^2 - d > t \right) \leq 2\exp\left(-c_1 t\right), \tag{79}$$

where c_1 is an absolute constant. By hypothesis, there exists a positive constant c_2 such that

$$\eta_k \leq \frac{2}{\gamma_n n(1+c_2)} = \frac{2}{d(1+c_2)}. \tag{80}$$

Then, we have

$$\mathbb{P}\left(\bar{\eta}_k \neq \eta_k\right) = \mathbb{P}\left(\|x_k\|_2^2 > \frac{2}{\eta_k} \right) \leq \mathbb{P}\left(\|x_k\|_2^2 > d(1+c_2) \right) \leq 2\exp\left(-c_3 d\right), \tag{81}$$

where the first step follows from the definition of $\bar{\eta}$ in Algorithm 1, the second step follows from (80), and last step follows from (79). Thus, the deisred result follows via a union bound over all $k \in [n]$, with probability at least $1 - 2n\exp\left(-c_3 d\right) \geq 1 - 2\exp\left(-c_4 d\right)$. ■

Lemma 17 *We have that, for any quadratic* $q \in Q$ *such that* $\|q\|_{C^2} \leq 1$*,*

$$\left| \sum_{k=1}^{\lfloor tn \rfloor} \mathbb{E}[\Delta\mathcal{E}_k^{\text{Noise}}(q) \mid \mathcal{F}_k] \right| \leq \frac{2\gamma_n}{\rho^2 n} C_{\eta,2}, \quad \textit{a.s.} \tag{82}$$

where $C_{\eta,2}$ *denotes the upper bound on the absolute value of the second derivative of* $\tilde{\eta}^2(t)$*. In addition, there is a constant* $C = C(c, \gamma_n, \rho, M) > 0$*, such that, for any* $y \in [1, n]$*,*

$$\sup_{1\leq k\leq (n\wedge\tau)} |\mathcal{M}_k^{\text{Noise}}(q)| \leq C n^{-\frac{1}{2}} y, \tag{83}$$

with probability at least $1 - e^{-y}$*.*

Proof Recall the definition in (51)

$$\begin{aligned}\Delta\mathcal{M}_k^{\text{Noise}}(q) = 2c\sqrt{d}\sigma_k \langle b_{k+1}, \nabla q(u_k)\rangle + \frac{1}{2\sqrt{n}}\langle 2c\sqrt{\frac{d}{n}}\sigma_k b_{k+1}\tilde{\eta}(k/n)\bar{g}_k^\top, \nabla^2 q(u_k)\rangle \\ + 2dc^2\sigma_k^2\langle (b_{k+1})^{\otimes 2}, \nabla^2 q(u_k)\rangle - 2dc^2\sigma_k^2\langle I_d, \nabla^2 q(u_k)\rangle.\end{aligned}$$

Then, for any $k \leq \tau$, we rewrite the martingale as a combination of the following three terms

$$\Delta\mathcal{M}_k^{\text{Noise}}(q) = \frac{1}{n^{3/2}}\langle b_{k+1}, A_{k,1} + A_{k,2}\rangle + \frac{1}{n^2}\langle b_{k+1}^{\otimes 2} - I_d, C_k\rangle, \tag{84}$$

where we introduced the shorthands

$$A_{k,1} := 2cn^{3/2}\sqrt{d}\sigma_k \nabla q(u_k),\ A_{k,2} := cn\sqrt{\frac{d}{n}}\sigma_k \tilde{\eta}(k/n)\nabla^2 q(u_k)\bar{g}_k,\ C_k := 2dn^2c^2\sigma_k^2\nabla^2 q(u_k). \tag{85}$$

We will separately bound the contribution of each term in terms of its Orlicz norm. Let us start with the second term, and consider $q \in Q$. This is a quadratic function of the iterates; its Hessian, therefore, does not depend on the iterates and explicitly on k. In addition we have that $\sup_{z\in\Omega}\|R(z;\Sigma)\|_{\text{op}} \le 2$. Since $b_{k+1} \sim \mathcal{N}(0, I_d)$ and independent, using the Hanson-Wright inequality (Theorem 6.2.1 in [58]) we have that, for some $c > 0$ and $K = \max_{i\in[d]}\|b_{k+1,i}\|_{\psi_2}$,

$$\begin{aligned}\mathbb{P}\left(\frac{1}{n^2}\langle b_{k+1}^{\otimes 2} - I_d, C_k\rangle \ge t\right) &\le 2\exp\left(-c\min\left(\frac{t^2n^4}{K^4\|C_k\|^2}, \frac{tn^2}{K^2\|C_k\|_{\text{op}}}\right)\right)\\ &\le 2\exp\left(-c'\min\left(\frac{t^2n^4}{d}, tn^2\right)\right).\end{aligned}$$

To justify the last passage, note that, by the structure of the noise,

$$\sigma_k^2 \le \frac{1}{\rho^2n^2}|\tilde{\eta}(k/n)^2 - \tilde{\eta}((k+1)/n)^2| \le \frac{2}{\rho^2n^3}\max_{x\in[\frac{k}{n},\frac{k+1}{n}]}\left|\frac{\mathrm{d}}{\mathrm{d}x}\tilde{\eta}^2(x)\right| \le \frac{4}{\rho^2n^3\gamma_n}C_{\eta,1},$$

as $\left|\frac{\mathrm{d}}{\mathrm{d}x}\tilde{\eta}^2(x)\right| \le C_{\eta,1}$ for some $C_{\eta,1}$ due to Assumption 2. Using the above bound, we obtain that $\|C_k\|_{\text{op}} \le 2dn^2c^2\sigma_k^2\|q\|_{C^2} \le 8c^2/\rho^2C_{\eta,1}$, and similarly $\|C_k\| \le 8\sqrt{d}c^2/\rho^2C_{\eta,1}$. We, therefore, have that $\|\frac{1}{n^2}\langle b_{k+1}^{\otimes 2} - I_d, C_k\rangle\|_{\psi_1} \le Cn^{-1}$, for some constant $C(\rho, c, C_{\eta,1}) > 0$.

For the first term, by Eq. (89) in [41], the norm of q is bounded for the stopped process u_k^τ as follows,

$$\|\nabla q(u)\| \le \|\nabla^2 q\|_{\text{op}}\|u\| + \|\nabla q(0)\| \le \|q\|_{C^2}(1 + \|u\|) \le C(1 + M). \tag{86}$$

Hence, we have that $\|A_{k,1}\| \le 2n^{3/2}c\sqrt{n}\sigma_k C(1 + M) \le 4\sqrt{d}c\sqrt{C_{\eta,1}/\gamma_n}C(1 + M)/\rho$, which implies that

$$\left\|\frac{1}{n^{3/2}}\langle b_{k+1}, A_{k,1}\rangle\right\|_{\psi_2} \le \frac{C}{n}, \tag{87}$$

with $C = C(M, \rho, \gamma_n, C_{\eta,1}, c) > 0$. Using a similar analysis as in Eq. (92-99) in [41], we have

$$\begin{aligned}\left|\frac{1}{n^{3/2}}\langle b_{k+1}, A_{k,2}\rangle\right| &\le \frac{1}{n^{3/2}}cn\sqrt{\gamma_n}\sigma_k\tilde{\eta}(k/n)|\langle b_{k+1}, \nabla^2 q(u_k)g_k\rangle 1_{\|g_k\|_2\le c\sqrt{d}}| \\ &+ \frac{1}{n^{3/2}}cn\sqrt{\gamma_n}\sigma_k\tilde{\eta}(k/n)c\sqrt{d}|\langle b_{k+1}, \nabla^2 q(u_k)\frac{g_k}{\|g_k\|_2}\rangle 1_{\|g_k\|_2>c\sqrt{d}}| \\ &\le \frac{4c\sqrt{C_{\eta,1}}}{\gamma_n n\rho}|\langle b_{k+1}, \nabla^2 q(u_k)g_k\rangle| + \frac{4c\sqrt{C_{\eta,1}}}{\gamma_n n\rho}|\langle b_{k+1}, \nabla^2 q(u_k)g_k\rangle| \\ &\le \frac{8c\sqrt{C_{\eta,1}}}{\gamma_n n\rho}|\langle b_{k+1}, \nabla^2 q(u_k)x_{k+1}\rangle(\langle x_{k+1}, u_k\rangle + z_{k+1})|.\end{aligned} \tag{88}$$

Due to (86) and Assumptions 1 and 2 we have $\|\langle b_{k+1}, \nabla^2 q(u_k)x_{k+1}\rangle\|_{\psi_1} \le C\|b_{k+1}\|_{\psi_2}\|x_{k+1}\|_{\psi_2} \le C$ for some $C > 0$ that depends on the $\|\Sigma\|_{\text{op}}$. Finally, we note that $\|\langle x_{k+1}, u_k\rangle\|_{\psi_2} \le C(1 + M)$,

and $\|z_k\|_{\psi_2} \leq C\zeta$ for any $k \leq \tau$. Combining the above, we obtain that there is a constant $C = C(c, C_{\eta,1}, \rho, \gamma_n, M, \zeta) > 0$ such that

$$\phi_{k,1} := \inf\{t > 0 : \mathbb{E}[\exp(|\Delta\mathcal{M}_k^{\text{Noise}}(q)|/t) \mid \mathcal{F}_{k-1}] \leq 2\} \leq Cn^{-1}. \tag{89}$$

We then apply Lemma 5 in [41] for some absolute constants $C > 0$ for all $t > 0$:

$$\begin{aligned}\mathbb{P}\left(\sup_{1\leq k\leq n\wedge\tau} |\mathcal{M}_k^{\text{Noise}}(q) - \mathbb{E}\mathcal{M}_0^{\text{Noise}}(q)| \geq t\right) &\leq 2\exp\left(-\min\{\frac{t}{C\max_{k\in[n]}\phi_{k,1}}, \frac{t^2}{C\sum_{i=1}^n \phi_{i,1}^2}\}\right) \\ &\leq 2\exp\left(-Cn\min\{t, t^2\}\right).\end{aligned} \tag{90}$$

As we assume that n is proportional to d and noting that $\mathbb{E}\mathcal{M}_0^{\text{Noise}}(q) = 0$ by our construction, we then have that, for any $y \in [1, n]$,

$$\sup_{1\leq k\leq(n\wedge\tau)} |\mathcal{M}_k^{\text{Noise}}(q)| \leq Cn^{-\frac{1}{2}}y, \tag{91}$$

with probability at least $1 - e^{-y}$ for any $y \in [1, n]$.

