ADAPTIVE METHODS ARE PREFERABLE IN HIGH PRIVACY SETTINGS: AN SDE PERSPECTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

Differential Privacy (DP) is becoming central to large-scale training as privacy regulations tighten. We revisit how DP noise interacts with *adaptivity* in optimization through the lens of *stochastic differential equations*, providing the first SDE-based analysis of private optimizers. Focusing on `DP-SGD` and `DP-SignSGD` under per-example clipping, we show a sharp contrast under fixed hyperparameters: `DP-SGD` converges at a privacy-utility trade-off $\mathcal{O}(1/\varepsilon^2)$ with speed independent of ε, while `DP-SignSGD` converges at a speed *linear in* ε with a $\mathcal{O}(1/\varepsilon)$ trade-off, dominating in high-privacy or high-noise regimes. Under optimal learning rates, both methods reach comparable theoretical asymptotic performance; however, the optimal learning rate of `DP-SGD` scales linearly with ε, while that of `DP-SignSGD` is essentially ε-independent. This makes adaptive methods far more practical, as their hyperparameters transfer across privacy levels with little or no re-tuning. Empirical results confirm our theory across training and test metrics, and extend from `DP-SignSGD` to `DP-Adam`.

1 INTRODUCTION

The rapid deployment of large-scale machine learning systems has intensified the demand for rigorous privacy guarantees. In sensitive domains such as healthcare or conversational agents, even the disclosure of a single training example can have serious consequences. Legislation and policy initiatives show that AI regulation is tightening rapidly. In the United States, the *Executive Order of October 30, 2023* mandates developers of advanced AI systems to share safety test results and promotes privacy-preserving techniques such as differential privacy (House, 2023). Complementing this, the National Institute of Standards and Technology (NIST), a U.S. federal agency, released draft guidance (SP 800-226) on privacy guarantees in AI (NITS, 2023a) and included "privacy-enhanced" as a key dimension in its AI Risk Management Framework (RMF 1.0) (NITS, 2023b). In Europe, the *EU AI Act* sets binding obligations for high-risk systems (EU, 2023), while ENISA recommends integrating data protection into AI development (EU, 2024). In this context, Differential Privacy (DP) (Dwork et al., 2006) is therefore emerging as the de facto standard for ensuring user-level confidentiality in stochastic optimization. By injecting carefully calibrated noise into the training process, DP optimizers protect individual data points while inevitably trading off some population-level utility.

A central open question is how differential privacy noise influences optimization dynamics, and in particular, how it interacts with adaptivity and batch noise. In this work, we revisit this problem through the lens of *stochastic differential equations* (SDEs), which, over the last decade, have proven to be a powerful tool for analyzing optimization algorithms (Li et al., 2017; Mandt et al., 2017; Compagnoni et al., 2023). While SDEs have not yet been applied to DP methods, here we use them to uncover a key and previously overlooked phenomenon: *DP noise affects adaptive and non-adaptive methods in structurally different ways*. We focus on two fundamental DP optimizers: `DP-SGD` (Abadi et al., 2016) and `DP-SignSGD`. The former serves as the baseline for DP optimization; Although the latter is not widely used in practice, it is substantially simpler to analyze than the popular DP optimizer `DP-Adam` (Gylberth et al., 2017; Zhou et al., 2020b; Li et al., 2021a; McKenna et al., 2025). Relying on `SignSGD` as a proxy for `Adam` is standard in prior work (Compagnoni et al., 2025c; Balles & Hennig, 2018; Zou et al., 2021; Peng et al., 2025; Li et al., 2025), and this motivates our focus on `DP-SignSGD` for the theoretical development. Importantly, setting $\beta_1 = \beta_2 = 0$ reduces `DP-Adam` to `DP-SignSGD`. We leave the study of more advanced DP optimizers to future work, as each would require a separate technical treatment. Under standard

assumptions and with per-example clipping, our analysis isolates how the privacy budget ε, which governs the overall privacy level, influences the dynamics.

In practice, private training is usually performed across a range of privacy budgets ε, and for each value one searches for the best-performing hyperparameters. A change in ε can therefore arise either from this exploratory sweep or from stricter regulatory requirements. To capture these situations, we study two complementary protocols. **Protocol A (fixed hyperparameters):** To examine the situation when re-tuning is not feasible, e.g., low budget, we we first fix a privacy budget ε and find the optimal configuration $(\eta, C, B, \dots)$ via grid search. Then, we analyze how performance changes if training were repeated under different ε, without adjusting hyperparameters, therefore isolating the impact of ε on the performance. **Protocol B (best-tuned per ε):** When re-tuning is allowed, we search the optimal hyperparameters (i.e., $(\eta, C, B, \dots)$) for each ε, thereby isolating the *intrinsic scaling of the optimal learning rates with respect to* ε.

Contributions. Our work makes the following contributions:

1. We provide the first SDE-based analysis of differentially private optimizers, using this framework to expose how DP noise interacts with adaptivity and batch noise;
2. **Protocol A:** We show that `DP-SGD` converges at a speed *independent* of ε, with a privacy-utility trade-off that scales as $\mathcal{O}\left(1/\varepsilon^2\right)$ (consistent with prior work);
3. **Protocol A:** We prove a novel result for `DP-SignSGD`: its convergence speed scales linearly in ε, while its privacy-utility trade-off scales as $\mathcal{O}\left(1/\varepsilon\right)$;
4. **Protocol A:** When batch noise is sufficiently large, `DP-SignSGD` always dominates. When batch noise is small, the outcome depends on the privacy budget: for strict privacy ($\varepsilon < \varepsilon^\star$), `DP-SignSGD` is preferable, while for looser privacy ($\varepsilon > \varepsilon^\star$), `DP-SGD` has better performance;
5. **Protocol B:** We theoretically derive that the optimal learning rate of `DP-SGD` scales as $\eta^\star \propto \varepsilon$, while the optimal learning rate of `DP-SignSGD` is ε-independent. This tuning allows the two methods to reach theoretically *comparable* asymptotic performance, including at very small ε;
6. We empirically validate all our theoretical insights on real-world tasks, and show that the qualitative insights extend from training to *test* loss and from `DP-SignSGD` to `DP-Adam`.

In summary, our results refine the privacy–utility landscape, which, to our knowledge, has not yet provided a definitive answer as to which of `DP-SGD` or `DP-Adam`/`DP-SignSGD` performs best, and under which conditions. Under Protocol A, adaptivity is preferable in stricter privacy regimes: `DP-SignSGD` converges more slowly but achieves better utility when ε is small or batch noise is large, whereas `DP-SGD` converges faster but suffers sharper degradation. Under Protocol B, both methods achieve comparable asymptotic performance; however, adaptive methods are far more practical, as their optimal learning rate is essentially ε-independent, allowing it to transfer across privacy levels with little or no re-tuning. This matters not only for computational cost but also for privacy, since each hyperparameter search consumes additional budget (Papernot & Steinke, 2021). In contrast, `DP-SGD` requires an ε-dependent learning rate tuned *ad hoc*, making it brittle if the sweep grid misses the "right" value. Intuitively, adaptive methods inherently adjust to the scale of DP noise, whereas non-adaptive methods require explicit tuning of the learning rate to counter the effect of privacy noise.

2 Related Work

SDE approximations. SDEs have long been used to analyze discrete-time optimization algorithms (Helmke & Moore, 1994; Kushner & Yin, 2003). Beyond their foundational role, these approximations have been applied to practical tasks such as learning-rate tuning (Li et al., 2017; 2019) and batch-size selection (Zhao et al., 2022). Other works have focused on deriving convergence bounds (Compagnoni et al., 2023; 2024; 2025c), uncovering scaling laws that govern optimization dynamics (Jastrzebski et al., 2018; Compagnoni et al., 2025c;a), and revealing implicit effects such as regularization (Smith et al., 2021; Compagnoni et al., 2023) and preconditioning (Xiao et al., 2025; Marshall et al., 2025). In particular, SDE-based techniques have been used to study a broad class of modern adaptive optimizers, including RMSProp, Adam, AdamW, and SignSGD, as well as minimax and distributed variants (Compagnoni et al., 2024; 2025c;a; Xiao et al., 2025). Most analyses rely on weak approximations, as rigorously formalized by Li et al. (2017), although some works have also considered heavy-tailed batch noise via Lévy-driven SDEs to capture non-Gaussianity (Simsekli et al., 2019; Zhou et al., 2020a). Despite this progress, prior work has exclusively focused on non-private optimization. To our knowledge, ours is the first to extend the

SDE lens to differentially private optimizers, including explicit convergence rates and stationary distributions as functions of the privacy budget.

Differential privacy in optimization. Differentially private training is most commonly implemented via DP−SGD (Abadi et al., 2016), which clips per-example gradients to a fixed norm bound to control sensitivity and injects calibrated Gaussian noise into the averaged update. Advanced accounting methods such as the moments accountant (Abadi et al., 2016) and Rényi differential privacy (Mironov, 2017; Wang et al., 2019), combined with privacy amplification by subsampling (Balle et al., 2018; 2020), allow practitioners to track the cumulative privacy cost tightly over many updates and have made large-scale private training feasible. A central challenge is that clipping, while essential for privacy, also alters the optimization dynamics: overly aggressive thresholds bias gradients and can stall convergence (Chen et al., 2020), prompting extensive work on how to set or adapt the clipping norm. Approaches include rule-based or data-driven thresholds, such as AdaCliP (Pichapati et al., 2019) and quantile-based adaptive clipping (Andrew et al., 2021), as well as recent analyses that characterize precisely how the clipping constant influences convergence (Koloskova et al., 2023). Together, these contributions have positioned DP−SGD and its variants as the standard backbone for differentially private optimization.

Adaptive DP optimizers. Adaptive methods such as AdaGrad (Duchi et al., 2011; McMahan & Streeter, 2010), RMSProp (Tieleman & Hinton, 2012), and Adam (Kingma & Ba, 2015) generally outperform non-adaptive SGD in non-private training. However, this performance *gap* under DP constraints; i) narrows considerably (Zhou et al., 2020b; Li et al., 2022); ii) essentially vanishes when both optimizers are carefully tuned, as observed for large-scale LLM fine-tuning in Li et al. (2021a, App. S). Consistently with non-DP training, non-adaptive methods are sometimes still preferred in vision tasks (De et al., 2022). Therefore, which of DP−SGD and DP−Adam is preferable remains an open question. Under assumptions that include bounded/convex domain, bounded gradient norm, bounded gradient noise, convexity of the loss, and possibly without performing clipping of the per-sample gradients, several strategies have been theoretically and empirically explored to mitigate the drop in performance of adaptive methods in DP. These include bias-corrected DP−Adam variants (Tang & Lécuyer, 2023; Tang et al., 2023), the use of non-sensitive auxiliary data (Asi et al., 2021), and scale-then-privatize techniques that exploit adaptivity before noise injection (Li et al., 2023; Ganesh et al., 2025). A most recent related work by (Jin & Dai, 2025) studies Noisy SignSGD: Conceptually, they investigate how the sign compressor amplifies privacy, and argue that the sign operator itself provides privacy amplification beyond the Gaussian mechanism. Their analysis establishes convergence guarantees in the distributed learning setting while relying on *bounded gradient norms and bounded variance* assumptions, thereby avoiding the need for clipping and explicitly leaving its study to future work.

We view these contributions as providing valuable theoretical and empirical advances in the design of adaptive private optimizers, clarifying many important aspects of their behavior as well as trying to restore the aforementioned performance *gap*. Yet, the fundamental question of *which privacy regimes are most favorable to adaptivity* remains largely unanswered, and addressing it could explain at least one aspect of the nature of this *gap*. Our work addresses this *open question* by analyzing *why and when adaptivity matters* under DP noise, identifying the regimes where adaptive methods dominate and where they match non-adaptive ones. Crucially, we incorporate *per-example clipping*, a central element of DP−SGD, and a heavy-tailed batch noise model that captures unbounded variance.

3 Preliminaries

General Setup and Noise Assumptions. We model the loss function with a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ with global minimum $f^* = 0$: This is not restrictive, as one can always consider the suboptimality $f(x) - f^*$ and rename it as f. Regarding noise assumptions, recent literature commonly assumes that the stochastic gradient of the loss function on a minibatch γ can be decomposed as $\nabla f_\gamma(x) = \nabla f(x) + Z_\gamma$ where batch noise Z_γ is modeled with a Gaussian (Ahn et al., 2012; Chen et al., 2014; Mandt et al., 2016; Stephan et al., 2017; Zhu et al., 2019; Jastrzebski et al., 2018; Wu et al., 2020; Xie et al., 2021), often with constant covariance matrix (Li et al., 2017; Mertikopoulos & Staudigl, 2018; Raginsky & Bouvrie, 2012; Zhu et al., 2019; Mandt et al., 2016; Ahn et al., 2012; Jastrzebski et al., 2018). In this work, we refine the standard noise assumption to distinguish the two regimes induced by per-example clipping in DP training. Since clipping is applied at the datapoint level, each mini-batch contains a mix of *clipped* and *unclipped* gradients. For unclipped datapoints, we follow the usual literature and model the batch-averaged noise as Gaussian. For

clipped datapoints, which do not benefit from batch averaging, we model the per-example noise as multivariate Student-t, $Z_\gamma \sim \sigma_\gamma t_\nu(0, I_d)$, capturing potentially heavy-tailed behaviour and recovering the Gaussian case as $\nu \to \infty$. See Assumption B.2 and Remark B.2 for more details. Finally, we use the following approximation, formally derived in Lemma A.2: $\mathbb{E}\left[\frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|}\right] \approx \frac{\nabla f(x)}{\sigma_\gamma \sqrt{d}}$. The approximation is valid under two assumptions: $i)$ The parameter dimension d is sufficiently large ($d = \Omega(10^4)$), consistent with modern deep learning models that often reach billions of trainable parameters; $ii)$ The signal-to-noise ratio satisfies $\frac{\|\nabla f(x)\|_2^2}{2\sigma_\gamma^2} \ll d$: This condition has been thoroughly empirically studied by Malladi et al. (2022) (Appendix G), who observed that across multiple tasks and architectures the ratio $\frac{\|\nabla f(x)\|_2^2}{2\sigma_\gamma^2}$ never exceeds $\mathcal{O}(10^2)$, well below typical values of d. Therefore, this signal-to-noise ratio need not be small: We simply require it to be smaller than d — See Remark A.1 for more details, including experimental validations. We highlight that our experiments confirm that the insights derived from our theoretical results carry over to real-world tasks. Importantly, while our theory is developed for `DP-SignSGD`, we further validate that the same insights hold empirically for `DP-Adam`, showing that our insights extend directly to this widely used private optimizer, as well as also transfer from training to test loss. This highlights both the mildness of the assumptions and the robustness of the analysis.

SDE approximation. The following definition formalizes in which sense a continuous-time model, such as a solution to an SDE, can accurately describe the dynamics of a discrete-time process, such as an optimizer. Drawn from the field of numerical analysis of SDEs (see Mil'shtein (1986)), it quantifies the disparity between the discrete and the continuous processes. Simply put, the approximation is meant in a *weak sense*, meaning in distribution rather than path-wise: We require their expectations to be close over a class of test functions with polynomial growth, meaning that all the moments of the processes become closer at a rate of η^α and thus their distributions.

Definition 3.1 *Let $0 < \eta < 1$ be the learning rate, $\tau > 0$ and $T = \lfloor \frac{\tau}{\eta} \rfloor$. We say that a continuous time process X_t over $[0, \tau]$, is an order-α weak approximation of a discrete process x_k, if for any polynomial growth function g, $\exists M > 0$, independent of the learning rate η, such that for all $k = 0, 1, \ldots, T$, $|\mathbb{E}g(X_{k\eta}) - \mathbb{E}g(x_k)| \leq M\eta^\alpha$.*

Remark 3.1 (Validity of the SDE approximation) *To guarantee that the SDE model is a first-order weak approximation of the optimizer dynamics in the sense of Definition 3.1, one shows that the first two moments of the one-step increments of the optimizer and of the SDE match up to $O(\eta^2)$, while all higher-order terms in the Taylor expansion are collected in an $O(\eta^2)$ remainder (see Appendix B). This implies that the discrepancy between the two processes scales as $O(\eta)$ for any test function of polynomial growth and any finite time horizon. The neglected $O(\eta^2)$ terms could, in principle, be retained by deriving higher-order SDEs. However, to the best of our knowledge, such second-order models have been derived (Li et al., 2017), but have not yet led to additional practical insight in the analysis of optimisation algorithms. Finally, Figure C.1 empirically compares the discrete algorithms with their SDE counterparts on quadratic and quartic objectives, confirming that for the step sizes used in our experiments, the first-order SDEs closely track the discrete dynamics.*

While we refer the reader to Appendix B for technical details, we illustrate with a basic example. The SGD iterates follow $x_{k+1} = x_k - \eta \nabla f_{\gamma_k}(x_k)$, and, as shown in Li et al. (2017), it can be approximated in continuous time by the first-order SDE

$$dX_t = -\nabla f(X_t)dt + \sqrt{\eta}\sqrt{\Sigma(X_t)}dW_t, \tag{1}$$

where $\Sigma(x) = \frac{1}{n}\sum_{i=1}^n (\nabla f(x) - \nabla f_i(x))(\nabla f(x) - \nabla f_i(x))^\top$ is the gradient noise covariance. Intuitively, the iterates drift along the gradient while the stochasticity scales with this covariance.