Next, we bound the error due to the discretization:

$$\begin{aligned}\left|\sum_{k=1}^{\lfloor tn\rfloor} \mathbb{E}[\Delta\mathcal{E}_k^{\text{Noise}}(q) \mid \mathcal{F}_k]\right| &\leq 2dc^2 \sum_{k=1}^{\lfloor tn\rfloor} \left|\sigma_k^2 - \frac{1}{n^3}\tilde{\sigma}(k/n)^2\right| \cdot \left|\text{tr}(\nabla^2 q(u_k))\right| \\ &\leq \frac{2d^2}{n^2\rho^2}c^2 \sum_{k=1}^{\lfloor tn\rfloor} \left|\tilde{\eta}(k/n)^2 - \tilde{\eta}((k+1)/n)^2 - \frac{1}{n}\tilde{\sigma}(k/n)^2\right| \\ &\leq \frac{2\gamma_n}{\rho^2 n^2} \sum_{k=1}^{\lfloor tn\rfloor} \max_{x\in(\frac{k}{n},\frac{k+1}{n})} \left|\frac{\mathrm{d}^2}{\mathrm{d}x^2}\tilde{\eta}(x)^2\right|,\end{aligned}$$

where we use the definition of the noise function as the derivative of the learning rate $\tilde{\sigma}(x)^2 = -\frac{\mathrm{d}}{\mathrm{d}x}\tilde{\eta}^2(x)$ at any point $x \in [0, 1)$. Then, as $|\frac{\mathrm{d}^2}{\mathrm{d}x^2}\tilde{\eta}^2(x)| \leq C_{\eta,2}$ for some constant $C_{\eta,2}$, the desired result follows. ■

Lemma 18 *Denote by $\mathcal{M}_t^{\text{H-DP-SGD},\tau}$ the H-DP-SGD martingale in which the stopping time is imposed. There is a constant $C = C(c, C_{\eta,1}, \gamma_n, \|\Sigma\|_{\text{op}}, M) > 0$, such that for any quadratic $q \in Q$ with $\|q\|_{C^2} \leq 1$, and for any $y \in [1, n]$, we have*

$$\sup_{0\leq t<1} |\mathcal{M}_t^{\text{H-DP-SGD},\tau}(q)| \leq Cn^{-\frac{1}{2}}y, \tag{92}$$

with probability at least $1 - e^{-y}$.

Proof To bound the martingale error from H-DP-SGD under some general statistic $q \in Q$, differently from the argument to obtain Eq. (72) in Lemma 2 in [41], we need to control the additional term due to the additive noise in Algorithm 1.

Using Itô's formula for any quadratic function q:

$$\begin{aligned}\mathrm{d}q(\Theta_t) = &- \tilde{\eta}(t)\mu_c(\Theta_t)\nabla\mathcal{P}(\Theta_t)^\top\nabla q(\Theta(t))\mathrm{d}t + \tilde{\eta}^2(t)\nu_c(\Theta_t)\mathcal{P}(\Theta_t)\frac{1}{n}\mathrm{tr}\left(\Sigma\nabla^2 q\right)\mathrm{d}t \\ &+ 2\frac{d}{n^2}c^2\tilde{\sigma}^2(t)\mathrm{tr}\left(\nabla^2 q\right)\mathrm{d}t + \mathrm{d}\mathcal{M}_t^{\text{H-DP-SGD}}, \\ \mathrm{d}\mathcal{M}_t^{\text{H-DP-SGD}} := &\langle\nabla q(V_t), \sqrt{\frac{2\tilde{\eta}(t)^2\nu_c(\Theta_t)\mathcal{P}(\Theta_t)\Sigma}{n} + 4\frac{d}{n^2}c^2\tilde{\sigma}(t)^2 I_d}\,\mathrm{d}B_t\rangle,\end{aligned} \tag{93}$$

with B_t being a standard d dimensional Brownian motion. The quadratic variation of the martingale is then bounded a.s.

$$\langle\mathcal{M}(q)\rangle_t \leq \frac{Ct}{n} + \frac{d}{n^2}(4c^2\bar{\eta}^2\|\Sigma\|_{\mathrm{op}}(1+M)^2),$$

for some constant $C = C(\|\Sigma\|_{\mathrm{op}}, M, C_{\eta,1}, c, \gamma_n) > 0$, where we used Assumption 2 which gives $|\tilde{\sigma}(t)| = |\frac{\mathrm{d}}{\mathrm{d}t}\tilde{\eta}^2(t)| \leq C_{\eta,1}$. The claim is then proved by an application of Gaussian concentration inequalities as in Section B.6 in [41]. ■

Lemma 19 *There exists a constant $C(\|\Sigma\|_{op}, c, v) > 0$ such that for any $r \geq 0$ with probability at least $1 - 2e^{-r^2/2}$ it holds that $\sup_{t\in[0,1]}\|V_t\|^2 \leq \|V_0\|^2 e^{Cd^{-1/2}r}$.*

Proof The proof follows a path similar to the one of Lemma 4 in [41]. In particular, consider the function $\varphi(V_t) = \log(1 + \|V_t\|^2)$. Then, by application of Itô's Lemma to (3),

$$\begin{aligned}\mathrm{d}\varphi(\Theta_t) = &- \tilde{\eta}(t)\frac{\mu_c(\Theta_t)}{1+\|V_t\|^2}\nabla\mathcal{P}(\Theta_t)^\top V_t\mathrm{d}t \\ &+ 2\tilde{\eta}^2(t)\frac{\nu_c(\Theta_t)\mathcal{P}(\Theta_t)}{(1+\|V_t\|^2)^2}\frac{1}{n}\mathrm{tr}\left(\Sigma(I_d(1+\|V_t\|^2) - V_t\otimes V_t)\right)\mathrm{d}t \\ &+ 2\frac{d}{n}c^2\tilde{\sigma}^2(t)\frac{1}{(1+\|V_t\|^2)^2}\mathrm{d}t + \mathrm{d}\mathcal{M}_t(\varphi)\end{aligned} \tag{94}$$

with

$$\mathrm{d}\mathcal{M}_t(\varphi) = \frac{1}{(1+\|V_t\|^2)}\langle V_t, \sqrt{\frac{2\tilde{\eta}(t)^2\nu_c(\Theta_t)\mathcal{P}(\Theta_t)\Sigma}{n} + 4\frac{d}{n^2}c^2\tilde{\sigma}(t)^2 I_d}\,\mathrm{d}B_t\rangle.$$

The drift terms and the quadratic variation terms can be bounded by some $C(c, \|\Sigma\|_{\mathrm{op}}, \zeta, \gamma_n)$. The quadratic variation of the martingale term is bounded $\langle M\rangle_t \leq \frac{Ct}{n}$. We then have by Gaussian concentration inequality,

$$\mathbb{P}\left(\sup_{0\leq t\leq 1}\varphi(V_t) \geq C(1 + r/\sqrt{n})\right) \leq e^{-r^2/2}, \tag{95}$$

which proves the claim. ■

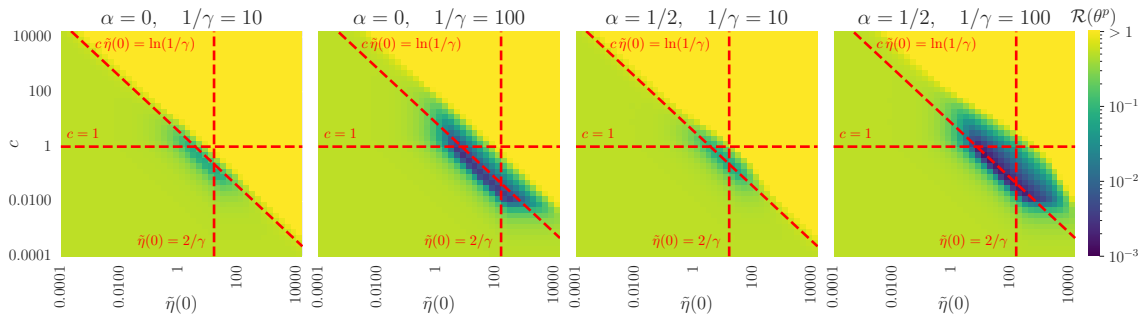

Figure 3: Numerical simulations for $\mathcal{R}(\theta^p)$ obtained via Algorithm 1 for $d = 1000$ and $\kappa = 1$, as a function of c and $\tilde{\eta}(0)$. We consider the schedules in (6) corresponding to output perturbation ($\alpha = 0$, first and second panel) and DP-SGD with constant noise ($\alpha = 1/2$, third and fourth panel), with fixed $\rho = 1$ and $\zeta = 0.3$. We set $\gamma = 0.1$ in the first and third panel, and $\gamma = 0.01$ in the second and fourth panel. The values of $\mathcal{R}(\theta^p)$ are capped at 1, and θ^* is chosen such that $\mathcal{R}(\theta_0) = 0.5$. We indicate with red dashed lines the curves $c = 1$, $\tilde{\eta}(0) = 2/\gamma$, and $c\tilde{\eta}(0) = \ln(1/\gamma)$, and we display the average over 10 independent trials.

Appendix E. Formal statements and proofs for Section 4

E.1. Formal statements

Theorem 20 *Let Assumptions 1 and 2 hold. Let θ_0^p and $\theta_{1/2}^p$ be the solutions obtained with Algorithm 1 with $\tilde{\eta}(t)$ given by (6) for $\alpha = 0$ and $\alpha = 1/2$, respectively. Consider the setting*

$$\gamma = \frac{d}{n} = o_\gamma(1), \qquad \frac{\ln^2(1/\gamma)\gamma}{\lambda_{\min}^2}\left(\lambda_{\max} + \frac{\gamma}{\rho^2}\right) = o_\gamma(1), \tag{96}$$

and pick

$$c = O_\gamma(1), \qquad \tilde{\eta}(0)c = \frac{C\ln(1/\gamma)}{\lambda_{\min}}, \qquad \tilde{\eta}(0) \le \frac{2}{\gamma}, \tag{97}$$

for a large enough constant C which does not depend on γ, ρ, Σ. Then, we have that, with overwhelming probability,

$$\begin{aligned} \mathcal{R}(\theta_0^p) &= O_\gamma\left(\frac{\lambda_{\max}}{\lambda_{\min}^2}\gamma\ln(1/\gamma) + \frac{1}{\lambda_{\min}^2}\frac{\gamma^2\ln^2(1/\gamma)}{\rho^2}\right), \\ \mathcal{R}(\theta_{1/2}^p) &= O_\gamma\left(\frac{\lambda_{\max}}{\lambda_{\min}^2}\gamma\ln^{2/3}(1/\gamma) + \frac{1}{\lambda_{\min}^2}\frac{\gamma^2\ln^{4/3}(1/\gamma)}{\rho^2}\right). \end{aligned} \tag{98}$$

Furthermore, assume that $\rho = \Omega_\gamma(\gamma^{1-h})$ for some $h > 0$. Then, for any choice of the hyperparameters c and $\tilde{\eta}(0)$ s.t. $\tilde{\eta}(0) < 2/\gamma$, we have that

$$\begin{aligned} \mathcal{R}(\theta_0^p) &= \Omega_\gamma\left(\frac{1}{\lambda_{\max}^2}\gamma\ln(1/\gamma) + \frac{1}{\lambda_{\max}^2}\frac{\gamma^2\ln^2(1/\gamma)}{\rho^2}\right), \\ \mathcal{R}(\theta_{1/2}^p) &= \Omega_\gamma\left(\frac{1}{\lambda_{\max}^2}\gamma\ln^{2/3}(1/\gamma) + \frac{1}{\lambda_{\max}^2}\frac{\gamma^2\ln^{4/3}(1/\gamma)}{\rho^2}\right). \end{aligned} \tag{99}$$

Theorem 20 (proved later in this appendix) gives upper and lower bounds for output perturbation ($\alpha = 0$) and DP-SGD with constant noise ($\alpha = 1/2$). Our analysis requires that neither the condition number κ nor the inverse of the privacy parameter $1/\rho$ are too large with respect to $1/\gamma$ (see the second condition in (96)), and the condition $\rho = \Omega_\gamma(\gamma^{1-h})$ needed for the lower bound in (99) is qualitatively similar. We highlight that for both values of α, if $\lambda_{\max}, \lambda_{\min} = \Theta_\gamma(1)$, then upper and lower bounds match. This has two remarkable consequences: *(i)* the hyper-parameters in (97) are optimal in terms of rate (assuming $\tilde{\eta}(0) < 2/\gamma$), and *(ii)* DP-SGD outperforms output perturbation.

We now comment on the optimal hyper-parameter choice in (97). First, the condition $\tilde{\eta}(0) \leq 2/\gamma$ implies that $\eta_k = \bar{\eta}_k$ for all $k \in [n]$ (see Lemma 16), i.e., the adaptive step on the learning rate in Algorithm 1 never happens. If that was not the case, the gradient update would be proportional to $\bar{\eta}_k < \eta_k$, while the private noise σ_k still depends on η_k via (2). This suggests the sub-optimality of having $\bar{\eta}_k < \eta_k$ and of the regime $\tilde{\eta}(0) > 2/\gamma$. Second, the choice $\tilde{\eta}(0)c = C\ln(1/\gamma)$ provides the optimal trade-off between two competing objectives: on the one hand, $\tilde{\eta}(t)c$ controls the size of the first term of the ODEs in (42) (for $c = O_\gamma(1)$ and bounded values of R, Lemma 12 gives that $\mu_c(R)/c$ is lower bounded by a constant), which in turn determines the speed of convergence towards 0 of the risk; on the other hand, a large product $\tilde{\eta}(t)c$ increases at least one of the last two terms of the ODEs, which have the opposite effect of increasing the risk. We remark that this scaling agrees with the empirical practice of using a small clipping constant, with a learning rate renormalized by its value [19, 42]. Third, the choice $c = O_\gamma(1)$ is motivated by the fact that increasing c beyond this point does not further increase $\mu_c(R) < 1$, which drives the risk to 0. However, larger values of c increase the private noise in DP-SGD, and hence the last term in the RHSs of (42), which increases the risk. Formally, it can be shown that, if $\rho/c = \Omega_\gamma(\gamma^{1-h})$ for some $h > 0$, then the lower bounds in (99) increase by a factor $\max\left(1, c^2\right)$, thus demonstrating the sub-optimality of the choice $c = \omega_\gamma(1)$.

These conclusions are supported by Figure 3: performance deteriorates if either c exceeds 1, (upper part of the heatmaps) or $\tilde{\eta}(0)$ exceeds $2/\gamma$ (right part of the heatmaps); the lowest values of the risk are roughly parallel to the line $c\tilde{\eta}(0) = \ln(1/\gamma)$. Combining the upper bound on $\tilde{\eta}(0)$ with this last condition gives the lower bound $c = \Omega(\gamma\ln(1/\gamma))$. This still allows for a wide range of aggressive clipping regimes s.t. $c = o_\gamma(1)$. We also remark that the lower bound on the optimal c comes from considering one-pass DP-SGD, and it may not hold in other settings (e.g., full batch DP-GD [9]).

Theorem 21 *Let Assumptions 1 and 2 hold, and let θ^p_α be the solution obtained with Algorithm 1, with $\tilde{\eta}(t)$ given by (6) for $\alpha \geq 1$. Consider the setting*

$$\gamma = \frac{d}{n} = o_\gamma(1), \qquad \frac{\ln^2(1/\gamma)\gamma}{\lambda_{\min}^2}\left(\lambda_{\max}\alpha + \frac{\gamma}{\rho^2}\right) = o_\gamma(1), \tag{100}$$

and pick

$$c = O_\gamma(1), \qquad \tilde{\eta}(0)c = \frac{C\alpha\ln(1/\gamma)}{\lambda_{\min}}, \qquad \tilde{\eta}(0) \leq \frac{2}{\gamma}, \tag{101}$$

for a large enough constant C which does not depend on $\gamma, \rho, \Sigma, \alpha$. Then, we have that, with overwhelming probability,

$$\mathcal{R}(\theta^p_\alpha) = O_\gamma\left(\frac{\lambda_{\max}}{\lambda_{\min}^2}\alpha\gamma\ln^{\frac{1}{1+\alpha}}(1/\gamma) + \frac{1}{\lambda_{\min}^2}\frac{\alpha^2\gamma^2\ln^{\frac{2}{1+\alpha}}(1/\gamma)}{\rho^2}\right). \tag{102}$$

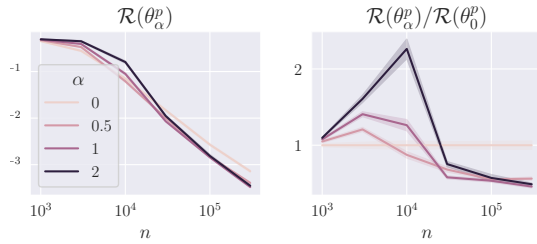

Figure 4: Numerical simulations for $\mathcal{R}(\theta^p_\alpha)$ obtained via Algorithm 1 for $d = 1000$, $\kappa = 1$ and $\rho = 1$, as a function of n. We consider the schedules in (6) for $\alpha \in \{0, 0.5, 1, 2\}$, we optimize w.r.t. c and $\tilde{\eta}(0)$, and we report the average over 10 independent trials, as well as the confidence interval corresponding to 1 standard deviation.

An immediate consequence of Theorem 21 (proved later in this section) is that, by taking $\alpha = \ln\ln(1/\gamma)$, in the setting where $\lambda_{\max}, \lambda_{\min} = \Theta_\gamma(1)$, we have

$$\mathcal{R}(\theta^p) = O_\gamma\left(\gamma(\ln\ln(1/\gamma)) + \frac{\gamma^2}{\rho^2}(\ln\ln(1/\gamma))^2\right). \tag{103}$$

Thus, for small γ, it is convenient to increase α and decay the noise faster during training, up to a level $\alpha = \Theta_\gamma(\ln\ln(1/\gamma))$. Values of α larger than that may then deteriorate performance. Figure 4 investigates the phenomenon by comparing different schedules after the hyper-parameters $\tilde{\eta}(0)$ and c have been optimized numerically. The left panel shows that, while all schedules rapidly give better results as n increases, larger values of α are optimal only for large enough n. This effect is clearly shown in the right panel, which reports the same results normalized by the loss of output perturbation ($\alpha = 0$). Hence, if n is sufficiently small compared to d, output perturbation can in fact be better than DP-SGD, in agreement with an observation made in [23] regarding the comparison between output and objective perturbation.

E.2. An auxiliary bound

Lemma 22 *Let $\beta < 1$ and $E_\beta(x) : \mathbb{R}^+ \to \mathbb{R}^+$ be the exponential integral function defined as*

$$E_\beta(x) := \int_1^{+\infty} \frac{e^{-xt}}{t^\beta}\,\mathrm{d}t. \tag{104}$$

Then, we have that

$$\lim_{x\to 0^+} x^{1-\beta} E_\beta(x) = \Gamma(1-\beta), \tag{105}$$

where $\Gamma(\cdot)$ denotes the Euler Gamma function

$$\Gamma(s) := \int_0^\infty e^{-t} t^{s-1}\,\mathrm{d}t. \tag{106}$$

Furthermore, we have that, for all $x \geq 0$,

$$E_\beta(x) \leq \Gamma(1-\beta) x^{-1+\beta}. \tag{107}$$

Finally, if either $\beta \geq 0$ or $x \geq -2\beta$, we also have that

$$E_\beta(x) \leq \frac{2e^{-x}}{x}. \tag{108}$$

Proof The change of variable $u = xt$ in the definition of $E_\beta(x)$ yields

$$\begin{aligned} E_\beta(x) &= x^{\beta-1} \int_x^{+\infty} \frac{e^{-u}}{u^\beta} \, \mathrm{d}u \\ &= x^{\beta-1} \left(\int_0^{+\infty} e^{-u} u^{-\beta} \, \mathrm{d}u - \int_0^x e^{-u} u^{-\beta} \, \mathrm{d}u \right) \\ &= x^{\beta-1} \left(\Gamma(1-\beta) - \int_0^x e^{-u} u^{-\beta} \, \mathrm{d}u \right). \end{aligned} \tag{109}$$

Then, (105) and (107) readily follow from the fact that the last term in the equation above is bounded by

$$0 \le \int_0^x e^{-u} u^{-\beta} \, \mathrm{d}u \le \int_0^x u^{-\beta} \, \mathrm{d}u = \frac{x^{1-\beta}}{1-\beta}. \tag{110}$$

For the upper bound, denoting with

$$\Gamma(s, x) := \int_x^\infty e^{-t} t^{s-1} \, \mathrm{d}t \tag{111}$$

the upper incomplete Euler gamma function, (109) allows us to write

$$E_\beta(x) = \frac{1}{x^{1-\beta}} \Gamma(1-\beta, x). \tag{112}$$

Via an integration by parts, we have

$$\Gamma(1-\beta, x) = \int_x^\infty e^{-t} t^{-\beta} \, \mathrm{d}t = e^{-x} x^{-\beta} - \beta \int_x^\infty e^{-t} t^{-\beta-1} \, \mathrm{d}t. \tag{113}$$