Differential Privacy. Here, we outline the relevant background of foundational prior work in DP optimization. We adopt the standard (ε, δ)-DP framework (Dwork et al., 2006).

Definition 3.2 *A random mechanism $\mathcal{M} : \mathcal{D} \to \mathcal{R}$ is said to be (ε, δ)-differentially private if for any two adjacent datasets $d, d' \in \mathcal{D}$ (i.e., they differ in 1 sample) and for any subset of outputs $S \subseteq \mathcal{R}$ it holds that $\mathbb{P}[\mathcal{M}(d) \in S] \leq e^\varepsilon \mathbb{P}[\mathcal{M}(d') \in S] + \delta$.*

In this work, we consider example-level differential privacy applied by a central trusted aggregator. We implement this using the sub-sampled Gaussian mechanism (Dwork & Roth, 2014; Mironov

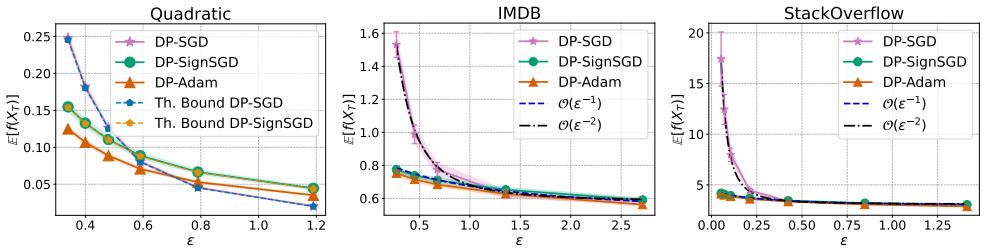

Figure 1: Empirical validation of the privacy-utility trade-off predicted by Thm. 4.1 and Thm. 4.3, comparing DP−SGD, DP−SignSGD, and DP−Adam: Our focus is on verifying the functional dependence of the asymptotic loss levels in terms of ε. **Left:** On a quadratic convex function $f(x) = \frac{1}{2}x^\top Hx$, the observed empirical loss values perfectly match the theoretical predictions (Eq. 7, Eq. 10). **Center and Right:** Logistic regressions on the IMDB dataset (center) and the StackOverflow dataset (right), confirm the same pattern: the utility of DP−SGD scales as $\frac{1}{\varepsilon^2}$, while the utility of DP−SignSGD scales linearly as $\frac{1}{\varepsilon}$. Across all settings, we observe that the insights obtained for DP−SignSGD extend to DP−Adam as well as to the test loss (see Figure C.4). For experimental details see Appendix C.2.

et al., 2019) to perturb the SGD updates: At each iteration, a random mini-batch is drawn, per-example gradients are clipped to a fixed bound to limit sensitivity, and Gaussian noise is added to the averaged clipped gradients. The following definition formalizes these mechanisms and provides the update rules for DP−SGD and DP−SignSGD.

Definition 3.3 *For $k \geq 0$, learning rate η, variance σ_{DP}^2, and batches γ_k of size B modeled as i.i.d. uniform random variables taking values in $\{1, \dots, n\}$. Let g_k be the private gradient, defined as*

$$g_k := \frac{1}{B}\sum_{i\in\gamma_k} \mathcal{C}(\nabla f_i(x_k)) + \frac{1}{B}\mathcal{N}(0, C^2\sigma_{DP}^2 I_d) \tag{2}$$

and $\mathcal{C}[\cdot]$ be the clipping function

$$\mathcal{C}(x) = \begin{cases} C\frac{x}{\|x\|_2} & \text{if } \|x\|_2 \geq C \\ x & \text{if } \|x\|_2 < C \end{cases}. \tag{3}$$

The iterates of DP−SGD are defined as

$$x_{k+1} = x_k - \eta g_k, \tag{4}$$

while those of DP−SignSGD are defined as

$$x_{k+1} = x_k - \eta \operatorname{sign}[g_k], \tag{5}$$

where sign$[\cdot]$ *is applied component-wise. Finally, those of DP-Adam are defined in Eq. 208.*

We say that an optimizer is in Phase 1 *if the argument of $\mathcal{C}$ is larger than C and Phase* 2 *otherwise.*

The following theorem from (Abadi et al., 2016) gives the conditions under which DP−SGD, and thus also DP−SignSGD, is a differentially-private algorithm.

Theorem 3.1 *For $q = \frac{B}{n}$ where B is the batch size, n is the number of training points, and number of iterations T, $\exists c_1, c_2$ s.t. $\forall \varepsilon < c_1 q^2 T$, if the noise multiplier σ_{DP} satisfies $\sigma_{DP} \geq c_2 \frac{q\sqrt{T\log(1/\delta)}}{\varepsilon}$, DP−SGD is (ε, δ)-differentially private for any $\delta > 0$. In the following, we will often use $\tilde{\sigma}_{DP} = \frac{\sqrt{T}\Phi}{\varepsilon}$, where $\Phi := q\sqrt{\log(1/\delta)}$ to indicate the DP noise multiplier.*

4 Theoretical Results

In this section, we investigate how the privacy budget ε influences convergence speed and shapes the privacy-utility trade-offs in both the loss and the gradient norm. To do so, we leverage SDE models for DP−SGD and DP−SignSGD, which can be found in Theorem B.5 and Theorem B.10, respectively, and are experimentally validated in Figure C.1. In addition, we provide the first stationary distributions for these optimizers, presented in Theorem B.9 and Theorem B.15 in the Appendix. This section is organized as follows:

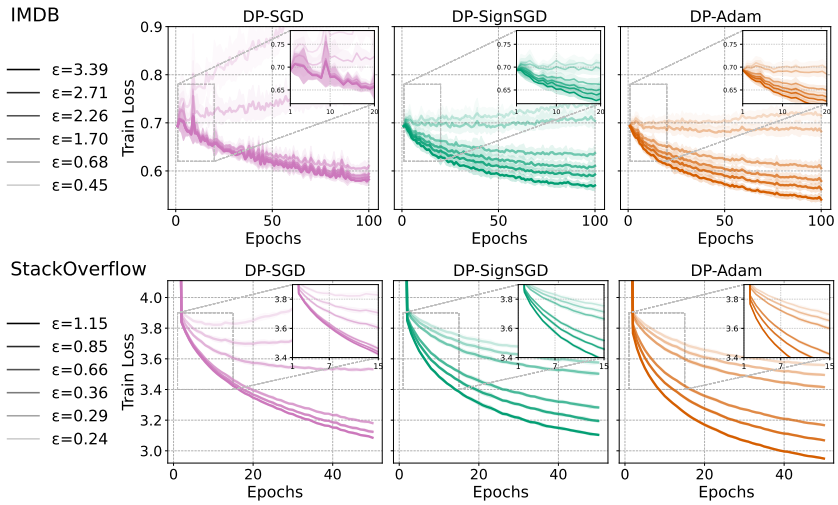

Figure 2: Empirical validation of the convergence speeds predicted by Thm. 4.1 and Thm. 4.3. We compare `DP-SGD`, `DP-SignSGD`, and `DP-Adam` as we train a logistic regression on the IMDB dataset (**Top Row**) and on the StackOverflow dataset (**Bottom Row**). In both tasks, we verify that when `DP-SGD` converges, its speed is unaffected by ε. As expected, it diverges when ε is too small. Regarding `DP-SignSGD` and `DP-Adam`, they are faster when ε is large and never diverge even when this is small. Crucially, Figure C.5 shows that these insights are also verified on the test loss. For experimental details see Appendix C.3.

1. **Protocol A (Section 4.1).** Section 4.1.1 analyzes `DP-SGD`, yielding bounds for the loss (Thm. 4.1) and the gradient norm (Thm. 4.2) in the μ-PL and L-smooth cases, respectively: We observe that the convergence speed is *independent* of ε, while the privacy-utility trade-off scales as $\mathcal{O}(1/\varepsilon^2)$. Section 4.1.2 analyzes `DP-SignSGD`, and Thm. 4.3 and Thm. 4.4) show a qualitatively different behavior: Convergence speed scales linearly with ε, while the privacy-utility terms scale as $\mathcal{O}(1/\varepsilon)$, making adaptivity preferable if the privacy budget is small enough. Finally, Theorem 4.5 in Section 4.1.3 shows that when batch noise is large enough, `DP-SignSGD` always dominates. When batch noise is small, the outcome depends on the privacy budget: There exists $\varepsilon^\star$ such that for strict privacy ($\varepsilon < \varepsilon^\star$), `DP-SignSGD` is preferable, while for looser privacy ($\varepsilon > \varepsilon^\star$), `DP-SGD` is better.
2. **Protocol B (Section 4.2).** In this section, we derive the optimal learning rates of `DP-SGD` and `DP-SignSGD`: That of `DP-SGD` scales linearly in ε, while that of `DP-SignSGD` is independent of it. Under these parameter choices, they achieve the same asymptotic neighbourhoods.

We empirically validate our theoretical insights on real datasets[1]. Crucially, the same insights derived from `DP-SignSGD` *empirically* extend to `DP-Adam` as well as to test metrics: This underscores the mildness of our assumptions and the depth of our analysis.

Notation. In the following, we use the symbol $\lesssim$ to suppress absolute numerical constants (e.g., 2, 4, etc.), and never problem-dependent quantities such as d, μ, L, or ε: This convention lightens the presentation. Finally, observe that $\Phi := q\sqrt{\log(1/\delta)} = \frac{B}{n}\sqrt{\log(1/\delta)} \Rightarrow \frac{\Phi}{B} = \frac{1}{n}\sqrt{\log(1/\delta)}$. We will often use ε to highlight the privacy budget in relevant formulas.

4.1 Protocol A: Fixed Hyperparameters

Following the tuning routine of Li et al. (2023), we conduce extensive grid search to select a configuration $(\eta, C, B, \ldots)$ for one ε and keep them unchanged as we vary ε. In particular, η does *not* depend on ε or on other hyperparameters. This absolute comparison exposes structural differences in how DP noise interacts with adaptive vs non-adaptive updates.

[1]For all our experiments, we use the official GitHub repository https://github.com/kenziyuliu/DP2 released with the Google paper Li et al. (2023).

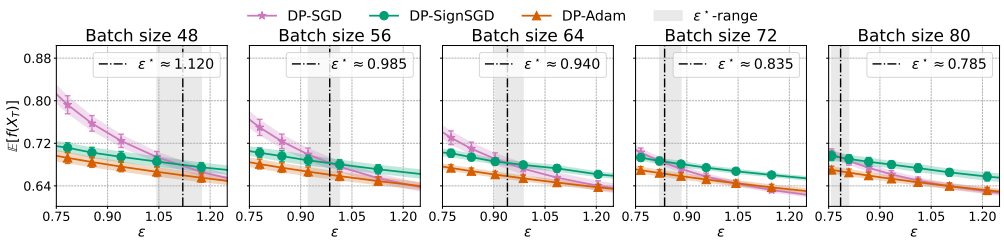

Figure 3: Logistic regression on IMDB Dataset: From left to right, we decrease the batch noise, i.e., increase the batch size, taking values $B \in \{48, 56, 64, 72, 80\}$: As per Theorem 4.5, the privacy threshold $\varepsilon^\star$ that determines when DP−SignSGD is more advantageous than DP−SGD shifts to the left. This confirms that if there is more noise due to the batch size, less privacy noise is needed for DP−SignSGD to be preferable over DP−SGD. For experimental details see Appendix C.4.

4.1.1 DP−SGD: THE PRIVACY-UTILITY TRADE-OFF IS $\mathcal{O}(1/\varepsilon^2)$

By definition, DP−SGD might alternate between a *clipped* and an *unclipped* phase. We first take a didactic perspective to analyze each phase separately to isolate the role of ε on the dynamics, while Theorem B.8 covers the case where these phases are mixed.

Theorem 4.1 *Let f be μ-PL and L-smooth, then we have that during*

• *Phase 1, i.e., when the gradient is clipped, the loss satisfies:*

$$\mathbb{E}[f(X_t)] \lesssim f(X_0) \underbrace{e^{-\frac{\mu C}{\sigma_\gamma \sqrt{d}} t}}_{\text{Decay}} + \left(1 - e^{-\frac{\mu C}{\sigma_\gamma \sqrt{d}} t}\right) \underbrace{\frac{T\eta d^{\frac{3}{2}} LC\sigma_\gamma}{\mu} \left(\frac{\varepsilon^2}{dT} + \frac{\Phi^2}{B^2}\right) \frac{1}{\varepsilon^2}}_{\text{Privacy-Utility Trade-off}}; \tag{6}$$

• *Phase 2, i.e., when the gradient is not clipped, the loss satisfies:*

$$\mathbb{E}[f(X_t)] \lesssim f(X_0) \underbrace{e^{-\mu t}}_{\text{Decay}} + (1 - e^{-\mu t}) \underbrace{\frac{T\eta dL}{\mu} \left(\frac{\varepsilon^2 {\sigma_\gamma}^2}{BT} + C^2 \frac{\Phi^2}{B^2}\right) \frac{1}{\varepsilon^2}}_{\text{Privacy-Utility Trade-off}}. \tag{7}$$

The decay rates are independent of ε: in Phase 2 they depend only on μ, while in Phase 1 normalization spreads the signal over the sphere of radius C (Vershynin, 2018, Ch. 3), giving a rate proportional to $C/(\sigma_\gamma \sqrt{d})$. In both phases, the privacy-utility term scales as $1/\varepsilon^2$.

We now turn to analyzing SDE dynamics assuming only L-smoothness of f. The following theorem presents a bound on the expected gradient norm *across* both phases **together**: We observe that the expected gradient norm admits the same $\mathcal{O}(1/\varepsilon^2)$ scaling.

Theorem 4.2 *Let f be L-smooth, $K_1 := \max\{1, \frac{\sigma_\gamma \sqrt{d}}{C}\}$, and $K_2 := \max\{\frac{\sigma_\gamma^2}{B}, \frac{C^2}{d}\}$. Then,*

$$\mathbb{E}\left[\|\nabla f(X_{\tilde{t}})\|_2^2\right] \lesssim K_1 \left(\frac{f(X_0)}{\eta T} + \eta dL \left(K_2 + \frac{C^2 \left(\frac{q}{B}\right)^2 T \log(1/\delta)}{\varepsilon^2}\right)\right), \tag{8}$$

where $\tilde{t}$ is a random time with uniform distribution over $[0, \tau]$.

Takeaway. Theorem 4.1 separates two effects: the *decay* terms, which determine the convergence speed, and the *privacy-utility* terms, which determine the asymptotic neighbourhood under DP. Our results show that the convergence speed of DP−SGD is unaffected by the privacy budget ε: Figure 2 confirms empirically that, whenever DP−SGD does not diverge, its convergence speed is independent of ε. Additionally, the privacy-utility trade-off scales as $\mathcal{O}(1/\varepsilon^2)$: This insight is validated in Figure 1: on a quadratic function (left panel) the observed loss matches the theoretical values from Theorem 4.1, and the same scaling is reproduced when training logistic regression on IMDB and StackOverflow (center and right panels). The behavior also persists on the test loss (Figure C.4).

4.1.2 DP−SIGNSGD: THE PRIVACY-UTILITY TRADE-OFF IS $\mathcal{O}(1/\varepsilon)$

As for DP−SGD, we isolate the effect of ε on the dynamics of DP−SignSGD and study the loss in each phase **separately**, while Theorem B.14 covers the case where these phases are mixed.

Theorem 4.3 *Let f be μ-PL and L-smooth. Then, we have that during*

• *Phase 1, i.e., when the gradient is clipped, the loss satisfies:*

$$\mathbb{E}[f(X_t)] \lesssim f(X_0)\underbrace{e^{\frac{-\mu B}{\sigma_\gamma\sqrt{dT}}\frac{\varepsilon}{\Phi}t}}_{\text{Decay}} + \left(1 - e^{\frac{-\mu B}{\sigma_\gamma\sqrt{dT}}\frac{\varepsilon}{\Phi}t}\right)\underbrace{\frac{\sqrt{T}\eta L d^{\frac{3}{2}}\sigma_\gamma}{\mu B}\frac{\Phi}{\varepsilon}}_{\text{Privacy-Utility Trade-off}}; \tag{9}$$

• *Phase 2, i.e., when the gradient is not clipped, the loss satisfies:*

$$\mathbb{E}[f(X_t)] \lesssim f(X_0)\underbrace{e^{\sqrt{\varepsilon^2\frac{\sigma_\gamma^2}{B}+\frac{C^2\Phi^2}{B^2}}\frac{-\mu\varepsilon t}{T}}}_{\text{Decay}} + \left(1 - e^{\sqrt{\varepsilon^2\frac{\sigma_\gamma^2}{B}+\frac{C^2\Phi^2}{B^2}}\frac{-\mu\varepsilon t}{T}}\right)\underbrace{\frac{\sqrt{T}\eta L d}{\mu}\sqrt{\frac{\varepsilon^2\sigma_\gamma^2}{BT}+\frac{C^2\Phi^2}{B^2}}\frac{1}{\varepsilon}}_{\text{Privacy-Utility Trade-off}}. \tag{10}$$

The decay rate scales proportionally with ε in both phases (Eq. 9 and Eq. 10), unlike DP−SGD, where it is independent of ε (Eq. 6 and Eq. 7). At the same time, the privacy-utility term in both phases scales as $\mathcal{O}(1/\varepsilon)$, which *might* be more favorable than the $\mathcal{O}(1/\varepsilon^2)$ scaling of DP−SGD in high-privacy regimes, e.g., if ε is sufficiently small.

Assuming only L-smoothness of f, the following theorem presents a bound on the expected gradient norm *across* both phases **together**. As the bound scales as $\mathcal{O}(1/\varepsilon)$, it suggests that adaptivity *might* mitigate the effect of large privacy noise on performance. Intuitively, the sign[·] effectively clips the privatized gradient signal, capping the update magnitude and reducing sensitivity to noise corruption.

Theorem 4.4 *Let f be L-smooth and $K_3 := \max\left\{\sqrt{\frac{\sigma_\gamma^2\varepsilon^2}{BT}+\frac{C^2\Phi^2}{B^2}}, \frac{\sigma_\gamma\Phi}{B}\sqrt{d}\right\}$. Then,*

$$\mathbb{E}\left[\|\nabla f(X_{\tilde{t}})\|_2^2\right] \lesssim K_3\left(\frac{f(X_0)}{\eta\sqrt{T}} + \eta d L\sqrt{T}\right)\frac{1}{\varepsilon}, \tag{11}$$

where $\tilde{t}$ is a random time with uniform distribution over $[0, \tau]$.

Takeaway: Theorem 4.3 suggests that the privacy noise directly enters the convergence dynamics of DP−SignSGD, making its behavior qualitatively different from DP−SGD: The center column of Figure 2 confirms that DP−SignSGD converges faster for larger ε. Additionally, it also shows that it never diverges as drastically as DP−SGD for small ε. This is better shown in Figure 1, where we validate that the asymptotic loss scales with $\frac{1}{\varepsilon}$, while that of DP−SGD scales with $\frac{1}{\varepsilon^2}$. Therefore, adaptive methods are preferable in high-privacy settings, and all these insights are verified also for DP−Adam and generalize to the test loss (Figure C.4).

4.1.3 When adaptivity really matters under fixed hyperparameters.

In this subsection, we quantify when an adaptive method such as DP−SignSGD achieves better utility than DP−SGD. To this end, we compare *Privacy-Utility* terms of Phase 2 for both methods and derive conditions on the two sources of noise that govern the dynamics: the batch noise size σ_γ and the privacy budget ε.

Theorem 4.5 *If $\sigma_\gamma^2 \geq B$, then DP−SignSGD always achieves a better privacy-utility trade-off than DP−SGD. If $\sigma_\gamma^2 < B$, there exists a critical privacy level $\varepsilon^\star = \sqrt{\frac{C^2TB}{n^2(B-\sigma_\gamma^2)}\log\left(\frac{1}{\delta}\right)}$ such that DP−SignSGD outperforms DP−SGD in utility whenever $\varepsilon < \varepsilon^\star$.*

Takeaway: This result makes the comparison explicit: *i)* Under **large batch noise** ($\sigma_\gamma^2 \geq B$), DP−SignSGD achieves a better utility than DP−SGD; *ii)* Under **small batch noise** ($\sigma_\gamma^2 < B$), the best optimizer depends on the privacy budget. For strict privacy ($\varepsilon < \varepsilon^\star$), DP−SignSGD has better utility, while for looser privacy ($\varepsilon > \varepsilon^\star$), DP−SGD achieves better overall performance. Thus, $\varepsilon^\star$ marks the threshold at which the advantage shifts from adaptive to non-adaptive methods when batch noise is small. By contrast, when batch noise is large, adaptive methods are already known to be more effective (Compagnoni et al., 2025b;a), and the effect of DP noise is only marginal relative to the intrinsic stochasticity of the gradients. We verify this result empirically in Figure 3: As we increase the batch size B, $\varepsilon^\star$ decreases, in accordance with our theoretical prediction.

Practical Implication. If hyperparameter re-tuning is infeasible and the target regime involves stronger privacy constraints, e.g., lower privacy budget ε, or high stochasticity from small batches, adaptive methods are preferable. Otherwise, DP−SGD is the method of choice.

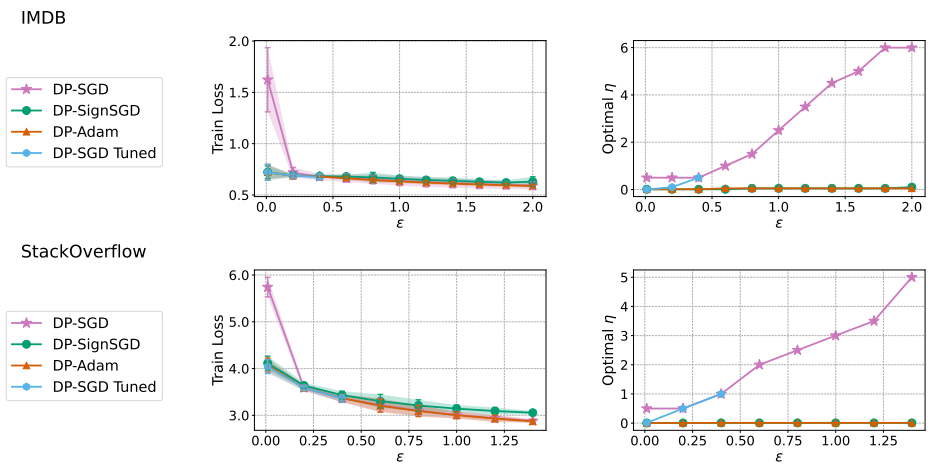

Figure 4: Empirical verification of Thm. 4.6 and Thm. 4.7 under Protocol B on the IMDB dataset **(Top Row)** and on the StackOverflow dataset **(Bottom Row)**. We tune (η, C) of each optimizer for each ε and confirm that: i) all methods achieve comparable performance across privacy budgets; ii) the optimal η of DP−SGD scales linearly with ε, while that of adaptive methods is essentially ε-independent; iii) failing to sweep over the "best" range of learning rates causes DP−SGD to severely underperform, whereas adaptive methods are resilient. On the **left**, DP−SGD degrades sharply for small ε. Indeed, the **right** panels shows that the selected optimal η flattens out, while the theoretical one would have linearly decayed more: The "best" η was simply missing from the grid. *A posteriori*, re-running the sweep with a larger grid (DP−SGD Tuned) recovers the scaling law and matches the performance of adaptive methods. For experimental details see Appendix C.5.

4.2 Protocol B: Best-tuned Hyperparameters

We now mirror standard practice by allowing (η, C) to be *tuned* over an extensive grid search for each target privacy budget ε. In contrast to Protocol A, this leads us to derive the theoretical optimal learning rates, which, just as in empirical tuning, are allowed to depend on ε explicitly.

To select the optimal learning rate $\eta^\star$ for DP−SGD, we minimize the bound in Thm. 4.2 and consequently derive the implied optimal privacy-utility trade-off for DP−SGD in the L-smooth case.

Theorem 4.6 (DP−SGD) *Let* $\eta^\star = \min\left\{\sqrt{\frac{f(X_0)}{dLT\sigma_\gamma^2}}, \sqrt{\frac{f(X_0)}{dL}}\frac{\varepsilon n}{CT}\right\}$, *then the expected gradient norm bound of* DP−SGD *is* $\widetilde{\mathcal{O}}\left(\frac{C\sqrt{dLf(X_0)}}{\varepsilon n}\right)$, *as we ignore logarithmic terms and those decaying in* T.

This result aligns with the best-known privacy-utility trade-off obtained in prior works in these settings (Koloskova et al., 2023; Bassily et al., 2014). Importantly, we notice that the optimal learning rate of DP−SGD scales linearly in ε and that the resulting asymptotic performance scales like $\frac{1}{\varepsilon}$.

To derive the optimal learning rate $\eta^\star$ of DP−SignSGD, we minimize the bound in Theorem 4.4, and derive a privacy-utility trade-off in the L-smooth case.

Theorem 4.7 (DP−SignSGD) *Let* $\eta^\star = \sqrt{\frac{f(X_0)}{dLT}}$. *The expected asymptotic gradient norm bound of* DP−SignSGD *is* $\widetilde{\mathcal{O}}\left(\frac{C\sqrt{dLf(X_0)}}{\varepsilon n}\right)$, *as we ignore logarithmic terms and those decaying in* T.

Importantly, we observe that the asymptotic neighborhood of DP−SignSGD matches that of DP−SGD, while the optimal learning rate is independent of ε. This suggests that adaptivity automatically handles the privacy noise injection: This facilitates the transferability of optimal parameters to setups that require higher privacy. In contrast, DP−SGD needs retuning of the hyperparameters.

Takeaway: Our theory shows that while optimal learning rate scalings differ, the induced neighborhoods match. As shown in Figure 4, our experiments verify that: i) DP−SGD, DP−SignSGD, and DP−Adam exhibit similar asymptotic performance across a broad range of ε, including very small values; ii) the optimal learning rate of DP−SGD is linear in ε, while those of adaptive methods are seemingly independent of it.

Practical implication. Hyperparameter searches are not free under DP: each evaluation consumes a portion of the privacy budget (Papernot & Steinke, 2021), making fine learning-rate grids costly. This asymmetrically impacts the two optimizers. For DP−SGD, the optimal step size scales linearly with ε (Thm. 4.6), so the "right" $\eta^\star$ *moves* as privacy tightens. If a fixed sweep grid misses a value close to $\eta^\star$, the performance of DP−SGD can degrade sharply. This is illustrated in our experiments (Fig. 4): in the left panel, the performance of DP−SGD collapses because the selected "optimal" η plateaus instead of decaying linearly as predicted (right panel) — the true $\eta^\star$ was simply absent from the grid. By contrast, the optimal step size of DP−SignSGD (and empirically DP−Adam) is essentially ε-invariant (Thm. 4.7), so a single well-chosen η transfers across privacy levels with little or no re-tuning. This mechanism also helps explain prior empirical reports that non-adaptive methods deteriorate more severely under stricter privacy (Zhou et al., 2020b, Fig. 1), (Li et al., 2023, Fig. 5), (Asi et al., 2021, Fig. 2): a plausible cause is that their fixed grids did not track the ε-dependent $\eta^\star$ for DP−SGD. Importantly, when both optimizers are carefully tuned, DP−SGD and DP−Adam achieve matching performance in large-scale LLM fine-tuning (Li et al., 2021a, App. S).

5 Conclusion

We studied how differential privacy noise interacts with adaptive compared to non-adaptive optimization through the lens of SDEs: To our knowledge, this is the first SDE-based analysis of DP optimizers. Our results include explicit upper bounds on the expected loss and gradient norm, optimal learning rates, as well as the first characterization of stationary distributions for DP optimizers.

Under a *fixed-hyperparameter* scenario (Protocol A), the analysis reveals a sharp contrast: $i)$ DP−SGD converges at a speed independent of the privacy budget ε while incurring a $\mathcal{O}\left(1/\varepsilon^2\right)$ privacy-utility trade-off; $ii)$ DP−SignSGD converges at a speed proportional to ε while exhibiting a $\mathcal{O}\left(1/\varepsilon\right)$ privacy-utility trade-off. Additionally, when batch noise is large, adaptive methods dominate in terms of utility, as the effect of DP noise is marginal compared to the intrinsic stochasticity of the gradients, confirming known insights from non-private optimization. When batch noise is small, the preferable method depends on the privacy budget: for strict privacy, DP−SignSGD yields better utility, while for looser privacy, DP−SGD achieves better overall performance.

Under a *best-tuned* scenario (Protocol B), the picture changes: theory and experiments agree that the optimal learning rate of DP−SGD scales linearly with ε, whereas the optimal learning rate of DP−SignSGD (and empirically DP−Adam) is approximately ε-independent. With this tuning, the induced privacy-utility trade-offs match in order and the methods achieve comparable asymptotic performance, including at very small ε. A practical implication is that adaptive methods require less re-tuning if regulations mandate tighter privacy budgets.

We validated these theoretical insights on both synthetic and real datasets. Importantly, we also demonstrated that the qualitative behavior observed for DP−SignSGD extends empirically to DP−Adam and to test metrics, underscoring the strength and generality of our framework.

Practitioner guidance. Under higher privacy requirements, e.g., regulations mandate a smaller ε, if per-ε re-tuning of the hyperparameters is impractical because retraining/tuning is expected to be costly (Protocol A), prefer an *adaptive* private optimizer such as DP−SignSGD (or DP−Adam): their performance scales more favorably as ε decreases compared to DP−SGD.

When re-tuning is feasible (Protocol B): Both DP−SGD and adaptive methods can reach comparable asymptotic performance. However, hyperparameter searches are not free under DP: each sweep consumes additional privacy budget (Papernot & Steinke, 2021), making fine grids expensive. This creates an asymmetric risk: DP−SGD requires an ε-dependent learning rate ($\eta^\star \propto \varepsilon$), so if the sweep grid does not track this scaling, its performance can degrade sharply. In contrast, adaptive methods retain a portable, ε-independent learning rate, making them more robust and less costly to tune across privacy levels.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS'16. ACM, October 2016. doi: 10.1145/2976749.2978318. URL http://dx.doi.org/10.1145/2976749.2978318.

Sungjin Ahn, Anoop Korattikara, and Max Welling. Bayesian posterior sampling via stochastic gradient fisher scoring. *arXiv preprint arXiv:1206.6380*, 2012.

Galen Andrew, Om Thakkar, Brendan McMahan, and Swaroop Ramaswamy. Differentially private learning with adaptive clipping. *Advances in Neural Information Processing Systems*, 34:17455–17466, 2021.

Hilal Asi, John Duchi, Alireza Fallah, Omid Javidbakht, and Kunal Talwar. Private adaptive gradient methods for convex optimization, 2021. URL https://arxiv.org/abs/2106.13756.

Borja Balle, Gilles Barthe, and Marco Gaboardi. Privacy amplification by subsampling: Tight analyses via couplings and divergences. *Advances in neural information processing systems*, 31, 2018.

Borja Balle, Gilles Barthe, and Marco Gaboardi. Privacy profiles and amplification by subsampling. *Journal of Privacy and Confidentiality*, 10(1), 2020.

Lukas Balles and Philipp Hennig. Dissecting adam: The sign, magnitude and variance of stochastic gradients. In *International Conference on Machine Learning*, pp. 404–413. PMLR, 2018.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Differentially private empirical risk minimization: Efficient algorithms and tight error bounds. *arXiv preprint arXiv:1405.7085*, 2014.

Tianqi Chen, Emily Fox, and Carlos Guestrin. Stochastic gradient hamiltonian monte carlo. In *International conference on machine learning*, pp. 1683–1691. PMLR, 2014.

Xiangyi Chen, Steven Z Wu, and Mingyi Hong. Understanding gradient clipping in private sgd: A geometric perspective. *Advances in Neural Information Processing Systems*, 33:13773–13782, 2020.