If $\beta \ge 0$, we have $\Gamma(1-\beta, x) \le e^{-x} x^{-\beta}$, which together with (112) gives (108). If $\beta < 0$, the second term in the equation above is positive, and since $t \ge x$, it can be upper bounded as

$$-\beta \int_x^\infty e^{-t} t^{-\beta-1} \, \mathrm{d}t \le -\frac{\beta}{x} \int_x^\infty e^{-t} t^{-\beta} \, \mathrm{d}t = -\frac{\beta}{x} \Gamma(1-\beta, x), \tag{114}$$

which, if plugged in (113), for $x \ge -2\beta$ gives

$$\Gamma(1-\beta, x) \le \frac{e^{-x} x^{-\beta}}{1 + \beta/x} \le 2 e^{-x} x^{-\beta}, \tag{115}$$

and the thesis again follows from (112). ■

E.3. Proof of Theorem 20

All the ODEs defined in this (and the following) section will be such that their RHS is uniformly Lipschitz in the dependent variable at all times $t \in [0, 1]$, which in turn guarantees they have a unique solution. Furthermore, the RHSs will also be uniformly Lipschitz with respect to the variable t due to Assumption 2. Thus, if $R(0) = R'(0)$, and

$$\mathrm{d}R = f(t, R), \qquad \mathrm{d}R' = f'(t, R'), \tag{116}$$

with

$$f'(t, R'(t)) \geq f(t, R'(t)), \tag{117}$$

for all $t \in [0, 1]$, we have that

$$R'(t) \geq R(t) \text{ for all } t \in [0, 1]. \tag{118}$$

The same statement holds also for the opposite inequality, and will be used extensively to bound the solutions of $R(t)$ for different schedules. This is a direct application of Theorem 1.3 in [56]

To ease the presentation, we introduce the notation $v = \tilde{\eta}(0)$. We will provide the proof separately for $\alpha = 0$ and $\alpha = 1/2$, and the proof for $\alpha \geq 1$ in the next section. We will also denote the test risk at initialization $R(0) = \left\|\Sigma^{1/2}\theta^*\right\|_2^2/2$, and all asymptotic notations will be with respect to the limit $\gamma \to 0$.

$\alpha = 0$**: output perturbation.** Recall that in the setting $\alpha = 0$, we have

$$\begin{aligned} \mathrm{d}\overline{R}(t) &= -2\lambda_{\min} vc \frac{\mu_c(\overline{R})}{c}\overline{R}\mathrm{d}t + \lambda_{\max}(vc)^2 \frac{\nu_c(\overline{R})(\overline{R} + \zeta^2/2)}{c^2}\gamma \mathrm{d}t \\ \mathrm{d}\underline{R}(t) &= -2\lambda_{\max} vc \frac{\mu_c(\underline{R})}{c}\underline{R}\mathrm{d}t + (vc)^2 \frac{\nu_c(\underline{R})(\underline{R} + \zeta^2/2)}{c^2}\gamma \mathrm{d}t \end{aligned} \tag{119}$$

Importantly, recall that the risk $\mathcal{R}(\theta^p)$ in this setting is not well approximated by $R(1)$, due to Proposition 5.

Theorem 23 *Let Assumptions 1 and 2 hold, and let θ^p be the solution obtained with Algorithm 1, with the schedule defined in (6) for $\alpha = 0$, in the setting $\gamma = d/n \to 0$. Furthermore, assume*

$$\frac{\lambda_{\max}}{\lambda_{\min}^2} \ln(1/\gamma)\gamma = o(1). \tag{120}$$

Then, by setting

$$c = O(1), \qquad vc = \frac{C \ln(1/\gamma)}{\lambda_{\min}}, \qquad v \leq 2/\gamma, \tag{121}$$

for a large enough constant C which does not depend on γ, ρ, Σ, we have that, with overwhelming probability,

$$\mathcal{R}(\theta^p) = O\left(\frac{\lambda_{\max}\gamma \ln(1/\gamma)}{\lambda_{\min}^2} + \frac{\gamma^2 \ln^2(1/\gamma)}{\rho^2 \lambda_{\min}^2}\right). \tag{122}$$

Furthermore, suppose there exists $h > 0$ such that $\rho = \Omega\left(\gamma^{1-h}\right)$. Then, for any choice of the hyper-parameters c and v such that $v \leq 2/\gamma$, we have that

$$\mathcal{R}(\theta^p) = \Omega\left(\frac{\gamma \ln(1/\gamma)}{\lambda_{\max}^2} + \frac{\gamma^2 \ln^2(1/\gamma)}{\rho^2 \lambda_{\max}^2}\right). \tag{123}$$

Proof Let us introduce the shorthands

$$\begin{aligned} \overline{f}(t, \overline{R}) &= -2\lambda_{\min} vc \frac{\mu_c(\overline{R})}{c}\overline{R}\mathrm{d}t + \lambda_{\max}(vc)^2 \frac{\nu_c(\overline{R})(\overline{R} + \zeta^2/2)}{c^2}\gamma \mathrm{d}t, \\ \underline{f}(t, \underline{R}) &= -2\lambda_{\max} vc \frac{\mu_c(\underline{R})}{c}\underline{R}\mathrm{d}t + \tilde{\eta}^2(t)(vc)^2 \frac{\nu_c(\underline{R})(\underline{R} + \zeta^2/2)}{c^2}\gamma \mathrm{d}t, \end{aligned} \tag{124}$$

corresponding to the RHSs of the ODEs of interest. Then, consider the auxiliary ODEs

$$\mathrm{d}\overline{R}' = -\overline{a}\overline{R}'\mathrm{d}t + \overline{b}\mathrm{d}t =: \overline{f}'(t, \overline{R}')\mathrm{d}t, \qquad \mathrm{d}\underline{R}' = -\underline{a}\,\underline{R}'\mathrm{d}t + \underline{b}\mathrm{d}t =: \underline{f}'(t, \underline{R}')\mathrm{d}t, \qquad \overline{a}, \overline{b}, \underline{a}, \underline{b} > 0, \tag{125}$$

with initial conditions $\overline{R}'(0) = \overline{R}(0) = R(0) = \underline{R}(0) = \underline{R}'(0)$.

Notice that $\overline{R}'(t)$ admits the closed form solution

$$\overline{R}'(t) = \left(R(0) - \overline{b}/\overline{a}\right) e^{-\overline{a}t} + \overline{b}/\overline{a}. \tag{126}$$

Similarly, we have that

$$\underline{R}'(t) = \left(R(0) - \underline{b}/\underline{a}\right) e^{-\underline{a}t} + \underline{b}/\underline{a}. \tag{127}$$

Since $c = O(1)$, by Lemma 12, we have that

$$\frac{\underline{c}_\mu(c,\zeta)}{\sqrt{2R+\zeta^2}} < \frac{\mu_c(R)}{c} < \frac{1}{\sqrt{\pi\left(R+\zeta^2/2\right)}}, \qquad \underline{c}_\nu(c,\zeta) < \frac{2\nu_c(R)\left(R+\zeta^2/2\right)}{c^2} < 1. \tag{128}$$

Then, let us set

$$\overline{a} = 2\lambda_{\min} vc \frac{\underline{c}_\mu(c,\zeta)}{\sqrt{R(0)+1+\zeta^2}}, \quad \overline{b} = \lambda_{\max}\frac{v^2c^2\gamma}{2}, \quad \underline{a} = 2vc\frac{\lambda_{\max}}{\sqrt{\pi\zeta^2/2}}, \quad \underline{b} = \frac{v^2c^2\gamma\underline{c}_\nu(c,\zeta)}{2}. \tag{129}$$

As $cv = C\ln(1/\gamma)/\lambda_{\min}$, the choice in (129) ensures that $\underline{b}/\underline{a} \le 1$ and $\overline{b}/\overline{a} \le 1$ as long as $\lambda_{\max}/\lambda_{\min}^2 \ln(1/\gamma)\gamma = o(1)$, which in turn guarantees that

$$\overline{R}'(t) \in [0, R(0)+1], \qquad \underline{R}'(t) \in [0, R(0)+1]. \tag{130}$$

Thus, (129) guarantees

$$\overline{f}(t, \overline{R}'(t)) < \overline{f}'(t, \overline{R}'(t)), \qquad \underline{f}(t, \underline{R}'(t)) > \underline{f}'(t, \underline{R}'(t)), \qquad \text{for all } t \in [0,1]. \tag{131}$$

Thus, by (118), we have

$$\underline{R}'(t) < \underline{R}(t), \qquad \overline{R}'(t) > \overline{R}(t) \qquad \text{for all } t \in (0,1]. \tag{132}$$

In particular, plugging $vc = C\ln(1/\gamma)/\lambda_{\min}$ and (129) in (126), we have that

$$\overline{R}(1) < \overline{R}'(1) < e^{-\overline{a}} + \frac{\overline{b}}{\overline{a}} = O\left(\frac{\lambda_{\max}}{\lambda_{\min}^2}\ln(1/\gamma)\gamma\right), \tag{133}$$

as long as C is chosen to be large enough. Then, the upper bounds follows from Proposition 13 and Proposition 5, as the risk at the last iterate roughly increases by $2c^2v^2\gamma^2/\rho^2 = O(\gamma^2\ln^2(1/\gamma)/(\rho^2\lambda_{\min}^2))$.

For the lower bound, let us first suppose $c \le 1$, which implies that (128) holds. Pick $\underline{a}, \underline{b}$ as in (129).