Enea Monzio Compagnoni, Luca Biggio, Antonio Orvieto, Frank Norbert Proske, Hans Kersting, and Aurelien Lucchi. An sde for modeling sam: Theory and insights. In *International Conference on Machine Learning*, pp. 25209–25253. PMLR, 2023.

Enea Monzio Compagnoni, Antonio Orvieto, Hans Kersting, Frank Proske, and Aurelien Lucchi. Sdes for minimax optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 4834–4842. PMLR, 2024.

Enea Monzio Compagnoni, Rustem Islamov, Frank Norbert Proske, and Aurelien Lucchi. Unbiased and sign compression in distributed learning: Comparing noise resilience via SDEs. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025a.

Enea Monzio Compagnoni, Tianlin Liu, Rustem Islamov, Frank Norbert Proske, Antonio Orvieto, and Aurelien Lucchi. Adaptive methods through the lens of sdes: Theoretical insights on the role of noise, 2025b. URL https://arxiv.org/abs/2411.15958.

Enea Monzio Compagnoni, Tianlin Liu, Rustem Islamov, Frank Norbert Proske, Antonio Orvieto, and Aurelien Lucchi. Adaptive methods through the lens of SDEs: Theoretical insights on the role of noise. In *The Thirteenth International Conference on Learning Representations*, 2025c.

Soham De, Leonard Berrada, Jamie Hayes, Samuel L. Smith, and Borja Balle. Unlocking high-accuracy differentially private image classification through scale, 2022. URL https://arxiv.org/abs/2204.13650.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 07 2011.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, August 2014. ISSN 1551-305X. doi: 10.1561/0400000042. URL https://doi.org/10.1561/0400000042.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In Serge Vaudenay (ed.), *Advances in Cryptology - EUROCRYPT 2006*, pp. 486–503, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-34547-3.

EU. Eu ai act: First regulation on artificial intelligence. https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence, 2023. Accessed: 2025-09-24.

EU. Artificial intelligence and next generation technologies. https://www.enisa.europa.eu/topics/artificial-intelligence-and-next-gen-technologies, 2024. Accessed: 2025-09-24.

Arun Ganesh, Brendan McMahan, and Abhradeep Thakurta. On design principles for private adaptive optimizers, 2025. URL https://arxiv.org/abs/2507.01129.

Roan Gylberth, Risman Adnan, Setiadi Yazid, and T. Basaruddin. Differentially private optimization algorithms for deep neural networks. In *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pp. 387–394, 2017. doi: 10.1109/ICACSIS.2017.8355063.

Uwe Helmke and John B Moore. *Optimization and Dynamical Systems*. Springer London, 1st edition, 1994.

White House. Fact sheet: President biden issues executive order on safe, secure, and trustworthy artificial intelligence. https://bidenwhitehouse.archives.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/, 2023. Accessed: 2025-09-24.

Stanisław Jastrzebski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Three factors influencing minima in sgd. *ICANN 2018*, 2018.

Richeng Jin and Huaiyu Dai. Noisy SIGNSGD is more differentially private than you (might) think. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=thCqMz1ZXw.

Kaggle. Stack overflow data on kaggle. https://www.kaggle.com/datasets/stackoverflow/stackoverflow, 2022.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Anastasia Koloskova, Hadrien Hendrikx, and Sebastian U. Stich. Revisiting Gradient Clipping: Stochastic bias and tight convergence guarantees. In *ICML 2023 - 40th International Conference on Machine Learning*, pp. 1–19, Honolulu, Hawaii,, United States, July 2023. URL https://openreview.net/pdf?id=C3DXiFTrve.

Harold Kushner and G George Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.

Bingrui Li, Wei Huang, Andi Han, Zhanpeng Zhou, Taiji Suzuki, Jun Zhu, and Jianfei Chen. On the optimization and generalization of two-layer transformers with sign gradient descent, 2025. URL https://arxiv.org/abs/2410.04870.

Qianxiao Li, Cheng Tai, and E Weinan. Stochastic modified equations and adaptive stochastic gradient algorithms. In *International Conference on Machine Learning*, pp. 2101–2110. PMLR, 2017.

Qianxiao Li, Cheng Tai, and E Weinan. Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations. *The Journal of Machine Learning Research*, 20 (1):1474–1520, 2019.

Tian Li, Manzil Zaheer, Sashank J. Reddi, and Virginia Smith. Private adaptive optimization with side information, 2022. URL https://arxiv.org/abs/2202.05963.

Tian Li, Manzil Zaheer, Ken Ziyu Liu, Sashank J. Reddi, H. Brendan McMahan, and Virginia Smith. Differentially private adaptive optimization with delayed preconditioners, 2023. URL https://arxiv.org/abs/2212.00309.

Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. *arXiv preprint arXiv:2110.05679*, 2021a.

Zhiyuan Li, Sadhika Malladi, and Sanjeev Arora. On the validity of modeling SGD with stochastic differential equations (SDEs). In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021b.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea (eds.), *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://aclanthology.org/P11-1015/.

Sadhika Malladi, Kaifeng Lyu, Abhishek Panigrahi, and Sanjeev Arora. On the SDEs and scaling rules for adaptive gradient algorithms. In *Advances in Neural Information Processing Systems*, 2022.

Stephan Mandt, Matthew Hoffman, and David Blei. A variational analysis of stochastic gradient algorithms. In *International conference on machine learning*, pp. 354–363. PMLR, 2016.

Stephan Mandt, Matthew D Hoffman, and David M Blei. Stochastic gradient descent as approximate bayesian inference. *JMLR 2017*, 2017.

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.

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, 2025. URL https://arxiv.org/abs/2501.18914.

H. Brendan McMahan and Matthew Streeter. Adaptive bound optimization for online convex optimization, 2010. URL https://arxiv.org/abs/1002.4908.

Panayotis Mertikopoulos and Mathias Staudigl. On the convergence of gradient-like flows with noisy gradient input. *SIAM Journal on Optimization*, 28(1):163–197, 2018.

GN Mil'shtein. Weak approximation of solutions of systems of stochastic differential equations. *Theory of Probability & Its Applications*, 30(4):750–766, 1986.

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

Ilya Mironov, Kunal Talwar, and Li Zhang. Rényi differential privacy of the sampled gaussian mechanism, 08 2019.

NITS. Nist offers draft guidance for evaluating privacy protection technique in the ai era. https://www.nist.gov/news-events/news/2023/12/nist-offers-draft-guidance-evaluating-privacy-protection-technique-ai-era, 2023a. Accessed: 2025-09-24.

NITS. Artificial intelligence risk management framework (ai rmf 1.0). Technical report, National Institute of Standards and Technology, 2023b. URL https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

Nicolas Papernot and Thomas Steinke. Hyperparameter tuning with renyi differential privacy. *arXiv preprint arXiv:2110.03620*, 2021.

Courtney Paquette, Kiwon Lee, Fabian Pedregosa, and Elliot Paquette. Sgd in the large: Average-case analysis, asymptotics, and stepsize criticality. In *Conference on Learning Theory*, pp. 3548–3626. PMLR, 2021.

Hanyang Peng, Shuang Qin, Yue Yu, Fangqing Jiang, Hui Wang, and Zhouchen Lin. Simple convergence proof of adam from a sign-like descent perspective, 2025. URL https://arxiv.org/abs/2507.05966.

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.

Maxim Raginsky and Jake Bouvrie. Continuous-time stochastic mirror descent on a network: Variance reduction, consensus, convergence. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pp. 6793–6800. IEEE, 2012.

Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, 2019.

Samuel L. Smith, Benoit Dherin, David G. T. Barrett, and Soham De. On the origin of implicit regularization in stochastic gradient descent. *ArXiv*, abs/2101.12176, 2021.

Mandt Stephan, Matthew D Hoffman, David M Blei, et al. Stochastic gradient descent as approximate bayesian inference. *Journal of Machine Learning Research*, 18(134):1–35, 2017.

Qiaoyue Tang and Mathias Lécuyer. Dp-adam: Correcting dp bias in adam's second moment estimation, 2023. URL https://arxiv.org/abs/2304.11208.

Qiaoyue Tang, Frederick Shpilevskiy, and Mathias Lécuyer. Dp-adambc: Your dp-adam is actually dp-sgd (unless you apply bias correction), 2023. URL https://arxiv.org/abs/2312.14334.

TensorFlow Federated. Tensorflow federated stack overflow dataset. https://www.tensorflow.org/federated/api_docs/python/tff/simulation/datasets/stackoverflow/load_data, 2022.

Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude., 2012. URL https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf.

Roman Vershynin. *High-dimensional probability : an introduction with applications in data science*. Cambridge series in statistical and probabilistic mathematics ; 47. Cambridge University Press, Cambridge, United Kingdom ;, 2018. ISBN 9781108415194.

Yu-Xiang Wang, Borja Balle, and Shiva Prasad Kasiviswanathan. Subsampled renyi differential privacy and analytical moments accountant. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1226–1235. PMLR, 16–18 Apr 2019.

Jingfeng Wu, Wenqing Hu, Haoyi Xiong, Jun Huan, Vladimir Braverman, and Zhanxing Zhu. On the noisy gradient descent that generalizes as sgd. In *International Conference on Machine Learning*, pp. 10367–10376. PMLR, 2020.

Ke Liang Xiao, Noah Marshall, Atish Agarwala, and Elliot Paquette. Exact risk curves of signSGD in high-dimensions: quantifying preconditioning and noise-compression effects. In *Forty-second International Conference on Machine Learning*, 2025.

Zeke Xie, Li Yuan, Zhanxing Zhu, and Masashi Sugiyama. Positive-negative momentum: Manipulating stochastic gradient noise to improve generalization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11448–11458. PMLR, 18–24 Jul 2021.

Jim Zhao, Aurelien Lucchi, Frank Norbert Proske, Antonio Orvieto, and Hans Kersting. Batch size selection by stochastic optimal control. In *Has it Trained Yet? NeurIPS 2022 Workshop*, 2022.

Pan Zhou, Jiashi Feng, Chao Ma, Caiming Xiong, Steven Chu Hong Hoi, et al. Towards theoretically understanding why sgd generalizes better than adam in deep learning. *Advances in Neural Information Processing Systems*, 33:21285–21296, 2020a.

Yingxue Zhou, Xiangyi Chen, Mingyi Hong, Zhiwei Steven Wu, and Arindam Banerjee. Private stochastic non-convex optimization: Adaptive algorithms and tighter generalization bounds, 2020b. URL https://arxiv.org/abs/2006.13501.

Zhanxing Zhu, Jingfeng Wu, Bing Yu, Lei Wu, and Jinwen Ma. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. *ICML*, 2019.

Difan Zou, Yuan Cao, Yuanzhi Li, and Quanquan Gu. Understanding the generalization of adam in learning neural networks with proper regularization, 2021. URL https://arxiv.org/abs/2108.11371.

Appendix

CONTENTS

A TECHNICAL RESULTS

In this section, we introduce some technical results used in the derivation of the SDEs.

Lemma A.1 *Let $X \sim \mathcal{N}(\mu, \sigma^2 I_d)$ and fix a tolerance $\epsilon > 0$. If $\frac{\|\mu\|_2^2}{2\sigma^2(d+2)} < \epsilon$, for $d \to \infty$, we have that $\mathbb{E}\left(\frac{X}{\|X\|_2}\right) = \sqrt{\frac{1}{d}}\frac{\mu}{\sigma} + \varepsilon\mathcal{O}\left(\frac{1}{d^{3/2}}\right)$.*

Proof: Let us remember that if $X \sim \mathcal{N}(\mu, \sigma^2 I_d)$,

$$\mathbb{E}\left(\frac{X}{\|X\|_2^k}\right) = \frac{\Gamma\left(\frac{d}{2}+1-\frac{k}{2}\right)}{(2\sigma^2)^{k/2}\,\Gamma\left(\frac{d}{2}+1\right)}\,{}_1F_1\left(\frac{k}{2}; \frac{d+2}{2}; -\frac{\|\mu\|_2^2}{2\sigma^2}\right)\mu, \tag{12}$$

where ${}_1F_1(a; b; z)$ is Kummer's confluent hypergeometric function. We know that

$$\lim_{d\to\infty} \frac{\Gamma\left(\frac{d}{2}+1-\frac{1}{2}\right)}{\Gamma\left(\frac{d}{2}+1\right)} \overset{k=1}{\sim} \sqrt{\frac{2}{d}} + \mathcal{O}\left(\frac{1}{d^{3/2}}\right). \tag{13}$$

Let $z = \frac{\|\mu\|_2^2}{2\sigma^2}$. If $d > z$, by expanding the series, we have

$${}_1F_1\left(\frac{1}{2}; \frac{d}{2}+1; -z\right) = \sum_{n\geq 0} \frac{a^{(n)}(-z)^n}{b^{(n)}n!} = 1 - \frac{z}{d+2} + \mathcal{O}\left(\frac{z}{d}\right)^2 < 1 - \epsilon. \tag{14}$$

Combining everything together, we obtain $\mathbb{E}\left(\frac{X}{\|X\|_2}\right) = \sqrt{\frac{1}{d}}\frac{\mu}{\sigma} + \epsilon\mathcal{O}\left(\frac{1}{d^{3/2}}\right)$. □

Lemma A.2 *Let $K(\nu) = \sqrt{\frac{2}{\nu}}\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}$ and $X \sim t_\nu(\mu, \sigma^2 I_d)$, for $\nu \geq 1$. Fix a tolerance $\epsilon > 0$: If $\frac{\|\mu\|_2^2}{2\sigma^2(d+2)} < \epsilon$, for $d \to \infty$, we have that $\mathbb{E}\left(\frac{X}{\|X\|_2}\right) = K(\nu)\sqrt{\frac{1}{d}}\frac{\mu}{\sigma} + \epsilon\mathcal{O}\left(\frac{1}{d^{3/2}}\right)$.*

Proof: One can write $X = \mu + \frac{\sigma Z}{\sqrt{S/\nu}}$, where $Z \sim \mathcal{N}(0, I_d)$ and $S \sim \chi^2_\nu$ are independent. Define $\tau = \frac{\sigma}{\sqrt{S/\nu}}$, then, conditioning on S and applying Lemma A.1, we have

$$\mathbb{E}\left[\frac{X}{\|X\|_2}\middle| S\right] = \sqrt{\frac{1}{d}}\frac{\mu}{\tau} + \epsilon\mathcal{O}\left(d^{-3/2}\right). \tag{15}$$

Remembering that $\mathbb{E}[\sqrt{S}] = \sqrt{2}\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}$, we have

$$\mathbb{E}\left[\frac{X}{\|X\|_2}\right] = \mathbb{E}_S\left[\left(\sqrt{\frac{1}{d}}\frac{\mu}{\tau} + \epsilon\mathcal{O}\left(d^{-3/2}\right)\right)\right] \tag{16}$$

$$= \left(\sqrt{\frac{1}{d}}\frac{\mu}{\sigma\sqrt{\nu}}\mathbb{E}_S[\sqrt{S}] + \epsilon\mathcal{O}\left(d^{-3/2}\right)\right) \tag{17}$$

$$= K(\nu)\sqrt{\frac{1}{d}}\frac{\mu}{\sigma} + \epsilon\mathcal{O}\left(\frac{1}{d^{3/2}}\right). \tag{18}$$

□

Remark A.1 *As discussed in the main paper, our analysis is based on the following two assumptions:*

i) The number of trainable parameters is large, specifically $d = \Omega(10^4)$;

ii) The signal-to-noise ratio satisfies $\frac{\|\nabla f(x)\|_2^2}{2\sigma_\gamma^2} \ll d$.

First, they ensure that the approximation of the confluent hypergeometric function in Equation 14 is highly accurate. Second, neither condition is restrictive for modern deep learning models. The dimensionality assumption is trivially satisfied by contemporary architectures, which routinely have millions of parameters. Regarding the second assumption, Malladi et al. (2022) empirically measured the signal-to-noise ratio $\frac{\|\nabla f(x)\|_2^2}{2\sigma_\gamma^2}$ across a wide range of large-scale architectures and datasets, and consistently found values of at most $O(10^2)$. Thus, the regime in which our approximation is valid closely matches the regime observed in practice. This is further supported by our experimental results, which confirm our theoretical predictions across all models and tasks considered in this paper. In Figure A.1, we numerically evaluate the confluent hypergeometric function for varying values of d, and show that, for sufficiently large parameter counts, the approximation remains tight throughout the realistic signal-to-noise ratio range reported in (Malladi et al., 2022).

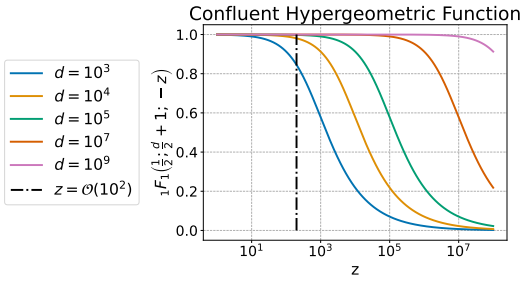

Figure A.1: Numerical validation of the approximation used in Equation 14. For several values of d, we plot the confluent hypergeometric function as a function of the signal-to-noise ratio z. In the realistic range observed in (Malladi et al., 2022), approximating this function by 1 is extremely accurate.

B THEORETICAL FRAMEWORK

In this section, we introduce the theoretical framework, assumptions, and notations used to formally derive the SDE models used in this paper. We briefly recall the definition of L-smoothness and μ-PL functions. Then we introduce the set of functions of polynomial growth G.

Definition B.1 *A function $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth if it is differentiable and its gradient is L-Lipschitz continuous, namely*

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2 \ \forall x, y. \tag{19}$$

Definition B.2 *A function $f : \mathbb{R}^d \to \mathbb{R}$ admitting a global minima x^* satisfies the Polyak-Łojasiewicz inequality if, for some $\mu > 0$ and for all $x \in \mathbb{R}^d$, it holds*

$$f(x) - f^* \leq \frac{1}{2\mu}\|\nabla f(x)\|_2^2. \tag{20}$$

In this case, we say that the function f is μ-PL.

Definition B.3 *Let G denote the set of continuous functions $g : \mathbb{R}^d \to \mathbb{R}$ of at most polynomial growth, namely such that there exist positive integers $k_1, k_2 > 0$ such that $|g(x)| < k_1(1 + \|x\|_2^2)^{k_2}$, for all $x \in \mathbb{R}^d$.*

To simplify the notation, we will write

$$b(x + \eta) = b_0(x) + \eta b_1(x) + \mathcal{O}(\eta^2),$$

whenever there exists $g \in G$, independent of η, such that

$$|b(x + \eta) - b_0(x) - \eta b_1(x)| \leq g(x)\eta^2.$$

We now introduce the definition of weak approximation, which formalizes in which sense the solution to an SDE, which is a continuous-time random process, models a discrete-time optimizer.

Definition B.4 *Let $0 < \eta < 1$, $\tau > 0$ and $T = \lfloor \frac{\tau}{\eta} \rfloor$. We say that a continuous time process X_t over $[0, \tau]$, is an order α weak approximation of a discrete process x_k, for $k = 0, \dots, N$, if for every $g \in G$, there exists M, independent of η, such that for all $k = 0, 1, \dots, N$*

$$|\mathbb{E}g(X_{k\eta}) - \mathbb{E}g(x_k)| \leq M\eta^\alpha.$$

This framework focuses on approximation in a *weak sense*, meaning in distribution rather than pathwise. Since G contains all polynomials, all the moments of both processes become closer at a rate of η^α and thus their distributions. Thus, while the processes exhibit similar average behavior, their sample paths may differ significantly, justifying the term weak approximation.

Remark B.1 *Our continuous-time models are derived using the standard order-1 weak-approximation framework (see Section 2). In particular, Definition B.4 together with Theorems B.5 and B.10 show that the drift and covariance of the discrete updates match those of the SDE up to $O(\eta^2)$, so that the weak error over finite horizons is $O(\eta)$. In line with the existing literature, we therefore restrict attention to first-order SDEs; higher-order SDEs have been derived in special cases but, to the best of our knowledge, have not led to additional practical insights. Finally, Figure C.1 empirically confirms that our SDEs closely track their corresponding algorithms on simple landscapes, following the standard validation practice in the field (Compagnoni et al., 2025c).*

The key ingredient for deriving the SDE is given by the following result (see Theorem 1, (Li et al., 2017)), which provides sufficient conditions to get a weak approximation in terms of the single step increments of both X_t and x_k. Before stating the theorem, we list the regularity assumption under which we are working.

Assumption B.1 *Assume that the following conditions are satisfied:*

- $f, f_i \in \mathcal{C}_b^8(\mathbb{R}^d, \mathbb{R})$;

- *f, f_i and its partial derivatives up to order 7 belong to G;*
- *$\nabla f, \nabla f_i$ satisfy the following Lipschitz condition: there exists $L > 0$ such that*

$$\|\nabla f(u) - \nabla f(v)\|_2 + \sum_{i=1}^{d} \|\nabla f_i(u) - \nabla f_i(v)\|_2 \leq L\|u - v\|_2;$$

- *$\nabla f, \nabla f_i$ satisfy the following growth condition: there exists $M > 0$ such that*

$$\|\nabla f(x)\|_2 + \sum_{i=1}^{n} \|\nabla f_i(x)\|_2 \leq M(1 + \|x\|_2).$$

Assumption B.2 *Assume that the stochastic gradient can be written as $\nabla f_\gamma = \nabla f + Z_\gamma$. In Phase 1 (clipping regime), the batch noise Z_γ is modeled as heavy-tailed, e.g., a Student-t distribution with ν degrees of freedom and scale σ_γ: for $\nu = \infty$ we recover the Gaussian case, while if $\nu < 2$ the variance is unbounded and if $\nu = 1$ the distribution becomes a Cauchy, therefore the expectation is unbounded as well. In Phase 2 (non-clipping regime), the batch noise is modeled as a Gaussian of variance $\frac{\sigma_\gamma^2}{B}$, reflecting the averaging effect of i.i.d., per-sample gradients.*

Remark B.2 *The distinction between the two phases stems from the effect of per-example clipping on the noise distribution. In Phase 1, clipping is applied* before *the batch average, so the noise of each individual stochastic gradient is not smoothed by averaging and can remain strongly heavy-tailed; in this regime, a Gaussian model is no longer appropriate. We therefore model the Phase 1 noise as a multivariate Student-t, which both captures this heavy-tailed behaviour and admits tractable expressions for our SDE analysis, while recovering the Gaussian model used in Phase 2 in the limit of large degrees of freedom.*

Lemma B.3 *Let $0 < \eta < 1$. Consider a stochastic process $X_t, t \geq 0$ satisfying the SDE*

$$dX_t = b(X_t)dt + \sqrt{\eta}\sigma(X_t)dW_t, \qquad X_0 = x \tag{21}$$

where b, σ together with their derivatives belong to G. Define the one-step difference $\Delta = X_\eta - x$, and indicate the i-th component of Δ with Δ_i. Then we have

1. $\mathbb{E}\Delta_i = b_i\eta + \frac{1}{2}\left[\sum_{j=1}^{d} b_j \partial_j b_i\right]\eta^2 + \mathcal{O}(\eta^3) \qquad \forall i = 1, \ldots, d;$
2. $\mathbb{E}\Delta_i\Delta_j = \left[b_i b_j + \sigma\sigma_{ij}^\top\right]\eta^2 + \mathcal{O}(\eta^3) \qquad \forall i, j = 1, \ldots, d;$
3. $\mathbb{E}\prod_{j=1}^{s} \Delta_{i_j} = \mathcal{O}(\eta^3) \qquad \forall s \geq 3,\ i_j = 1, \ldots, d.$

All functions above are evaluated at x.

Theorem B.4 *Let $0 < \eta < 1$, $\tau > 0$ and set $T = \lfloor \tau/\eta \rfloor$. Let Assumption B.1 hold and let X_t be a stochastic process as in Lemma B.3. Define $\bar{\Delta} = x_1 - x$ to be the increment of the discrete-time algorithm, and indicate the i-th component of $\bar{\Delta}$ with $\bar{\Delta}_i$. If in addition there exist $K_1, K_2, K_3, K_4 \in G$ so that*

1. $\left|\mathbb{E}\Delta_i - \mathbb{E}\bar{\Delta}_i\right| \leq K_1(x)\eta^2, \qquad \forall i = 1, \ldots, d;$
2. $\left|\mathbb{E}\Delta_i\Delta_j - \mathbb{E}\bar{\Delta}_i\bar{\Delta}_j\right| \leq K_2(x)\eta^2, \qquad \forall i, j = 1, \ldots, d;$
3. $\left|\mathbb{E}\prod_{j=1}^{s} \Delta_{i_j} - \mathbb{E}\prod_{j=1}^{s} \bar{\Delta}_{i_j}\right| \leq K_3(x)\eta^2, \qquad \forall s \geq 3, \forall i_j = 1, \ldots, d;$
4. $\mathbb{E}\prod_{j=1}^{s} \left|\bar{\Delta}_{i_j}\right| \leq K_4(x)\eta^2, \qquad \forall i_j = 1, \ldots, d.$

Then, there exists a constant C so that for all $k = 0, 1, \ldots, N$ we have

$$|\mathbb{E}g(X_{k\eta}) - \mathbb{E}g(x_k)| \leq C\eta. \tag{22}$$

We say Eq. 21 is an order 1 weak approximation of the update step of x_k.

B.1 DP-SGD

This subsection provides the formal derivation of the SDE model for DP-SGD and formal statements of Theorem 4.1, Theorem 4.2, and Theorem B.9. Since, by construction, the dynamic of the method shifts stochastically between two phases, we first model and study each phase separately.

Theorem B.5 *Let $0 < \eta < 1$, $\tau > 0$ and set $T = \lfloor \tau/\eta \rfloor$ and $K(\nu) = \sqrt{\frac{2}{\nu}}\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}$. Let $x_k \in \mathbb{R}^d$ denote a sequence of DP-SGD iterations defined in Eq. 4. Assume Assumption B.1 and Assumption B.2. Let X_t be the solution of the following SDEs with initial condition $X_0 = x_0$:*

• *Phase 1:*

$$dX_t = -\frac{CK(\nu)}{\sigma_\gamma\sqrt{d}}\nabla f(X_t)dt + \sqrt{\eta}\sqrt{\bar{\Sigma}(X_t)}dW_t, \tag{23}$$

where $\bar{\Sigma}(x) = C^2\left(\mathbb{E}\left[\frac{\nabla f_\gamma(x)\nabla f_\gamma(x)^\top}{\|\nabla f_\gamma(x)\|_2^2}\right] - \frac{K(\nu)^2}{\sigma_\gamma\sqrt{d}}\nabla f(x)\nabla f(x)^\top + \frac{\sigma_{DP}^2}{B^2}I_d\right)$.

• *Phase 2:*

$$dX_t = -\nabla f(X_t)dt + \sqrt{\eta}\sqrt{\bar{\Sigma}(X_t)}dW_t, \tag{24}$$

where $\bar{\Sigma}(x) = \left(\frac{\sigma_\gamma^2}{B} + \frac{C^2\sigma_{DP}^2}{B^2}\right)I_d$.

Then, Eq. 23 and Eq. 24 are an order 1 *approximation of the discrete update of Phase 1 and Phase 2 of DP-SGD, respectively.*

Proof: • Phase 1: Let $Z_{DP} \sim \mathcal{N}\left(0, \frac{C^2\sigma_{DP}^2}{B^2}I_d\right)$ be the differentially-private noise injected via Gaussian Mechanism and denote with $\bar{\Delta} = x_1 - x$ the one-step increment for Phase 1. Applying Lemma A.2 with tolerance $\epsilon = \eta$ and by definition we have

$$\mathbb{E}[\bar{\Delta}] = -\eta\mathbb{E}_{\gamma,DP}\left[C\frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|_2} + Z_{DP}\right] = -\eta\frac{CK(\nu)}{\sigma_\gamma\sqrt{d}}\nabla f(x) + \mathcal{O}(\eta^2). \tag{25}$$

Then, the second moment becomes

$$\begin{aligned}
\mathrm{Cov}(\bar{\Delta}) =& \mathbb{E}\bar{\Delta}\bar{\Delta}^\top - \mathbb{E}\left[\bar{\Delta}\right]\mathbb{E}\left[\bar{\Delta}^\top\right] && (26)\\
=& \eta^2\mathbb{E}_{\gamma,DP}\left[\left(C\frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|_2} + Z_{DP} - \frac{CK(\nu)}{\sigma_\gamma\sqrt{d}}\nabla f(x) + \mathcal{O}(\eta^2)\right)\right. && (27)\\
& \left.\left(C\frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|_2} + Z_{DP} - \frac{CK(\nu)}{\sigma_\gamma\sqrt{d}}\nabla f(x) + \mathcal{O}(\eta^2)\right)^\top\right] && (28)\\
=& \eta^2\left(C^2\mathbb{E}_{\gamma,DP}\left[\frac{\nabla f_\gamma(x)\nabla f_\gamma(x)^\top}{\|\nabla f_\gamma(x)\|_2^2}\right] + Z_{DP}Z_{DP}^\top\right. && (29)\\
& \left. - \frac{C^2K(\nu)^2}{\sigma_\gamma^2 d}\nabla f(x)\nabla f(x)^\top\right) + \mathcal{O}(\eta^4) && (30)\\
=& \eta^2\left(C^2\mathbb{E}\left[\frac{\nabla f_\gamma(x)\nabla f_\gamma(x)^\top}{\|\nabla f_\gamma(x)\|_2^2}\right] - \frac{C^2K(\nu)^2}{\sigma_\gamma\sqrt{d}}\nabla f(x)\nabla f(x)^\top + \frac{C^2\sigma_{DP}^2}{B^2}I_d\right) + \mathcal{O}(\eta^4). && (31)
\end{aligned}$$

Define now

$$\begin{aligned}
b(x) &:= -\frac{CK(\nu)}{\sigma_\gamma\sqrt{d}}\nabla f(x) && (32)\\
\bar{\Sigma}(x). &:= C^2\mathbb{E}\left[\frac{\nabla f_\gamma(x)\nabla f_\gamma(x)^\top}{\|\nabla f_\gamma(x)\|_2^2}\right] - \frac{C^2K(\nu)^2}{\sigma_\gamma\sqrt{d}}\nabla f(x)\nabla f(x)^\top + \frac{C^2\sigma_{DP}^2}{B^2}I_d. && (33)
\end{aligned}$$

Then, from Lem B.3 and Thm. B.4 the claim follows.