First, let us consider the case $\underline{a} \le 1$. We have that (132) gives

$$\underline{R}(1) > \underline{R}'(1) = R(0)e^{-\underline{a}} + \frac{\underline{b}}{\underline{a}} - \frac{\underline{b}}{\underline{a}}e^{-\underline{a}} > R(0)e^{-\underline{a}} \ge R(0)/e = \Omega(1). \tag{134}$$

In the case $\underline{a} > 1$, (132) gives

$$\underline{R}(1) > \underline{R}'(1) = R(0)e^{-\underline{a}} + \frac{\underline{b}}{\underline{a}} - \frac{\underline{b}}{\underline{a}}e^{-\underline{a}} > R(0)e^{-\underline{a}} + \frac{\underline{b}}{\underline{a}}(1-e^{-1}) = R(0)\left(e^{-vc\lambda_{\max}a} + \frac{vcb\gamma}{\lambda_{\max}a}\right), \tag{135}$$

where a, b are positive constants which do not depend on γ, v, c and the spectrum of Σ. Then, let us consider the following quantity

$$\underline{\mathcal{R}} := R(0)\left(e^{-vc\lambda_{\max}a} + \frac{vcb\gamma}{\lambda_{\max}a}\right) + 2v^2c^2\frac{\gamma^2}{\rho^2}. \tag{136}$$

We have that

$$\arg\min_{vc} e^{-vc\lambda_{\max}a} + \frac{vcb\gamma}{\lambda_{\max}a} = \frac{1}{\lambda_{\max}a}\ln\left(\frac{(\lambda_{\max}a)^2}{b\gamma}\right), \tag{137}$$

which implies

$$\underline{\mathcal{R}} > R(0)\left(e^{-vc\lambda_{\max}a} + \frac{vcb\gamma}{\lambda_{\max}a}\right) = \Omega\left(\frac{\gamma\ln(1/\gamma)}{\lambda_{\max}^2}\right). \tag{138}$$

Note that, assuming $\rho = \Omega(\gamma^{1-h})$, we have

$$\begin{aligned}
&\min_{vc\geq\frac{\ln(\rho^2/\gamma^2)}{2a\lambda_{\max}}} e^{-vc\lambda_{\max}a} + v^2c^2\frac{\gamma^2}{\rho^2} \geq \min_{vc\geq\frac{\ln(\rho^2/\gamma^2)}{2a\lambda_{\max}}} v^2c^2\frac{\gamma^2}{\rho^2} = \Omega\left(\frac{\gamma^2\ln^2(1/\gamma)}{\rho^2\lambda_{\max}^2}\right),\\
&\min_{0\leq vc\leq\frac{\ln(\rho^2/\gamma^2)}{2a\lambda_{\max}}} e^{-vc\lambda_{\max}a} + v^2c^2\frac{\gamma^2}{\rho^2} \geq \frac{\gamma}{\rho} = \Omega\left(\frac{\gamma^2\ln^2(1/\gamma)}{\rho^2}\right) = \Omega\left(\frac{\gamma^2\ln^2(1/\gamma)}{\rho^2\lambda_{\max}^2}\right).
\end{aligned} \tag{139}$$

Then, merging (136), (138) and (139) yields

$$\underline{\mathcal{R}} = \Omega\left(\frac{\gamma\ln(1/\gamma)}{\lambda_{\max}^2} + \frac{\gamma^2\ln^2(1/\gamma)}{\rho^2\lambda_{\max}^2}\right). \tag{140}$$

The result in (135) together with Proposition 13 and Proposition 5 implies that, with overwhelming probability,

$$\mathcal{R}(\theta^p) = \Omega\left(\frac{\gamma\ln(1/\gamma)}{\lambda_{\max}^2} + \frac{\gamma^2\ln^2(1/\gamma)}{\rho^2\lambda_{\max}^2}\right). \tag{141}$$

It remains to prove the lower bound for $c > 1$. From (18), one readily has that $0 < \mu_c(R) < 1$. Note that

$$\nu_c(R)(R+\zeta^2/2) \geq \max\left(\frac{\nu_c(R)(R+\zeta^2/2)}{c^2}, \nu_c(R)\zeta^2/2\right), \tag{142}$$

which combined with (34) gives that $\nu_c(R)(R+\zeta^2/2) > b_1$, for a positive constant b_1 independent of γ, v, c and the spectrum of Σ. Furthermore, (32) implies that $\nu_c(R)(R+\zeta^2/2) < c^2/2$. Thus, the solution of $\underline{R}$ is lower bounded by that of the ODE below:

$$\mathrm{d}\underline{R}'' = -2\lambda_{\max}v\underline{R}''\mathrm{d}t + v^2b_1\gamma\mathrm{d}t, \tag{143}$$

with initial condition $\underline{R}''(0) = R(0)$. Thus, following the same steps we used to achieve (140), it can be shown that, as $c \geq 1$, we have, with overwhelming probability,

$$\mathcal{R}(\theta^p) = \Omega\left(\frac{\gamma\ln(1/\gamma)}{\lambda_{\max}^2} + \frac{\gamma^2\ln^2(1/\gamma)}{\rho^2\lambda_{\max}^2}\right), \tag{144}$$

which gives the desired result.

■

DP-SGD with constant noise. In the setting $\alpha = 1/2$, (6) gives

$$\begin{aligned}
\mathrm{d}\overline{R}(t) &= -2\lambda_{\min} vc\sqrt{1-t}\frac{\mu_c(\overline{R})}{c}\overline{R}\mathrm{d}t + \lambda_{\max}(vc)^2(1-t)\frac{\nu_c(\overline{R})(\overline{R}+\zeta^2/2)}{c^2}\gamma\mathrm{d}t + 2(vc)^2\frac{\gamma^2}{\rho^2}\mathrm{d}t,\\
\mathrm{d}\underline{R}(t) &= -2\lambda_{\max} vc\sqrt{1-t}\frac{\mu_c(\underline{R})}{c}\underline{R}\mathrm{d}t + (vc)^2(1-t)\frac{\nu_c(\underline{R})(\underline{R}+\zeta^2/2)}{c^2}\gamma\mathrm{d}t + 2(vc)^2\frac{\gamma^2}{\rho^2}\mathrm{d}t.
\end{aligned} \tag{145}$$

Importantly, recall that the risk $\mathcal{R}(\theta^p)$ in this setting is well approximated by $R(1)$, due to Proposition 5, since $\tilde{\eta}(1) = 0$.

Theorem 24 *Let Assumptions 1 and 2 hold, and let θ^p be the solution obtained with Algorithm 1, with the schedule defined in (6) for $\alpha = 1/2$, in the setting $\gamma = d/n \to 0$. Furthermore, assume*

$$\frac{\ln^2(1/\gamma)\gamma}{\lambda_{\min}^2}\left(\lambda_{\max} + \frac{\gamma}{\rho^2}\right) = o(1). \tag{146}$$

Then, by setting

$$c = O(1), \qquad vc = \frac{C\ln(1/\gamma)}{\lambda_{\min}}, \qquad v \le 2/\gamma, \tag{147}$$

for a large enough constant C which does not depend on γ, ρ, Σ, we have that, with overwhelming probability,

$$\mathcal{R}(\theta^p) = O\left(\frac{\lambda_{\max}\gamma\ln^{2/3}(1/\gamma)}{\lambda_{\min}^2} + \frac{\gamma^2\ln^{4/3}(1/\gamma)}{\rho^2\lambda_{\min}^2}\right). \tag{148}$$

Furthermore, suppose there exists $h > 0$ such that $\rho = \Omega\left(\gamma^{1-h}\right)$. Then, for any choice of the hyper-parameters c and v such that $v \le 2/\gamma$, we have that

$$\mathcal{R}(\theta^p) = \Omega\left(\frac{\gamma\ln^{2/3}(1/\gamma)}{\lambda_{\max}^2} + \frac{\gamma^2\ln^{4/3}(1/\gamma)}{\rho^2\lambda_{\max}^2}\right). \tag{149}$$

Proof

As before, let us introduce the shorthands

$$\begin{aligned}
\overline{f}(t,\overline{R}) &= -2\lambda_{\min} vc\sqrt{1-t}\frac{\mu_c(\overline{R})}{c}\overline{R} + \lambda_{\max}(vc)^2(1-t)\frac{\nu_c(\overline{R})(\overline{R}+\zeta^2/2)}{c^2}\gamma + 2(vc)^2\frac{\gamma^2}{\rho^2},\\
\underline{f}(t,\underline{R}) &= -2\lambda_{\max} vc\sqrt{1-t}\frac{\mu_c(\underline{R})}{c}\underline{R} + (vc)^2(1-t)\frac{\nu_c(\underline{R})(\underline{R}+\zeta^2/2)}{c^2}\gamma + 2(vc)^2\frac{\gamma^2}{\rho^2}.
\end{aligned} \tag{150}$$

Then, consider the auxiliary ODEs

$$\begin{aligned}
\mathrm{d}\overline{R}' &= -\overline{a}\sqrt{1-t}\overline{R}'\mathrm{d}t + \overline{b}_1(1-t)\mathrm{d}t + \overline{b}_2\mathrm{d}t =: \overline{f}'(t,\overline{R})\mathrm{d}t, \qquad \overline{a},\overline{b}_1,\overline{b}_2 > 0,\\
\mathrm{d}\underline{R}' &= -\underline{a}\sqrt{1-t}\underline{R}'\mathrm{d}t + \underline{b}_1(1-t)\mathrm{d}t + \underline{b}_2\mathrm{d}t =: \underline{f}'(t,\underline{R})\mathrm{d}t, \qquad \underline{a},\underline{b}_1,\underline{b}_2 > 0,
\end{aligned} \tag{151}$$

with initial conditions $\overline{R}'(0) = \overline{R}(0) = R(0) = \underline{R}(0) = \underline{R}'(0)$, which admit the closed-form solutions

$$\begin{aligned}\overline{R}'(t) = \frac{R(0)}{3} e^{-2\overline{a}(1-(1-t)^{3/2})/3} \Bigg(3 + \\ - \frac{2}{R(0)} \overline{b}_1 e^{2\overline{a}/3} \left(E_{-1/3}\left(\frac{2\overline{a}}{3}\right) - (1-t)^2 E_{-1/3}\left(\frac{2\overline{a}(1-t)^{3/2}}{3}\right)\right) \\ - \frac{2}{R(0)} \overline{b}_2 e^{2\overline{a}/3} \left(E_{1/3}\left(\frac{2\overline{a}}{3}\right) - (1-t) E_{1/3}\left(\frac{2\overline{a}(1-t)^{3/2}}{3}\right)\right)\Bigg),\end{aligned} \tag{152}$$

$$\begin{aligned}\underline{R}'(t) = \frac{R(0)}{3} e^{-2\underline{a}(1-(1-t)^{3/2})/3} \Bigg(3 + \\ - \frac{2}{R(0)} \underline{b}_1 e^{2\underline{a}/3} \left(E_{-1/3}\left(\frac{2\underline{a}}{3}\right) - (1-t)^2 E_{-1/3}\left(\frac{2\underline{a}(1-t)^{3/2}}{3}\right)\right) \\ - \frac{2}{R(0)} \underline{b}_2 e^{2\underline{a}/3} \left(E_{1/3}\left(\frac{2\underline{a}}{3}\right) - (1-t) E_{1/3}\left(\frac{2\underline{a}(1-t)^{3/2}}{3}\right)\right)\Bigg),\end{aligned} \tag{153}$$

expressed in terms of the exponential integral functions $E_{-1/3}(\cdot)$, $E_{1/3}(\cdot)$ defined in (104).

Note that, for $t \in [0,1]$, $\overline{R}'(t) \geq 0$. In fact, if this is not the case, by continuity of $\overline{R}'$, there exists an interval $(t^*, t^* + \delta) \subseteq [0,1]$ s.t. $\overline{R}'(t^*) = 0$ and $\overline{R}'(t) < 0$ for all $t \in (t^*, t^* + \delta)$. However, if $\overline{R}'(t^*) = 0$, then the derivative of $\overline{R}'$ evaluated at t^* is $\geq b_2 > 0$ by (151), which is a contradiction. A similar argument gives that, for $t \in [0,1]$, $\underline{R}'(t) \geq 0$.