• Phase 2: Following the same steps as above, one obtains:

$$\mathbb{E}[\bar{\Delta}] = -\eta \mathbb{E}_{\gamma,DP}\left[\nabla f_\gamma(x) + Z_{DP}\right] = -\eta \nabla f(x), \tag{34}$$

and

$$\mathrm{Cov}(\bar{\Delta}) = \eta^2 \mathbb{E}\left[\left(\nabla f_\gamma(x) + Z_{DP} - \nabla f(x)\right)\left(\nabla f_\gamma(x) + Z_{DP} - \nabla f(x)\right)^\top\right] \tag{35}$$

$$= \eta^2 \mathbb{E}\left[(\nabla f_\gamma(x) - \nabla f(x))(\nabla f_\gamma(x) - \nabla f(x))^\top\right] + \eta^2 \frac{C^2\sigma_{DP}^2}{B^2} I_d \tag{36}$$

$$= \eta^2 \left(\frac{\sigma_\gamma^2}{B} + \frac{C^2\sigma_{DP}^2}{B^2}\right) I_d \tag{37}$$

Define

$$b(x) \coloneqq -\nabla f(x). \tag{38}$$

$$\bar{\Sigma}(x) \coloneqq \left(\frac{\sigma_\gamma^2}{B} + \frac{C^2\sigma_{DP}^2}{B^2}\right) I_d. \tag{39}$$

Finally, from Lem B.3 and Thm. B.4 the claim follows. □

Theorem B.6 *Let f be L-smooth and μ-PL. Then, for $t \in [0, \tau]$, we have that*

• *Phase 1, i.e., when the gradient is clipped, the loss satisfies:*

$$\mathbb{E}[f(X_t)] \lesssim f(X_0) e^{-\frac{\mu C}{\sigma_\gamma \sqrt{d}} t} + \left(1 - e^{-\frac{\mu C}{\sigma_\gamma \sqrt{d}} t}\right) \frac{T\eta d^{\frac{3}{2}} LC\sigma_\gamma}{\mu} \left(\frac{\varepsilon^2}{dT} + \frac{\Phi^2}{B^2}\right) \frac{1}{\varepsilon^2}; \tag{40}$$

• *Phase 2, i.e., when the gradient is not clipped, the loss satisfies:*

$$\mathbb{E}[f(X_t)] \lesssim f(X_0) e^{-\mu t} + \left(1 - e^{-\mu t}\right) \frac{T\eta dL}{\mu} \left(\frac{\varepsilon^2 {\sigma_\gamma}^2}{BT} + C^2 \frac{\Phi^2}{B^2}\right) \frac{1}{\varepsilon^2}. \tag{41}$$

Proof: • Phase 1: By construction we have

$$\mathrm{Tr}\left(\bar{\Sigma}(x)\right) \leq C^2 + d\frac{C^2\sigma_{DP}^2}{B^2}. \tag{42}$$

Since f is μ-PL and L-smooth it follows that $2\mu f(x) \leq \|\nabla f(x)\|_2^2$ and $\nabla^2 f(x) \preceq L I_d$. Hence, by applying the Itô formula we have

$$df(X_t) = -\frac{CK(\nu)}{\sigma_\gamma \sqrt{d}} \|\nabla f(X_t)\|_2^2 dt + \frac{\eta}{2} \mathrm{Tr}\left(\nabla^2 f(X_t) \bar{\Sigma}(X_t)\right) dt + \mathcal{O}(\text{Noise}) \tag{43}$$

$$\leq -2\mu \frac{CK(\nu)}{\sigma_\gamma \sqrt{d}} f(X_t) dt + \frac{\eta dL}{2} \left(\frac{C^2}{d} + \frac{C^2\sigma_{DP}^2}{B^2}\right) dt + \mathcal{O}(\text{Noise}). \tag{44}$$

Therefore,

$$\mathbb{E}[f(X_t)] \leq f(X_0) e^{-2\mu \frac{K(\nu)C}{\sigma_\gamma \sqrt{d}} t} + \left(1 - e^{-2\mu \frac{K(\nu)C}{\sigma_\gamma \sqrt{d}} t}\right) \frac{\eta d^{\frac{3}{2}} L\sigma_\gamma}{4\mu CK(\nu)} \left(\frac{C^2}{d} + \frac{C^2\sigma_{DP}^2}{B^2}\right). \tag{45}$$

Let us now remind that

$$\sigma_{DP} = \frac{q\sqrt{T\log(1/\delta)}}{\varepsilon}, \tag{46}$$

then

$$\mathbb{E}[f(X_t)] \leq f(X_0) e^{-2\mu \frac{K(\nu)C}{\sigma_\gamma \sqrt{d}} t} + \left(1 - e^{-2\mu \frac{K(\nu)C}{\sigma_\gamma \sqrt{d}} t}\right) \frac{\eta d^{\frac{3}{2}} LC\sigma_\gamma}{4\mu K(\nu)} \left(\frac{1}{d} + \frac{Tq^2\log(1/\delta)}{B^2\varepsilon^2}\right). \tag{47}$$

• Phase 2: Similarly to Phase 1, we have

$$\operatorname{Tr}\left(\bar{\Sigma}(x)\right) = d\left(\frac{\sigma_\gamma^2}{B} + \frac{C^2\sigma_{DP}^2}{B^2}\right). \tag{48}$$

Again using the fact that f is μ-PL and L-smooth and by applying the Itô formula, one obtains

$$df(X_t) \leq -\|\nabla f(X_t)\|_2^2 dt + \frac{\eta dL}{2}\left(\frac{\sigma_\gamma^2}{B} + \frac{C^2\sigma_{DP}^2}{B^2}\right) + \mathcal{O}(\text{Noise}) \tag{49}$$

from which we have

$$\mathbb{E}[f(X_t)] \leq f(X_0)e^{-2\mu t} + \left(1 - e^{-2\mu t}\right)\frac{\eta dL}{4\mu}\left(\frac{\sigma_\gamma^2}{B} + \frac{C^2\sigma_{DP}^2}{B^2}\right). \tag{50}$$

Hence, by expanding σ_{DP}

$$\mathbb{E}[f(X_t)] \leq f(X_0)e^{-2\mu t} + \left(1 - e^{-2\mu t}\right)\frac{\eta dL}{4\mu}\left(\frac{\sigma_\gamma^2}{B} + \frac{C^2q^2T\log(1/\delta)}{B^2\varepsilon^2}\right). \tag{51}$$

Finally, let $\Phi = q\sqrt{\log(1/\delta)}$ and suppress all problem-independent constants, such as $2, \pi, K(\nu)$, to obtain the claim.

□

Theorem B.7 *Let f be L-smooth and define*

$$K_1 \coloneqq \max\left\{1, \frac{\sigma_\gamma\sqrt{d}}{CK(\nu)}\right\} \quad K_2 \coloneqq \max\left\{\frac{C^2}{d}, \frac{\sigma_\gamma^2}{B}\right\}. \tag{52}$$

then

$$\mathbb{E}\left[\|\nabla f(X_{\tilde{t}})\|_2^2\right] \lesssim K_1\left(\frac{f(X_0)}{\eta T} + \frac{\eta dL}{2}\left(K_2 + \frac{C^2\left(\frac{q}{B}\right)^2 T\log(1/\delta)}{\varepsilon^2}\right)\right), \tag{53}$$

where $\tilde{t} \sim \operatorname{Unif}(0, \tau)$.

Proof: Since f is L-smooth and by applying the Itô formula to Phase 1 we have:

$$df(X_t) \leq -\frac{CK(\nu)}{\sigma_\gamma\sqrt{d}}\|\nabla f(X_t)\|_2^2 dt + \frac{\eta}{2}\operatorname{Tr}\left(L\bar{\Sigma}(X_t)\right)dt + \mathcal{O}(\text{Noise}) \tag{54}$$

$$\leq -\frac{CK(\nu)}{\sigma_\gamma\sqrt{d}}\|\nabla f(X_t)\|_2^2 dt + \frac{\eta dL}{2}\left(\frac{C^2}{d} + \frac{C^2\sigma_{DP}^2}{B^2}\right)dt + \mathcal{O}(\text{Noise}). \tag{55}$$

Similarly, in Phase 2 we obtain

$$df(X_t) \leq -\|\nabla f(X_t)\|_2^2 dt + \frac{\eta}{2}\operatorname{Tr}\left(L\bar{\Sigma}(X_t)\right)dt + \mathcal{O}(\text{Noise}) \tag{56}$$

$$\leq -\|\nabla f(X_t)\|_2^2 dt + \frac{\eta dL}{2}\left(\frac{\sigma_\gamma^2}{B} + \frac{C^2\sigma_{DP}^2}{B^2}\right)dt + \mathcal{O}(\text{Noise}). \tag{57}$$

Let K_1 and K_2 as in Eq. 52. Then, by integrating and taking the expectation, we have

$$\mathbb{E}\int_0^\tau \|\nabla f(X_t)\|_2^2 dt \leq K_1\left(f(X_0) - f(X_\tau) + \frac{\tau\eta dL}{2}\left(K_2 + \frac{C^2\sigma_{DP}^2}{B^2}\right)\right) \tag{58}$$

$$\implies \mathbb{E}\int_0^\tau \frac{1}{\tau}\|\nabla f(X_t)\|_2^2 dt \leq K_1\left(\frac{f(X_0) - f(X_\tau)}{\tau} + \frac{\eta dL}{2}\left(K_2 + \frac{C^2\sigma_{DP}^2}{B^2}\right)\right) \tag{59}$$

$$\implies \mathbb{E}\left[\|\nabla f(X_{\tilde{t}})\|_2^2\right] \leq K_1\left(\frac{2\varepsilon^2(f(X_0) - f(X_\tau)) + \eta^2 dLTK_2}{2\eta T} + \frac{\eta dLTC^2q^2\log(1/\delta)}{B^2}\right)\frac{1}{\varepsilon^2}$$

where the last step follows from the Law of the Unconscious Statistician and $\tilde{t} \sim \operatorname{Unif}(0, \tau)$. Finally, by suppressing problem-independent constants, $2, \pi$, we obtain the claim.

□

B.1.1 MIXED-PHASE GRADIENT BOUND

In this section, we extend the two-phase SDE derivation to a single mixed setting. This is important because, at any point during training, some per-example gradients may exceed the clipping threshold while others remain below it. We show that in this scenario the same bound holds as in Theorem B.7, where it was previously derived under a worst-case approach.

Theorem B.8 *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth. Then, we can write the SDE of* DP−SGD *as*

$$dX_t = b_{\text{mix}}(X_t)\,dt + \sqrt{\eta}\,\Sigma_{\text{mix}}(X_t)^{1/2} dW_t, \tag{60}$$

where the drift and covariance satisfy, for all x,

$$\langle \nabla f(x), b_{\text{mix}}(x) \rangle \le -\frac{1}{K_1}\|\nabla f(x)\|^2, \tag{61}$$

$$\operatorname{Tr} \Sigma_{\text{mix}}(x) \le d\left(K_2 + \frac{C^2\sigma_{\text{DP}}^2}{B^2}\right), \tag{62}$$

where K_1, K_2 are defined in Equation 52. Therefore, for $\tilde{t} \sim \text{Unif}(0, \tau)$,

$$\mathbb{E}\|f(X_{\tilde{t}})\|_2^2 \le K_1\left(\frac{f(X_0)}{\eta T} + \frac{\eta dL}{2}\left(K_2 + \frac{C^2\sigma_{\text{DP}}^2}{B^2}\right)\right), \tag{63}$$

i.e., the same L-smooth convergence bound as in Theorem B.7 holds for any mixture *of clipped and unclipped samples in each mini-batch.*

Proof: We proceed in three steps: i) drift under mixed clipping, ii) covariance under mixed clipping using an explicit decomposition of G_k into clipped and unclipped parts, and iii) Itô's formula and the final bound.

• Step 1: Drift of the mixed batch. The DP-SGD update can be written as

$$x_{k+1} = x_k - \eta\left(G_k + \frac{1}{B}Z_{\text{DP}}\right), \quad \text{where} \quad G_k := \frac{1}{B}\sum_{i\in\gamma_k} \mathcal{C}(\nabla f_i(x_k)). \tag{64}$$

Let

$$S_{1,k} := \{i \in \gamma_k : \|\nabla f_i(x_k)\| \ge C\}, \qquad S_{2,k} := \{i \in \gamma_k : \|\nabla f_i(x_k)\| < C\}, \tag{65}$$

$$B_k := |S_{1,k}|, \qquad p_k := \frac{B_k}{B} \in [0, 1]. \tag{66}$$

Intuitively, p_k represents the probability of a sample being in Phase 1. Define the per-sample contributions

$$Y_i := C(\nabla f_i(x_k)), \quad i \in S_{1,k}, \qquad X_i := \nabla f_i(x_k), \quad i \in S_{2,k}, \tag{67}$$

and the corresponding batch averages

$$g_k^{(1)} := \frac{1}{B}\sum_{i\in S_{1,k}} Y_i, \qquad g_k^{(2)} := \frac{1}{B}\sum_{i\in S_{2,k}} X_i, \tag{68}$$

so that

$$G_k = g_k^{(1)} + g_k^{(2)}. \tag{69}$$

In the same way as in the proof of Theorem B.5, we have that

$$\mathbb{E}[Y_i] = a_1 \nabla f(x_k), \qquad a_1 := \frac{CK(\nu)}{\sigma_\gamma\sqrt{d}}, \tag{70}$$

and

$$\mathbb{E}[X_i] = \nabla f(x_k). \tag{71}$$

Conditioned on the sets $S_{1,k}, S_{2,k}$, the Y_i are i.i.d. over $S_{1,k}$ and the X_i are i.i.d. over $S_{2,k}$, so

$$\mathbb{E}[g_k^{(1)} \mid S_{1,k}] = \frac{1}{B}\sum_{i\in S_{1,k}} \mathbb{E}[Y_i] = \frac{B_k}{B}\mathbb{E}[Y_i] = p_k a_i \nabla f(x_k), \tag{72}$$

$$\mathbb{E}[g_k^{(2)} \mid S_{2,k}] = \frac{1}{B} \sum_{i \in S_{2,k}} \mathbb{E}[X_i] = \frac{B - B_k}{B} \mathbb{E}[X_i] = (1 - p_k) \nabla f(x_k). \tag{73}$$

Thus

$$\mathbb{E}[G_k \mid S_{1,k}, S_{2,k}] = p_k a_1 \nabla f(x_k) + (1 - p_k) \nabla f(x_k). \tag{74}$$

Recall the definition of K_1 from Equation 52

$$K_1 := \max \left\{ 1, \ \frac{\sigma_\gamma \sqrt{d}}{CK(\nu)} \right\}. \tag{75}$$

Then, it holds $a_1 = CK(\nu)/(\sigma_\gamma \sqrt{d}) \geq 1/K_1$ and $a_2 := 1 \geq 1/K_1$. Since $a_{\text{mix}}(p_k)$ is a convex combination of a_1 and a_2,

$$a_{\text{mix}}(p_k) = p_k a_1 + (1 - p_k) a_2 \ \geq \ \min\{a_1, a_2\} \ \geq \ \frac{1}{K_1} \qquad \forall\, p_k \in [0, 1]. \tag{76}$$

Therefore, the drift in the SDE limit satisfies

$$b_{\text{mix}}(x) = -a_{\text{mix}}(p(x))\, \nabla f(x), \qquad \langle \nabla f(x), b_{\text{mix}}(x) \rangle \leq -\frac{1}{K_1} \|\nabla f(x)\|^2. \tag{77}$$

• Step 2: Covariance of the mixed batch. We now compute the gradient noise covariance and show it is a convex combination of the pure-phase covariances, which are already derived in Theorem B.5. Define the centered contributions

$$U_i := Y_i - \mathbb{E}[Y_i], \quad i \in S_{1,k}, \qquad V_j := X_j - \mathbb{E}[X_j], \quad j \in S_{2,k}. \tag{78}$$

Then

$$g_k^{(1)} - \mathbb{E}[g_k^{(1)} \mid S_{1,k}] = \frac{1}{B} \sum_{i \in S_{1,k}} U_i, \qquad g_k^{(2)} - \mathbb{E}[g_k^{(2)} \mid S_{2,k}] = \frac{1}{B} \sum_{j \in S_{2,k}} V_j. \tag{79}$$

Hence

$$G_k - \mathbb{E}[G_k \mid S_{1,k}, S_{2,k}] = \frac{1}{B} \sum_{i \in S_{1,k}} U_i + \frac{1}{B} \sum_{j \in S_{2,k}} V_j. \tag{80}$$

Let

$$\Sigma_1^{\text{single}}(x_k) := \text{Cov}(Y_i) = \text{Cov}(U_i), \qquad \Sigma_2^{\text{single}}(x_k) := \text{Cov}(X_j) = \text{Cov}(V_j). \tag{81}$$

be the covariances of a single data-point. Conditioned on the sets $S_{1,k}, S_{2,k}$, the random vectors $\{U_i : i \in S_{1,k}\}$ and $\{V_j : j \in S_{2,k}\}$ are independent and zero-mean. Thus

$$\text{Cov}\big(G_k \mid S_{1,k}, S_{2,k}\big) = \text{Cov}\left(\frac{1}{B} \sum_{i \in S_{1,k}} U_i + \frac{1}{B} \sum_{j \in S_{2,k}} V_j \,\Big|\, S_{1,k}, S_{2,k} \right) \tag{82}$$

$$= \frac{1}{B^2} \text{Cov}\Big(\sum_{i \in S_{1,k}} U_i \Big) + \frac{1}{B^2} \text{Cov}\Big(\sum_{j \in S_{2,k}} V_j \Big), \tag{83}$$

where cross terms vanish by independence. Using i.i.d. within each group, we have

$$\text{Cov}\Big(\sum_{i \in S_{1,k}} U_i \Big) = B_k\, \Sigma_1^{\text{single}}(x_k), \qquad \text{Cov}\Big(\sum_{j \in S_{2,k}} V_j \Big) = (B - B_k)\, \Sigma_2^{\text{single}}(x_k), \tag{84}$$

therefore

$$\Sigma_{\text{grad}}(x_k; S_{1,k}, S_{2,k}) := \text{Cov}(G_k \mid x_k, S_{1,k}, S_{2,k}) = \frac{B_k}{B^2} \Sigma_1^{\text{single}}(x_k) + \frac{B - B_k}{B^2} \Sigma_2^{\text{single}}(x_k). \tag{85}$$

Since $p_k = B_k/B$ and $1 - p_k = (B - B_k)/B$, we obtain

$$\Sigma_{\text{grad}}(x_k; S_{1,k}, S_{2,k}) = \frac{p_k}{B}\, \Sigma_1^{\text{single}}(x_k) + \frac{1 - p_k}{B}\, \Sigma_2^{\text{single}}(x_k). \tag{86}$$

In the pure-phase SDEs of Theorem B.5, the *batch-level* (gradient and DP) covariances are given by

$$\Sigma_1(\bar{x}) = \frac{1}{B} \Sigma_1^{\text{single}}(x) + \frac{C^2 \sigma_{\text{DP}}^2}{B^2} I_d, \qquad \Sigma_2(\bar{x}) = \frac{1}{B} \Sigma_2^{\text{single}}(x) + \frac{C^2 \sigma_{\text{DP}}^2}{B^2} I_d. \tag{87}$$

From equation 86, the *gradient* part of the mixed-phase covariance is

$$\Sigma_{\text{grad}}(x_k; S_{1,k}, S_{2,k}) = p_k \left(\Sigma_1(x_k) - \frac{C^2\sigma_{\text{DP}}^2}{B^2} I_d \right) + (1 - p_k) \left(\Sigma_2(x_k) - \frac{C^2\sigma_{\text{DP}}^2}{B^2} I_d \right). \quad (88)$$

Adding the DP noise term $\frac{C^2\sigma_{\text{DP}}^2}{B^2} I_d$ back in, the *full* covariance of the DP-SGD increment in the mixed batch is

$$\Sigma_{\text{mix}}(x_k; S_{1,k}, S_{2,k}) = \Sigma_{\text{grad}}(x_k; S_{1,k}, S_{2,k}) + \frac{C^2\sigma_{\text{DP}}^2}{B^2} I_d \quad (89)$$

$$= p_k \, \Sigma_1(x_k) + (1 - p_k) \, \Sigma_2(x_k). \quad (90)$$

Thus, at the SDE level, the mixed-phase covariance is exactly a convex combination of the pure-phase covariances Σ_1 and Σ_2. From Theorem B.6, we have the trace bounds

$$\operatorname{Tr} \Sigma_1(\bar{x}) \;\le\; C^2 + d \, \frac{C^2\sigma_{\text{DP}}^2}{B^2}, \quad (91)$$

$$\operatorname{Tr} \Sigma_2(\bar{x}) \;=\; d \left(\frac{\sigma_\gamma^2}{B} + \frac{C^2\sigma_{\text{DP}}^2}{B^2} \right). \quad (92)$$

Let K_2 as in Equation 52 we can write, for $r \in \{1, 2\}$,

$$\operatorname{Tr} \Sigma_r(\bar{x}) \;\le\; d \left(K_2 + \frac{C^2\sigma_{\text{DP}}^2}{B^2} \right). \quad (93)$$

Using equation 90, for any $p_k \in [0, 1]$,

$$\operatorname{Tr} \Sigma_{\text{mix}}(x_k) = p_k \, \operatorname{Tr} \Sigma_1(x_k) + (1 - p_k) \, \operatorname{Tr} \Sigma_2(x_k) \le d \left(K_2 + \frac{C^2\sigma_{\text{DP}}^2}{B^2} \right). \quad (94)$$

Hence, the mixed-phase covariance satisfies exactly the same worst-case trace bound as the pure-phase covariances.

• Step 3: Itô bound and convergence. Finally, we can rewrite the SDE of `DP-SGD` as follows:

$$dX_t = b_{\text{mix}}(X_t) \, dt + \sqrt{\eta} \, \Sigma_{\text{mix}}(X_t)^{1/2} dW_t, \quad (95)$$

where, for all x,

$$\langle \nabla f(x), b_{\text{mix}}(x) \rangle \le -\frac{1}{K_1} \, \|\nabla f(x)\|^2, \quad (96)$$

$$\operatorname{Tr} \Sigma_{\text{mix}}(x) \le d \left(K_2 + \frac{C^2\sigma_{\text{DP}}^2}{B^2} \right). \quad (97)$$

Since f is L-smooth, $\nabla^2 f(x) \preceq L I_d$. By Itô's formula,

$$df(X_t) = \langle \nabla f(X_t), b_{\text{mix}}(X_t) \rangle \, dt + \frac{\eta}{2} \operatorname{Tr}\big(\nabla^2 f(X_t) \Sigma_{\text{mix}}(X_t) \big) \, dt + \mathcal{O}(\text{Noise}). \quad (98)$$

Using the drift and covariance bounds,

$$df(X_t) \le -\frac{1}{K_1} \, \|\nabla f(X_t)\|_2^2 \, dt + \frac{\eta dL}{2} \left(K_2 + \frac{C^2\sigma_{\text{DP}}^2}{B^2} \right) dt + \mathcal{O}(\text{Noise}). \quad (99)$$

Integrating from 0 to $\tau := \eta T$,

$$f(X_\tau) - f(X_0) \le -\frac{1}{K_1} \int_0^\tau \|\nabla f(X_t)\|_2^2 \, dt + \frac{\eta dL}{2} \left(K_2 + \frac{C^2\sigma_{\text{DP}}^2}{B^2} \right) \tau + \mathcal{O}(\text{Noise}). \quad (100)$$

Rearranging,

$$\frac{1}{K_1} \int_0^\tau \|\nabla f(X_t)\|_2^2 \, dt \le f(X_0) - f(X_\tau) + \frac{\eta dL}{2} \left(K_2 + \frac{C^2\sigma_{\text{DP}}^2}{B^2} \right) \tau + \mathcal{O}(\text{Noise}). \quad (101)$$

Taking expectations,

$$\frac{1}{K_1} \, \mathbb{E} \int_0^\tau \|\nabla f(X_t)\|_2^2 \, dt \le \mathbb{E}[f(X_0) - f(X_\tau)] + \frac{\eta dL}{2} \left(K_2 + \frac{C^2\sigma_{\text{DP}}^2}{B^2} \right) \tau. \quad (102)$$

Let $\tilde{t} \sim \mathrm{Unif}(0, \tau)$. Then, by the Law on Unconscious Statistician,

$$\mathbb{E}\|\nabla f(X_{\tilde{t}})\|^2 = \frac{1}{\tau}\,\mathbb{E}\int_0^\tau \|\nabla f(X_t)\|_2^2\,dt \le K_1\left(\frac{f(X_0)}{\tau} + \frac{\eta dL}{2}\left(K_2 + \frac{C^2\sigma_{\mathrm{DP}}^2}{B^2}\right)\right). \tag{103}$$

Since $\tau = \eta T$, this is exactly the gradient-norm bound as in Theorem B.7, with the same constants K_1, K_2, now rigorously shown to hold under arbitrary mixtures of clipped and unclipped samples at each iteration.

□

We now derive the stationary distribution of DP−SGD at convergence: We empirically validate this result in Figure C.3.

Theorem B.9 *Let $f(x) = \frac{1}{2}x^\top Hx$ where $H = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$. The stationary distribution at convergence of DP−SGD is*

$$(\mathbb{E}[X_\tau], \mathrm{Cov}(X_\tau)) = \left(X_0 e^{-H\tau}, \frac{T\eta}{2\varepsilon^2}\left(\frac{\varepsilon^2\sigma_\gamma^2}{BT} + \frac{C^2q^2\log(1/\delta)}{B^2}\right)(1 - e^{-2H\tau})H^{-1}\right). \tag{104}$$

Proof: Since H is diagonal, we can isolate each component. Furthermore, since $f(\cdot)$ is quadratic we can rewrite the SDE as:

$$dX_{t,i} = -\lambda_i X_{t,i} + \sqrt{\eta}\sqrt{\frac{\sigma_\gamma^2}{B} + \frac{C^2\sigma_{DP}^2}{B^2}}dW_{t,i}. \tag{105}$$

We have immediately that

$$\mathbb{E}[X_{t,i}] = X_{0,i}e^{-\lambda_i t}. \tag{106}$$

Applying the Itô isometry, we obtain:

$$\begin{aligned}
&\mathbb{E}[(X_{t,i} - \mathbb{E}[X_{t,i}])^2] \\
&= \eta\mathbb{E}\left[\int_0^t \left(e^{-\lambda_i(t-s)}\sqrt{\frac{\sigma_\gamma^2}{B} + \frac{C^2\sigma_{DP}^2}{B^2}}dW_s\right)^\top \left(e^{-\lambda_i(t-s)}\sqrt{\frac{\sigma_\gamma^2}{B} + \frac{C^2\sigma_{DP}^2}{B^2}}dW_s\right)\right] \\
&= \eta\left(\frac{\sigma_\gamma^2}{B} + \frac{C^2\sigma_{DP}^2}{B^2}\right)\int_0^t e^{-2\lambda_i(t-s)}ds \\
&= \frac{\eta}{2\lambda_i}\left(\frac{\sigma_\gamma^2}{B} + \frac{C^2q^2T\log(1/\delta)}{B^2\varepsilon^2}\right)(1 - e^{-2\lambda_i t}) \\
&= \frac{T\eta}{2\varepsilon^2\lambda_i}\left(\frac{\varepsilon^2\sigma_\gamma^2}{BT} + \frac{C^2q^2\log(1/\delta)}{B^2}\right)(1 - e^{-2\lambda_i t}).
\end{aligned}$$

□

B.2 DP−SignSGD

This subsection provides the formal derivation of the SDE model for DP−SignSGD and formal statements of Theorem 4.3, Theorem 4.4, and Theorem B.15. Similarly to DP−SGD, the dynamics of the method shifts again between two phases; we first model and study each phase separately.

Theorem B.10 *Let $0 < \eta < 1$, $\tau > 0$ and set $T = \lfloor \tau/\eta \rfloor$ and $K(\nu) = \sqrt{\frac{2}{\nu}}\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}$, $\nu \ge 1$. Let $x_k \in \mathbb{R}^d$ denote a sequence of DP−SignSGD iterations defined in Eq. 5. Assume Assumption B.1 and Assumption B.2. Let X_t be the solution of the following SDEs with initial condition $X_0 = x_0$:*

• *Phase 1:*

$$dX_t = -\mathbb{E}_\gamma\left[\mathrm{Erf}\left(\frac{B}{\sigma_{DP}\sqrt{2}}\frac{\nabla f_\gamma(X_t)}{\|\nabla f_\gamma(X_t)\|_2}\right)\right]dt + \sqrt{\eta}\sqrt{\bar{\Sigma}(X_t)}dW_t, \tag{107}$$

where $\bar{\Sigma}(x) = I_d - \mathbb{E}_\gamma\left[\mathrm{Erf}\left(\frac{B}{\sigma_{DP}\sqrt{2}}\frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|_2}\right)\right]^2$.

• *Phase 2:*

$$dX_t = -\,\mathrm{Erf}\left(\frac{\nabla f(X_t)}{\sqrt{2\left(\frac{C^2\sigma_{DP}^2}{B^2}+\frac{\sigma_\gamma^2}{B}\right)}}\right)dt + \sqrt{\eta}\sqrt{\bar{\Sigma}(X_t)}dW_t, \tag{108}$$

where $\bar{\Sigma}(x) = I_d - \mathrm{Erf}\left(\frac{\nabla f(x)}{\sqrt{2\left(\frac{C^2\sigma_{DP}^2}{B^2}+\frac{\sigma_\gamma^2}{B}\right)}}\right)^2$ *and* $\mathrm{Erf}(\cdot)$ *is applied component-wise.*

Then, Eq. 107 and Eq. 108 are an order 1 *approximation of the discrete update of Phase 1 and Phase 2 of* `DP-SignSGD`*, respectively.*

Proof: The proof is virtually identical to that of Theorem B.5. Hence, we highlight only the necessary details for each phase. Let $\bar{\Delta} = x_1 - x$ be the one-step increment.

• Phase 1: We begin by computing the first moment:

$$\mathbb{E}[\bar{\Delta}] = -\eta\mathbb{E}_{\gamma,DP}\left[\mathrm{sign}\left(C\frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|_2} + Z_{DP}\right)\right]. \tag{109}$$

Remember that, for any random variable Y, we have

$$\mathbb{E}[\mathrm{sign}(Y)] = 1 - 2\mathbb{P}(Y < 0), \tag{110}$$

and that if furthermore $Y \sim \mathcal{N}(0,1)$, then

$$\Phi(y) = \frac{1}{2}\left(1 + \mathrm{Erf}\left(\frac{y}{\sqrt{2}}\right)\right). \tag{111}$$

Since $Z_{DP} \sim \mathcal{N}\left(0, \frac{C^2\sigma_{DP}^2}{B^2}\right)$, we have that

$$1 - 2\mathbb{P}\left(C\frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|_2} + Z_{DP} < 0\right) = 1 - 2\Phi\left(-\frac{B}{C\sigma_{DP}}C\frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|_2}\right) \tag{112}$$

$$= 1 - \left(1 + \mathrm{Erf}\left(-\frac{B}{\sigma_{DP}\sqrt{2}}\frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|_2}\right)\right) \tag{113}$$

$$= \mathrm{Erf}\left(\frac{B}{\sigma_{DP}\sqrt{2}}\frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|_2}\right). \tag{114}$$

Thus

$$\mathbb{E}[\bar{\Delta}] = -\eta\mathbb{E}_\gamma\left[\mathrm{Erf}\left(\frac{B}{\sigma_{DP}\sqrt{2}}\frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|_2}\right)\right]. \tag{115}$$

The second moment is instead

$$\mathrm{Cov}(\bar{\Delta})_{ij} = \eta^2\mathbb{E}_{\gamma,DP}\left[\left(\mathrm{sign}\left(C\frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|_2} + Z_{DP}\right) - \mathbb{E}_\gamma\left[\mathrm{Erf}\left(\frac{B}{\sigma_{DP}\sqrt{2}}\frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|_2}\right)\right]\right)_i\right.$$
$$\left.\left(\mathrm{sign}\left(C\frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|_2} + Z_{DP}\right) - \mathbb{E}_\gamma\left[\mathrm{Erf}\left(\frac{B}{\sigma_{DP}\sqrt{2}}\frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|_2}\right)\right]\right)_j\right] \tag{116}$$

$$= \eta^2\mathbb{E}_{\gamma,DP}\left[\mathrm{sign}\left(C\frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|_2} + Z_{DP}\right)_i \mathrm{sign}\left(C\frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|_2} + Z_{DP}\right)_j\right] \tag{117}$$

$$- \eta^2\mathbb{E}_\gamma\left[\mathrm{Erf}\left(\frac{B}{\sigma_{DP}\sqrt{2}}\frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|_2}\right)_i\right]\mathbb{E}_\gamma\left[\mathrm{Erf}\left(\frac{B}{\sigma_{DP}\sqrt{2}}\frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|_2}\right)_j\right]. \tag{118}$$

If $i = j$, we have

$$\bar{\Delta}_{ii} = \eta^2 - \eta^2 \mathbb{E}_\gamma \left[\text{Erf}\left(\frac{B}{\sigma_{DP}\sqrt{2}} \frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|_2} \right) \right]^2. \tag{119}$$

Otherwise, we have

$$\begin{aligned}
&\mathbb{E}_{\gamma,DP}\left[\text{sign}\left(C\frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|_2} + Z_{DP} \right)_i \text{sign}\left(C\frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|_2} + Z_{DP} \right)_j \right] && (120)\\
&= \mathbb{E}_\gamma\left[\mathbb{E}_{DP}\left[\text{sign}\left(C\frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|_2} + Z_{DP} \right)_i \right] \mathbb{E}_{DP}\left[\text{sign}\left(C\frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|_2} + Z_{DP} \right)_j \right] \right] \\
&= \mathbb{E}_\gamma\left[\text{Erf}\left(\frac{B}{\sigma_{DP}\sqrt{2}} \frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|_2} \right)_i \text{Erf}\left(\frac{B}{\sigma_{DP}\sqrt{2}} \frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|_2} \right)_j \right] && (121)\\
&= \mathbb{E}_\gamma\left[\text{Erf}\left(\frac{B}{\sigma_{DP}\sqrt{2}} \frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|_2} \right)_i \right] \mathbb{E}_\gamma\left[\text{Erf}\left(\frac{B}{\sigma_{DP}\sqrt{2}} \frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|_2} \right)_j \right]. && (122)
\end{aligned}$$

Where we used the independence of the i-th and j-th components. Hence

$$\text{Cov}(\bar{\Delta})_{ij} = 0. \tag{123}$$

Finally, we have

$$\text{Cov}(\bar{\Delta}) = \eta^2 I_d - \eta^2 \mathbb{E}_\gamma \left[\text{Erf}\left(\frac{B}{\sigma_{DP}\sqrt{2}} \frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|_2} \right) \right]^2. \tag{124}$$

Define now

$$b(x) = -\mathbb{E}_\gamma \left[\text{Erf}\left(\frac{B}{\sigma_{DP}\sqrt{2}} \frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|_2} \right) \right] \tag{125}$$

$$\bar{\Sigma}(x) = I_d - \mathbb{E}_\gamma \left[\text{Erf}\left(\frac{B}{\sigma_{DP}\sqrt{2}} \frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|_2} \right) \right]^2. \tag{126}$$

Then, from Lem B.3 and Thm. B.4 the claim follows.