Next, we upper bound $\overline{R}'(t)$ as

$$\begin{aligned}\overline{R}'(t) \leq R(0) + \frac{2}{3}\overline{b}_1 e^{2\overline{a}(1-t)^{3/2}/3}(1-t)^2 E_{-1/3}\left(\frac{2\overline{a}(1-t)^{3/2}}{3}\right) \\ + \frac{2}{3}\overline{b}_2 e^{2\overline{a}(1-t)^{3/2}/3}(1-t) E_{1/3}\left(\frac{2\overline{a}(1-t)^{3/2}}{3}\right).\end{aligned} \tag{154}$$

If $2\overline{a}(1-t)^{3/2}/3 \geq 2/3$, an application of (108) gives that

$$\begin{aligned} e^{2\overline{a}(1-t)^{3/2}/3}(1-t)^2 E_{-1/3}\left(\frac{2\overline{a}(1-t)^{3/2}}{3}\right) \leq \frac{3}{\overline{a}}\sqrt{1-t} \leq \frac{3}{\overline{a}}, \\ e^{2\overline{a}(1-t)^{3/2}/3}(1-t) E_{1/3}\left(\frac{2\overline{a}(1-t)^{3/2}}{3}\right) \leq \frac{3}{\overline{a}}(1-t)^{-1/2} \leq \frac{3}{\overline{a}^{2/3}}.\end{aligned} \tag{155}$$

Instead, if $2\overline{a}(1-t)^{3/2}/3 < 2/3$, by using (107) we have

$$\begin{aligned} e^{2\overline{a}(1-t)^{3/2}/3}(1-t)^2 E_{-1/3}\left(\frac{2\overline{a}(1-t)^{3/2}}{3}\right) \leq e^{2/3}\Gamma\left(\frac{4}{3}\right)\left(\frac{2\overline{a}}{3}\right)^{-4/3}, \\ e^{2\overline{a}(1-t)^{3/2}/3}(1-t) E_{1/3}\left(\frac{2\overline{a}(1-t)^{3/2}}{3}\right) \leq e^{2/3}\Gamma\left(\frac{2}{3}\right)\left(\frac{2\overline{a}}{3}\right)^{-2/3}.\end{aligned} \tag{156}$$

Thus, the following upper bound holds for $t \in [0,1]$:

$$\overline{R}'(t) \leq R(0) + \frac{2\bar{b}_1}{\overline{a}} + \frac{2\bar{b}_2}{\overline{a}^{2/3}} + e^{2/3}\left(\frac{2}{3}\right)^{-1/3}\Gamma\left(\frac{4}{3}\right)\frac{\bar{b}_1}{\overline{a}^{4/3}} + e^{2/3}\left(\frac{2}{3}\right)^{1/3}\Gamma\left(\frac{2}{3}\right)\frac{\bar{b}_2}{\overline{a}^{2/3}}. \quad (157)$$

Following the same passages, we have that, for $t \in [0,1]$,

$$\underline{R}'(t) \leq R(0) + \frac{2\underline{b}_1}{\underline{a}} + \frac{2\underline{b}_2}{\underline{a}^{2/3}} + e^{2/3}\left(\frac{2}{3}\right)^{-1/3}\Gamma\left(\frac{4}{3}\right)\frac{\underline{b}_1}{\underline{a}^{4/3}} + e^{2/3}\left(\frac{2}{3}\right)^{1/3}\Gamma\left(\frac{2}{3}\right)\frac{\underline{b}_2}{\underline{a}^{2/3}}. \quad (158)$$

Let us now pick

$$\begin{aligned} \overline{a} &= 2vc\lambda_{\min}\frac{\underline{c}_\mu(c,\zeta)}{\sqrt{R(0)+1+\zeta^2/2}}, \qquad \bar{b}_1 = \lambda_{\max}\frac{v^2c^2\gamma}{2}, \qquad \bar{b}_2 = 4v^2c^2\frac{\gamma^2}{\rho^2}, \\ \underline{a} &= 2vc\lambda_{\max}\frac{1}{\sqrt{\pi\zeta^2/2}}, \qquad \underline{b}_1 = \frac{v^2c^2\gamma\underline{c}_\nu(c,\zeta)}{2}, \qquad \underline{b}_2 = v^2c^2\frac{\gamma^2}{\rho^2}. \end{aligned} \quad (159)$$

Since $c = O(1)$, by Lemma 12, we have that (128) holds. As $cv = C\ln(1/\gamma)/\lambda_{\min}$, the choice in (159) ensures that $0 \leq \underline{R}'(t) \leq \overline{R}'(t) \leq 1 + R(0)$ for $t \in [0,1]$, as long as we have $\bar{b}_1 + \bar{b}_2 = o(1)$, which holds due to (146).

Thus, (159) guarantees

$$\overline{f}(t,\overline{R}'(t)) < \overline{f}'(t,\overline{R}'(t)), \qquad \underline{f}(t,\underline{R}'(t)) > \underline{f}'(t,\underline{R}'(t)), \qquad \text{for all } t \in [0,1]. \quad (160)$$

Note that $\overline{f}(t,R)$ and $\underline{f}(t,R)$ are continuous in both variables in the intervals $t \in [0,1]$ and $R \in [0,1]$. Furthermore, they are also Lipschitz in R in these same intervals. Thus, (118) gives

$$\underline{R}'(t) < \underline{R}(t), \qquad \overline{R}(t) < \overline{R}'(t), \qquad \text{for all } t \in (0,1]. \quad (161)$$

To prove the upper bound, due to Propositions 13 and 5, it suffices to give an upper bound on $\overline{R}'(1)$. To this aim, note that Lemma 22 yields

$$\begin{aligned} \lim_{x\to 0} x^2 E_{-1/3}\left(\frac{2ax^{3/2}}{3}\right) &= \left(\frac{3}{2}\right)^{4/3}\Gamma\left(\frac{4}{3}\right)a^{-4/3}, \\ \lim_{x\to 0} x E_{1/3}\left(\frac{2ax^{3/2}}{3}\right) &= \left(\frac{3}{2}\right)^{2/3}\Gamma\left(\frac{2}{3}\right)a^{-2/3}, \end{aligned} \quad (162)$$

where $\Gamma(\cdot)$ denotes the Euler Gamma function (defined in (106)). Thus,

$$\begin{aligned} \overline{R}'(1) &= \frac{R(0)}{3}e^{-2\overline{a}/3}\Bigg(3 - \frac{2}{R(0)}\bar{b}_1 e^{2\overline{a}/3}\left(E_{-1/3}\left(\frac{2\overline{a}}{3}\right) - \left(\frac{3}{2}\right)^{4/3}\Gamma\left(\frac{4}{3}\right)\overline{a}^{-4/3}\right) \\ &\qquad - \frac{2}{R(0)}\bar{b}_2 e^{2\overline{a}/3}\left(E_{1/3}\left(\frac{2\overline{a}}{3}\right) - \left(\frac{3}{2}\right)^{2/3}\Gamma\left(\frac{2}{3}\right)\overline{a}^{-2/3}\right)\Bigg) \\ &\leq R(0)e^{-2\overline{a}/3} + \left(\frac{3}{2}\right)^{1/3}\Gamma\left(\frac{4}{3}\right)\bar{b}_1\overline{a}^{-4/3} + \left(\frac{3}{2}\right)^{-1/3}\Gamma\left(\frac{2}{3}\right)\bar{b}_2\overline{a}^{-2/3}, \end{aligned} \quad (163)$$

where in the last line we have used the non-negativity of the exponential integral functions. For C sufficiently large, due to (159), the first term in the RHS is $o(\gamma)$ (recall that $vc = C\ln(1/\gamma)/\lambda_{\min}$). The other two terms are $O(\lambda_{\max}\gamma \ln^{2/3}(1/\gamma)/\lambda_{\min}^2)$ and $O(\gamma^2 \ln^{4/3}(1/\gamma)/(\rho^2\lambda_{\min}^2)$ respectively. Note that $\overline{R}(1) < \overline{R}'(1)$ and $\tilde{\eta}(1) = 0$. Thus, the desired result follows from Propositions 13 and 5.

To prove the lower bound, due to Propositions 13 and 5, it suffices to show that the inequality in the thesis holds for $\underline{R}(1)$. Let us first suppose $c \leq 1$, which implies that (128) holds. Pick $\underline{a}, \underline{b}_1, \underline{b}_2$ as in (159), and consider the ODE $\underline{R}'(t)$ defined in (151) with the initial condition $\underline{R}'(0) = R(0)$, which is a lower bound on $\underline{R}(t)$ due to (161), *i.e.*,

$$\underline{R}(1) > \underline{R}'(1) = \frac{R(0)}{3}e^{-2\underline{a}/3}\Bigg(3 - \frac{2}{R(0)}\underline{b}_1 e^{2\underline{a}/3}\left(E_{-1/3}\left(\frac{2\underline{a}}{3}\right) - \left(\frac{3}{2}\right)^{4/3}\Gamma\left(\frac{4}{3}\right)\underline{a}^{-4/3}\right) - \frac{2}{R(0)}\underline{b}_2 e^{2\underline{a}/3}\left(E_{1/3}\left(\frac{2\underline{a}}{3}\right) - \left(\frac{3}{2}\right)^{2/3}\Gamma\left(\frac{2}{3}\right)\underline{a}^{-2/3}\right)\Bigg). \tag{164}$$

We consider two additional cases depending on the value of $\underline{a}$. If $\underline{a} \geq 2$, then (108) gives that

$$E_{-1/3}\left(\frac{2\underline{a}}{3}\right) \leq \frac{2e^{-2\underline{a}/3}}{\frac{2\underline{a}}{3}}, \qquad E_{1/3}\left(\frac{2\underline{a}}{3}\right) \leq \frac{2e^{-2\underline{a}/3}}{\frac{2\underline{a}}{3}}, \tag{165}$$

which implies that

$$\begin{aligned}\underline{R}(1) \geq R(0)e^{-2\underline{a}/3} &+ \left(\frac{3}{2}\right)^{1/3}\Gamma\left(\frac{4}{3}\right)\underline{b}_1\underline{a}^{-4/3} - 2e^{-2\underline{a}/3}\underline{b}_1\underline{a}^{-1} \\ &+ \left(\frac{3}{2}\right)^{-1/3}\Gamma\left(\frac{2}{3}\right)\underline{b}_2\underline{a}^{-2/3} - 2e^{-2\underline{a}/3}\underline{b}_2\underline{a}^{-1}.\end{aligned} \tag{166}$$

Note that

$$\begin{aligned}&\left(\frac{3}{2}\right)^{1/3}\Gamma\left(\frac{4}{3}\right)\underline{b}_1\underline{a}^{-4/3} - 2e^{-2\underline{a}/3}\underline{b}_1\underline{a}^{-1} \geq \underline{b}_1\underline{a}^{-4/3} - 2e^{-2\underline{a}/3}\underline{b}_1\underline{a}^{-1} \geq \frac{1}{3}\underline{b}_1\underline{a}^{-4/3}, \\ &\left(\frac{3}{2}\right)^{-1/3}\Gamma\left(\frac{2}{3}\right)\underline{b}_2\underline{a}^{-2/3} - 2e^{-2\underline{a}/3}\underline{b}_2\underline{a}^{-1} \geq \underline{b}_2\underline{a}^{-2/3} - 2e^{-2\underline{a}/3}\underline{b}_2\underline{a}^{-1} \geq \frac{1}{3}\underline{b}_2\underline{a}^{-2/3},\end{aligned} \tag{167}$$

where the inequalities on the right hold for $\underline{a} \geq 2$. Thus,

$$\begin{aligned}\underline{R}(1) \geq &\left(\frac{R(0)}{2}e^{-avc\lambda_{\max}} + \frac{b_1}{3}\lambda_{\max}^{-4/3}a^{-4/3}(vc)^{2/3}\gamma\right) \\ &+ \left(\frac{R(0)}{2}e^{-avc\lambda_{\max}} + \frac{b_2}{3}\lambda_{\max}^{-2/3}a^{-2/3}(vc)^{4/3}\frac{\gamma^2}{\rho^2}\right),\end{aligned} \tag{168}$$

where a, b_1, b_2 are positive constants which do not depend on γ, v, c, ρ or the spectrum of Σ.

Denoting with $a' = a\lambda_{\max}$, $\tau = vc$, we have that for a fixed $\beta \in \{2/3, 4/3\}$, and for any $\delta = o(1)$,

$$\min_{\tau \geq 0} e^{-\tau a'} + \frac{\tau^{2-\beta}\delta}{(a')^{\beta}} = \Omega\left(\frac{\delta \ln^{2-\beta}(1/\delta)}{(a')^2}\right), \tag{169}$$

which follows from the following calculations:

$$\begin{aligned}\min_{\tau\geq\ln(1/\delta)/(2a')} e^{-\tau a'} + \frac{\tau^{2-\beta}\delta}{(a')^\beta} &\geq \min_{\tau\geq\ln(1/\delta)/(2a')} \frac{\tau^{2-\beta}\delta}{(a')^\beta} = \Omega\left(\frac{\delta\ln^{2-\beta}(1/\delta)}{(a')^2}\right),\\ \min_{0\leq\tau\leq\ln(1/\delta)/(2a')} e^{-\tau a'} + \frac{\tau^{2-\beta}\delta}{(a')^\beta} &\geq \delta^{1/2} = \Omega(\delta\ln^{2-\beta}(1/\delta)) = \Omega\left(\frac{\delta\ln^{2-\beta}(1/\delta)}{(a')^2}\right).\end{aligned} \tag{170}$$

Then, since $\rho = \Omega(\gamma^{1-h})$, we can set $\delta = \gamma$ and $\delta = \gamma^2/\rho^2$ to obtain

$$\underline{R}(1) = \Omega\left(\frac{\gamma\ln^{2/3}(1/\gamma)}{\lambda_{\max}^2} + \frac{\gamma^2\ln^{4/3}(\rho^2/\gamma^2)}{\rho^2\lambda_{\max}^2}\right), \tag{171}$$

which implies the desired result (due to Propositions 13 and 5).

If $\underline{a} \leq 2$, then (107) gives that

$$E_{-1/3}\left(\frac{2\underline{a}}{3}\right) \leq \Gamma\left(\frac{4}{3}\right)\left(\frac{2\underline{a}}{3}\right)^{-4/3}, \qquad E_{1/3}\left(\frac{2\underline{a}}{3}\right) \leq \Gamma\left(\frac{2}{3}\right)\left(\frac{2\underline{a}}{3}\right)^{-2/3}, \tag{172}$$

which implies that

$$\underline{R}(1) \geq R(0)e^{-2\underline{a}/3} = \Omega(1), \tag{173}$$

thus again proving the desired claim (due to Propositions 13 and 5).

Finally, for $c > 1$, due to the same argument used to show (142), the solution of the original ODE is lower bounded by that of the ODE below:

$$\mathrm{d}\underline{R}'' = -2\lambda_{\max}v\sqrt{1-t}\underline{R}''\mathrm{d}t + v^2 b_1\gamma(1-t)\mathrm{d}t + 2v^2c^2\frac{\gamma^2}{\rho^2}\mathrm{d}t, \tag{174}$$

with initial condition $\underline{R}''(0) = R(0)$, where b_1 is a positive constant independent of γ, v, c, ρ and the spectrum of Σ. Thus, following the same steps above with $\underline{a} = 2v\lambda_{\max}, \underline{b}_1 = v^2 b_1\gamma, \underline{b}_2 = 2v^2c^2\gamma^2/\rho^2$, one readily shows that

$$\underline{R}''(1) = \Omega\left(\frac{\gamma\ln^{2/3}(1/\gamma)}{\lambda_{\max}^2} + \frac{\gamma^2\ln^{4/3}(1/\gamma)}{\rho^2\lambda_{\max}^2}\right), \tag{175}$$

concluding the proof (due to Propositions 13 and 5). ■

Proof of Theorem 20. The result is a consequence of Theorems 23 and 24. ■

Sub-optimality of $c = \omega_\gamma(1)$. Note that, if $c > 1$, the ODE in (143) maps to the one in (127), with $\underline{a}, \underline{b}$ defined as in (129) (except for absolute constants), with $vc \mapsto v$. This mapping also holds when defining (136), via $\rho \mapsto \rho/c$. Then, if we consider the further condition $\rho/c = \Omega(\gamma^{1-h})$, for some $h > 0$, the lower bound in Theorem 20 becomes

$$\mathcal{R}(\theta_0^p) = \Omega\left(\frac{\gamma\ln(1/\gamma)}{\lambda_{\max}^2} + \frac{\max(1,c^2)\gamma^2\ln^2(1/\gamma)}{\rho^2\lambda_{\max}^2}\right). \tag{176}$$

The same argument holds also for the ODE in (174), providing analogous expression for $\mathcal{R}(\theta_{1/2}^p)$. This quantifies the negative effects of using a large clipping constant $c = \omega_\gamma(1)$.

E.4. Proof of Theorem 21

Let $\overline{R}(t)$ be defined as in (42), with the learning rate schedule in (6) for a generic $\alpha \geq 1$. Then, we have

$$\begin{aligned} \mathrm{d}\overline{R} = &-2vc\lambda_{\min}(1-t)^{\alpha}\frac{\mu_c(\overline{R})}{c}\overline{R}\mathrm{d}t + (vc)^2\lambda_{\max}(1-t)^{2\alpha}\frac{\nu_c(\overline{R})(\overline{R}+\zeta^2/2)}{c^2}\gamma\mathrm{d}t \\ &+ 4(vc)^2\alpha(1-t)^{2\alpha-1}\frac{\gamma^2}{\rho^2}\mathrm{d}t. \end{aligned} \tag{177}$$

The argument is similar to the one used to obtain Theorem 20, so we only highlight differences. Using (118), we have that $\overline{R}(t)$ is strictly bounded by the auxiliary ODEs

$$\begin{aligned} \mathrm{d}\overline{R}' &= -\overline{a}(1-t)^{\alpha}\overline{R}'\mathrm{d}t + \overline{b}_1(1-t)^{2\alpha}\mathrm{d}t + \overline{b}_2(1-t)^{2\alpha-1}\mathrm{d}t =: \overline{f}'(t,\overline{R}')\mathrm{d}t, \qquad \overline{a},\overline{b}_1,\overline{b}_2 > 0, \\ \mathrm{d}\underline{R}' &= -\underline{a}(1-t)^{\alpha}\underline{R}'\mathrm{d}t + \underline{b}_1(1-t)^{2\alpha}\mathrm{d}t + \underline{b}_2(1-t)^{2\alpha-1}\mathrm{d}t =: \underline{f}'(t,\underline{R}')\mathrm{d}t, \qquad \underline{a},\underline{b}_1,\underline{b}_2 > 0, \end{aligned} \tag{178}$$

with initial conditions $\overline{R}'(0) = \overline{R}(0) = R(0)$.

To establish a closed form solution for the ODEs in (178), we start by analyzing the ODE

$$\mathrm{d}\tilde{R} = -\overline{a}(1-t)^{\alpha}\tilde{R}\mathrm{d}t + \overline{b}_1(1-t)^{2\alpha}\mathrm{d}t, \tag{179}$$

with initial condition $\tilde{R}(0) = R(0)$, which admits the closed form solution

$$\begin{aligned} \tilde{R}(t) = \frac{R(0)}{(1+\alpha)}e^{-\frac{\overline{a}(1-(1-t)^{1+\alpha})}{1+\alpha}}\Bigg(1+\alpha - \frac{\overline{b}_1}{R(0)}e^{\frac{\overline{a}}{1+\alpha}}E_{-1+\frac{1}{1+\alpha}}\left(\frac{\overline{a}}{1+\alpha}\right) \\ + \frac{\overline{b}_1}{R(0)}e^{\frac{\overline{a}}{1+\alpha}}(1-t)^{1+2\alpha}E_{-1+\frac{1}{1+\alpha}}\left(\frac{\overline{a}(1-t)^{1+\alpha}}{1+\alpha}\right)\Bigg). \end{aligned} \tag{180}$$

Let

$$w(t) = \overline{R}'(t) - \tilde{R}(t), \tag{181}$$

and note that

$$\begin{aligned} \frac{\mathrm{d}w}{\mathrm{d}t} &= \frac{\mathrm{d}\overline{R}'}{\mathrm{d}t} - \frac{\mathrm{d}\tilde{R}}{\mathrm{d}t} \\ &= -\overline{a}(1-t)^{\alpha}\overline{R} + \overline{a}(1-t)^{\alpha}\tilde{R} + \overline{b}_2(1-t)^{2\alpha-1} \\ &= -\overline{a}(1-t)^{\alpha}w + \overline{b}_2(1-t)^{2\alpha-1}. \end{aligned} \tag{182}$$

Thus, by using the initial condition $w(0) = \overline{R}'(0) - \tilde{R}(0) = 0$, we have

$$w(t) = \frac{\overline{b}_2}{1+\alpha}e^{\frac{\overline{a}(1-t)^{1+\alpha}}{1+\alpha}}\left((1-t)^{2\alpha}E_{\frac{1-\alpha}{1+\alpha}}\left(\frac{\overline{a}(1-t)^{1+\alpha}}{1+\alpha}\right) - E_{\frac{1-\alpha}{1+\alpha}}\left(\frac{\overline{a}}{1+\alpha}\right)\right), \tag{183}$$

which implies that

$$\begin{aligned} \overline{R}'(t) = \frac{R(0)}{(1+\alpha)}e^{-\frac{\overline{a}(1-(1-t)^{1+\alpha})}{1+\alpha}}\Bigg(1+\alpha - \frac{\overline{b}_1}{R(0)}e^{\frac{\overline{a}}{1+\alpha}}E_{-1+\frac{1}{1+\alpha}}\left(\frac{\overline{a}}{1+\alpha}\right) \\ + \frac{\overline{b}_1}{R(0)}e^{\frac{\overline{a}}{1+\alpha}}(1-t)^{1+2\alpha}E_{-1+\frac{1}{1+\alpha}}\left(\frac{\overline{a}(1-t)^{1+\alpha}}{1+\alpha}\right) \\ + \frac{\overline{b}_2}{R(0)}e^{\frac{\overline{a}}{1+\alpha}}\left((1-t)^{2\alpha}E_{\frac{1-\alpha}{1+\alpha}}\left(\frac{\overline{a}(1-t)^{1+\alpha}}{1+\alpha}\right) - E_{\frac{1-\alpha}{1+\alpha}}\left(\frac{\overline{a}}{1+\alpha}\right)\right)\Bigg). \end{aligned} \tag{184}$$

Similarly, we have that

$$\begin{aligned}\underline{R}'(t) = \frac{1}{(1+\alpha)} e^{-\frac{\underline{a}(1-(1-t)^{1+\alpha})}{1+\alpha}} \Bigg(& R(0)(1+\alpha) - \underline{b}_1 e^{\frac{\underline{a}}{1+\alpha}} E_{-1+\frac{1}{1+\alpha}}\left(\frac{\underline{a}}{1+\alpha}\right) \\ & + \underline{b}_1 e^{\frac{\underline{a}}{1+\alpha}} (1-t)^{1+2\alpha} E_{-1+\frac{1}{1+\alpha}}\left(\frac{\underline{a}(1-t)^{1+\alpha}}{1+\alpha}\right) \\ & + \underline{b}_2 e^{\frac{\underline{a}}{1+\alpha}} \left((1-t)^{2\alpha} E_{\frac{1-\alpha}{1+\alpha}}\left(\frac{\underline{a}(1-t)^{1+\alpha}}{1+\alpha}\right) - E_{\frac{1-\alpha}{1+\alpha}}\left(\frac{\underline{a}}{1+\alpha}\right)\right)\Bigg).\end{aligned} \tag{185}$$

For $t \in [0,1]$, $\overline{R}(t), \underline{R}(t) \geq 0$. Furthermore, the following upper bounds hold for $t \in [0,1]$:

$$\begin{aligned}\overline{R}'(t) \leq R(0) + \frac{2\overline{b}_1}{\overline{a}} + \frac{\overline{b}_2}{\alpha} + \overline{b}_1 \left(\frac{1}{1+\alpha}\right)^{-1+\frac{1}{1+\alpha}} \Gamma\left(2 - \frac{1}{1+\alpha}\right) \overline{a}^{-2+\frac{1}{1+\alpha}} \\ + \overline{b}_2 \Gamma\left(\frac{2\alpha}{1+\alpha}\right)(1+\alpha)^{\frac{\alpha-1}{1+\alpha}} \overline{a}^{-\frac{2\alpha}{1+\alpha}}.\end{aligned} \tag{186}$$

Let us take

$$\overline{a} = C_1 v c \lambda_{\min}, \qquad \overline{b}_1 = C_2 \gamma v^2 c^2 \lambda_{\max}, \qquad \overline{b}_2 = C_3 v^2 c^2 \frac{\gamma^2}{\rho^2}\alpha, \tag{187}$$

with C_1, C_2, C_3 constants independent from $\gamma, \rho, \alpha, v, c$ or the spectrum of Σ. Then, we have that $\overline{R}'(t) \leq 1 + R(0)$ for $t \in [0,1]$, as $\overline{a} = \Omega(\ln(1/\gamma))$ and

$$\frac{\ln^2(1/\gamma)\gamma}{\lambda_{\min}^2}\left(\lambda_{\max}\alpha + \frac{\gamma}{\rho^2}\right) = o(1). \tag{188}$$

As a result, we can apply the argument in (118) to obtain that

$$\overline{R}(1) \leq \overline{R}'(1) = \tilde{R}(1) + w(1). \tag{189}$$

Note that Lemma 22 yields

$$\lim_{x\to 0} x^{1+2\alpha} E_{-1+\frac{1}{1+\alpha}}\left(\frac{\overline{a}x^{1+\alpha}}{1+\alpha}\right) = \left(\frac{1}{1+\alpha}\right)^{-2+\frac{1}{1+\alpha}} \Gamma\left(2 - \frac{1}{1+\alpha}\right)\overline{a}^{-2+\frac{1}{1+\alpha}}. \tag{190}$$

Thus,

$$\begin{aligned}\tilde{R}(1) &= \frac{1}{(1+\alpha)} e^{-\frac{\overline{a}}{1+\alpha}} \Bigg(R(0)(1+\alpha) - \overline{b}_1 e^{\frac{\overline{a}}{1+\alpha}} E_{-1+\frac{1}{1+\alpha}}\left(\frac{\overline{a}}{1+\alpha}\right) \\ &\qquad + \overline{b}_1 e^{\frac{\overline{a}}{1+\alpha}} \left(\frac{1}{1+\alpha}\right)^{-2+\frac{1}{1+\alpha}} \Gamma\left(2 - \frac{1}{1+\alpha}\right)\overline{a}^{-2+\frac{1}{1+\alpha}}\Bigg) \\ &\leq R(0) e^{-\frac{\overline{a}}{1+\alpha}} + \left(\frac{1}{1+\alpha}\right)^{-1+\frac{1}{1+\alpha}} \Gamma\left(2 - \frac{1}{1+\alpha}\right)\overline{b}_1 \overline{a}^{-2+\frac{1}{1+\alpha}} \\ &\leq R(0) e^{-\frac{\overline{a}}{1+\alpha}} + (1+\alpha)\overline{b}_1 \overline{a}^{-2+\frac{1}{1+\alpha}},\end{aligned} \tag{191}$$

where in the second line we have used the non-negativity of the exponential integral function and in the third line we have used that $\Gamma\left(2-\frac{1}{1+\alpha}\right) \leq \Gamma(2) = 1$ for all $\alpha \geq 1$. Next, we bound $w(1)$ as

$$\begin{aligned} w(1) &= \frac{\bar{b}_2}{1+\alpha}\left(\Gamma\left(\frac{2\alpha}{1+\alpha}\right)(1+\alpha)^{\frac{2\alpha}{1+\alpha}}\bar{a}^{-\frac{2\alpha}{1+\alpha}} - E_{\frac{1-\alpha}{1+\alpha}}\left(\frac{\bar{a}}{1+\alpha}\right)\right) \\ &\leq \frac{\bar{b}_2}{1+\alpha}\Gamma\left(\frac{2\alpha}{1+\alpha}\right)(1+\alpha)^{\frac{2\alpha}{1+\alpha}}\bar{a}^{-\frac{2\alpha}{1+\alpha}} \\ &\leq (1+\alpha)\bar{b}_2\bar{a}^{-\frac{2\alpha}{1+\alpha}}, \end{aligned} \tag{192}$$

where in the last line we have used that $\Gamma\left(\frac{2\alpha}{1+\alpha}\right) \leq 1$ for $\alpha \geq 1$. By combining (189), (191) and (192), we conclude that

$$\begin{aligned} R(1) &\leq R(0)e^{-\frac{\bar{a}}{1+\alpha}} + (1+\alpha)^{1/3}\bar{b}_1\bar{a}^{-\frac{2\alpha+1}{1+\alpha}} + (1+\alpha)\bar{b}_2\bar{a}^{-\frac{2\alpha}{1+\alpha}} \\ &= O\left(\frac{\lambda_{\max}\alpha\gamma\ln^{\frac{1}{1+\alpha}}(1/\gamma)}{\lambda_{\min}^2} + \frac{\alpha^2\gamma^2\ln^{\frac{2}{1+\alpha}}(1/\gamma)}{\rho^2\lambda_{\min}^2}\right), \end{aligned} \tag{193}$$

which gives the desired result, after applying Propositions 13 and 5, since $\tilde{\eta}(1) = 0$. ■

Proof of (103)**.** By taking $\alpha = \ln\ln(1/\gamma)$, we have

$$\begin{aligned} &\alpha\gamma\ln^{\frac{1}{1+\alpha}}(1/\gamma) + \frac{\alpha^2\gamma^2\ln^{\frac{2}{1+\alpha}}(1/\gamma)}{\rho^2} \\ &\quad= \gamma\ln\ln(1/\gamma)e^{\frac{\ln\ln(1/\gamma)}{1+\ln\ln(1/\gamma)}} + \frac{\gamma^2}{\rho^2}(\ln\ln(1/\gamma))^2 e^{\frac{2\ln\ln(1/\gamma)}{1+\ln\ln(1/\gamma)}} \\ &\quad\leq \gamma(\ln\ln(1/\gamma))e + \frac{\gamma^2}{\rho^2}(\ln\ln(1/\gamma))^2 e^2. \end{aligned} \tag{194}$$

Thus, in the setting where $\lambda_{\max}, \lambda_{\min} = \Theta_\gamma(1)$, this choice (together with Propositions 13 and 5) yields

$$\mathcal{R}(\theta^p) = O\left(\gamma(\ln\ln(1/\gamma)) + \frac{\gamma^2}{\rho^2}(\ln\ln(1/\gamma))^2\right). \tag{195}$$

Appendix F. Future work

Our work opens the door to a number of interesting future directions. The first consists in a tighter characterization of the test risk with respect to the condition number $\kappa = \lambda_{\max}/\lambda_{\min}$ of the data covariance. This is done e.g. in [39, 57] which are however unable to handle the proportional regime considered in our work. The second direction regards the study of the optimal scheduling of learning rate and private noise beyond the regime $\gamma = d/n \to 0$. In fact, Figure 4 suggests that different values of γ lead to different optimal schedules. Finally, homogenized DP-SGD has the potential to offer a powerful tool to study differentially private optimization beyond linear regression. A concrete setting for future work is provided e.g. by logistic regression, where the SGD dynamics (in the absence of clipping and private noise) has been considered in [17].