• Phase 2: Remember that, from Assumption B.2, $\nabla f_\gamma = \nabla f + Z_\gamma$, where $Z_\gamma \sim \mathcal{N}\left(0, \frac{\sigma_\gamma^2}{B}\right)$. We calculate the expected increment

$$\begin{aligned}
\mathbb{E}[\bar{\Delta}] &= -\eta \mathbb{E}\left[\text{sign}(\nabla f_\gamma(x) + Z_{DP}) \right] && (127)\\
&= -\eta \mathbb{E}[\text{sign}(\nabla f(x) + Z_\gamma + Z_{DP}] && (128)\\
&= -\eta \,\text{Erf}\left(\frac{\nabla f(x)}{\sqrt{2\left(\frac{C^2\sigma_{DP}^2}{B^2} + \frac{\sigma_\gamma^2}{B} \right)}} \right). && (129)
\end{aligned}$$

Instead, the covariance becomes

$$\mathrm{Cov}(\bar{\Delta})_{ij} = \eta^2 \mathbb{E}_{\gamma,DP}\left[\left(\mathrm{sign}(\nabla f_\gamma + Z_{DP}) - \mathrm{Erf}\left(\frac{\nabla f(x)}{\sqrt{2\left(\frac{C^2\sigma_{DP}^2}{B^2}+\frac{\sigma_\gamma^2}{B}\right)}}\right)\right)_i \tag{130}$$

$$\left(\mathrm{sign}(\nabla f_\gamma + Z_{DP}) - \mathrm{Erf}\left(\frac{\nabla f(x)}{\sqrt{2\left(\frac{C^2\sigma_{DP}^2}{B^2}+\frac{\sigma_\gamma^2}{B}\right)}}\right)\right)_j\Bigg] \tag{131}$$

$$= \eta^2 \mathbb{E}_{\gamma,DP}\left[\mathrm{sign}(\nabla f_\gamma + Z_{DP})_i \,\mathrm{sign}(\nabla f_\gamma + Z_{DP})_j\right] \tag{132}$$

$$- \eta^2 \,\mathrm{Erf}\left(\frac{\partial_i f(x)}{\sqrt{2\left(\frac{C^2\sigma_{DP}^2}{B^2}+\frac{\sigma_\gamma^2}{B}\right)}}\right)\mathrm{Erf}\left(\frac{\partial_j f(x)}{\sqrt{2\left(\frac{C^2\sigma_{DP}^2}{B^2}+\frac{\sigma_\gamma^2}{B}\right)}}\right). \tag{133}$$

If $i = j$, we have

$$\mathrm{Cov}(\bar{\Delta})_{ii} = \eta^2\left(1 - \mathrm{Erf}\left(\frac{\partial_i f(x)}{\sqrt{2\left(\frac{C^2\sigma_{DP}^2}{B^2}+\frac{\sigma_\gamma^2}{B}\right)}}\right)^2\right); \tag{134}$$

while if $i \neq j$

$$\mathrm{Cov}(\bar{\Delta})_{ij} = \eta^2 \,\mathrm{Erf}\left(\frac{\partial_i f(x)}{\sqrt{2\left(\frac{C^2\sigma_{DP}^2}{B^2}+\frac{\sigma_\gamma^2}{B}\right)}}\right)\mathrm{Erf}\left(\frac{\partial_j f(x)}{\sqrt{2\left(\frac{C^2\sigma_{DP}^2}{B^2}+\frac{\sigma_\gamma^2}{B}\right)}}\right) \tag{135}$$

$$- \eta^2 \,\mathrm{Erf}\left(\frac{\partial_i f(x)}{\sqrt{2\left(\frac{C^2\sigma_{DP}^2}{B^2}+\frac{\sigma_\gamma^2}{B}\right)}}\right)\mathrm{Erf}\left(\frac{\partial_j f(x)}{\sqrt{2\left(\frac{C^2\sigma_{DP}^2}{B^2}+\frac{\sigma_\gamma^2}{B}\right)}}\right) = 0, \tag{136}$$

Define now

$$b(x) = -\,\mathrm{Erf}\left(\frac{\nabla f(x)}{\sqrt{2\left(\frac{C^2\sigma_{DP}^2}{B^2}+\frac{\sigma_\gamma^2}{B}\right)}}\right) \tag{137}$$

$$\bar{\Sigma}(x) = I_d - \mathrm{Erf}\left(\frac{\nabla f(x)}{\sqrt{2\left(\frac{C^2\sigma_{DP}^2}{B^2}+\frac{\sigma_\gamma^2}{B}\right)}}\right)^2. \tag{138}$$

Then, from Lem B.3 and Thm. B.4 the claim follows.

□

Corollary B.11 *Under our assumptions, the SDEs (Eq. 107 and Eq. 108) modelling the two phases of* `DP-SignSGD` *as follows:*

• *Phase 1:*

$$dX_t = -\sqrt{\frac{2}{d\pi}}\frac{BK(\nu)}{\sigma_{DP}\sigma_\gamma}\nabla f(X_t)dt + \sqrt{\eta}\sqrt{I_d - \frac{2}{d\pi}\frac{B^2K(\nu)^2}{\sigma_{DP}^2\sigma_\gamma^2}\,\mathrm{diag}(\nabla f(X_t))^2}dW_t; \tag{139}$$

• *Phase 2:*

$$dX_t = -\sqrt{\frac{2}{\pi}}\frac{1}{\sqrt{\frac{C^2\sigma_{DP}^2}{B^2}+\frac{\sigma_\gamma^2}{B}}}\nabla f(X_t) + \eta\sqrt{I_d - \frac{2}{\pi}\frac{1}{\frac{C^2\sigma_{DP}^2}{B^2}+\frac{\sigma_\gamma^2}{B}}\operatorname{diag}(\nabla f(X_t))^2}dW_t. \quad (140)$$

Proof: Let us w.l.o.g. assume that that $\left|\frac{\partial_i f_\gamma(x)}{\|\nabla f_\gamma(x)\|_2}\right| \ll 1$ when $\|\nabla f_\gamma(x)\|_2 \geq C$: This is not restrictive when the number of trainable parameters d is large as it is under our assumptions. Additionally, we recall that under our assumptions, $\frac{|\partial_i f(x)|}{\sqrt{2(\sigma_{DP}^2+\sigma_\gamma^2/B)}} \ll 1$. Then one can write:

• Phase 1: Since $\left|\frac{\partial_i f_\gamma(x)}{\|\nabla f_\gamma(x)\|_2}\right| \ll 1$, one can approximate the error function in a neighborhood of 0 as follows: $\operatorname{Erf}(x) \sim \frac{2}{\sqrt{\pi}}x$. Thanks to Lemma A.1, we have

$$\mathbb{E}_\gamma\left[\operatorname{Erf}\left(\frac{B}{\sigma_{DP}\sqrt{2}}\frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|_2}\right)\right] = \mathbb{E}_\gamma\left[\sqrt{\frac{2}{\pi}}\frac{B}{\sigma_{DP}}\frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|_2}\right] = \sqrt{\frac{2}{d\pi}}\frac{BK(\nu)}{\sigma_{DP}\sigma_\gamma}\nabla f(x) \quad (141)$$

Therefore, Eq. 107 becomes

$$dX_t = -\sqrt{\frac{2}{d\pi}}\frac{BK(\nu)}{\sigma_{DP}\sigma_\gamma}\nabla f(X_t)dt + \sqrt{\eta}\sqrt{I_d - \frac{2}{d\pi}\frac{B^2K(\nu)^2}{\sigma_{DP}^2\sigma_\gamma^2}\operatorname{diag}(\nabla f(X_t))^2}dW_t. \quad (142)$$

• Phase 2: Since $\left|\frac{\partial_i f(x)}{\sqrt{2(\sigma_{DP}^2+\sigma_\gamma^2/B)}}\right| \ll 1$ for $i = 1, \ldots, d$, one can use the same argument as before to use a linear approximation of the error function. In detail, one has

$$\operatorname{Erf}\left(\frac{\nabla f(x)}{\sqrt{2\left(\frac{C^2\sigma_{DP}^2}{B^2}+\frac{\sigma_\gamma^2}{B}\right)}}\right) = \frac{2}{\sqrt{\pi}}\frac{\nabla f(x)}{\sqrt{2\left(\frac{C^2\sigma_{DP}^2}{B^2}+\frac{\sigma_\gamma^2}{B}\right)}} = \sqrt{\frac{2}{\pi}}\frac{1}{\sqrt{\frac{C^2\sigma_{DP}^2}{B^2}+\frac{\sigma_\gamma^2}{B}}}\nabla f(x). \quad (143)$$

Therefore, Eq. 108 becomes

$$dX_t = -\sqrt{\frac{2}{\pi}}\frac{1}{\sqrt{\frac{C^2\sigma_{DP}^2}{B^2}+\frac{\sigma_\gamma^2}{B}}}\nabla f(X_t) + \eta\sqrt{I_d - \frac{2}{\pi}\frac{1}{\frac{C^2\sigma_{DP}^2}{B^2}+\frac{\sigma_\gamma^2}{B}}\operatorname{diag}(\nabla f(X_t))^2}dW_t. \quad (144)$$

□

Theorem B.12 *Let f be L-smooth and μ-PL. Then, for $t \in [0, \tau]$, we have that*

• *Phase 1, i.e., when the gradient is clipped, the loss satisfies:*

$$\mathbb{E}[f(X_t)] \lesssim f(X_0)e^{\frac{-\mu B}{\sigma_\gamma\sqrt{dT}}\frac{\varepsilon}{\Phi}t} + \left(1 - e^{\frac{-\mu B}{\sigma_\gamma\sqrt{dT}}\frac{\varepsilon}{\Phi}t}\right)\frac{\sqrt{T}\eta L d^{\frac{3}{2}}\sigma_\gamma}{\mu B}\frac{\Phi}{\varepsilon}; \quad (145)$$

• *Phase 2, i.e., when the gradient is not clipped, the loss satisfies:*

$$\mathbb{E}[f(X_t)] \lesssim f(X_0)e^{\sqrt{\frac{-\mu\varepsilon t}{\varepsilon^2\frac{\sigma_\gamma^2}{B}+\frac{C^2\Phi^2}{B^2}T}}} + \left(1 - e^{\sqrt{\frac{-\mu\varepsilon t}{\varepsilon^2\frac{\sigma_\gamma^2}{B}+\frac{C^2\Phi^2}{B^2}T}}}\right)\frac{\sqrt{T}\eta L d}{\mu}\sqrt{\frac{\varepsilon^2\sigma_\gamma^2}{BT}+\frac{C^2\Phi^2}{B^2}}\frac{1}{\varepsilon}. \quad (146)$$

Proof: First of all, observe that, in both phases, it holds that $\bar{\Sigma}(x) \preceq I_d$.

• Phase 1: Since f is μ-PL and L-smooth it follows that $2\mu f(x) \leq \|\nabla f(x)\|^2$ and $\nabla^2 f(x) \preceq LI_d$. Then, By applying the Itô formula, we have

$$df(X_t) \leq -\sqrt{\frac{2}{d\pi T}}\frac{K(\nu)}{\sigma_\gamma}\frac{B\varepsilon}{q\log(1/\delta)}\|\nabla f(X_t)\|_2^2 dt + \frac{\eta}{2}\operatorname{Tr}\left(\nabla^2 f(X_t)I_d\right)dt + \mathcal{O}(\text{Noise}) \quad (147)$$

$$\leq -2\mu\sqrt{\frac{2}{d\pi T}}\frac{K(\nu)}{\sigma_\gamma}\frac{B\varepsilon}{q\log(1/\delta)}f(X_t)dt + \frac{\eta dL}{2}dt + \mathcal{O}(\text{Noise}). \quad (148)$$

Therefore,

$$\mathbb{E}[f(X_t)] \leq f(X_0)e^{-2\mu\left(\sqrt{\frac{2}{d\pi T}}\frac{K(\nu)}{\sigma_\gamma}\frac{B\varepsilon}{q\log(1/\delta)}\right)t} \tag{149}$$

$$+\left(1-e^{-2\mu\left(\sqrt{\frac{2}{d\pi T}}\frac{K(\nu)}{\sigma_\gamma}\frac{B\varepsilon}{q\log(1/\delta)}\right)t}\right)\sqrt{\frac{\pi T}{2}}\frac{\eta d^{\frac{3}{2}}L\sigma_\gamma}{4\mu K(\nu)}\frac{q\log(1/\delta)}{B\varepsilon}. \tag{150}$$

• Phase 2: As for Phase 1, by applying the Itô formula one has

$$df(X_t) \leq -\sqrt{\frac{2}{\pi}}\frac{1}{\sqrt{\varepsilon^2 B^{-1}\sigma_\gamma^2 + B^{-2}C^2q^2\log(1/\delta)T}}\varepsilon\|\nabla f(X_t)\|_2^2 dt \tag{151}$$

$$+\frac{\eta}{2}\operatorname{Tr}\left(\nabla^2 f(X_t)I_d\right)dt + \mathcal{O}(\text{Noise}) \tag{152}$$

$$\leq -2\mu\sqrt{\frac{2}{\pi}}\frac{1}{\sqrt{\varepsilon^2 B^{-1}\sigma_\gamma^2 + B^{-2}C^2q^2\log(1/\delta)T}}\varepsilon f(X_t)dt + \frac{\eta dL}{2}dt + \mathcal{O}(\text{Noise}). \tag{153}$$

Therefore

$$\mathbb{E}[f(X_t)] \leq f(X_0)e^{-\sqrt{\frac{2}{\pi}}\frac{2\mu}{\sqrt{\varepsilon^2 B^{-1}\sigma_\gamma^2+B^{-2}C^2q^2\log(1/\delta)T}}\varepsilon t} \tag{154}$$

$$+\left(1-e^{-\sqrt{\frac{2}{\pi}}\frac{2\mu}{\sqrt{\varepsilon^2 B^{-1}\sigma_\gamma^2+B^{-2}C^2q^2\log(1/\delta)T}}\varepsilon t}\right)\sqrt{\frac{\pi T}{2}}\frac{\eta dL}{4\mu}\sqrt{\frac{\varepsilon^2\sigma_\gamma^2}{BT}+\frac{C^2q^2\log(1/\delta)}{B^2}}\frac{1}{\varepsilon}.$$

Finally, by suppressing problem-independent constants, such as $2, \pi, K(\nu)$, the thesis follows.

□

Theorem B.13 *Let f be an L-smooth function. Define*

$$K_3 = \max\left\{\sqrt{\frac{d\pi}{2}}\frac{\sigma_\gamma q\sqrt{\log(1/\delta)}}{BK(\nu)}, \sqrt{\frac{\pi}{2}}\sqrt{\frac{\varepsilon^2\sigma_\gamma^2}{BT}+\frac{C^2q^2\log(1/\delta)}{B^2}}\right\}. \tag{155}$$

Then

$$\mathbb{E}\left[\|\nabla f(X_{\tilde{t}})\|_2^2\right] \lesssim K_3\left(\frac{f(X_0)}{\eta\sqrt{T}}+\eta dL\sqrt{T}\right)\frac{1}{\varepsilon}, \tag{156}$$

where $\tilde{t} \sim \text{Unif}(0,\tau)$.

Proof: Since in both phases the diffusion coefficient $\bar{\Sigma}(x) \preceq I_d$, the drift is the only term worth comparing for a worst-case analysis. Let then K_3 as in Eq. 155. Applying the Itô formula to the worst-case SDE we have

$$df(X_t) \leq -\varepsilon(\sqrt{T}K_3)^{-1}\|\nabla f(X_t)\|_2^2 dt + \frac{\eta}{2}\operatorname{Tr}\left(\nabla^2 f(X_t)I_d\right)dt + \mathcal{O}(\text{Noise}) \tag{157}$$

$$\leq -\varepsilon(\sqrt{T}K_3)^{-1}\|\nabla f(X_t)\|_2^2 dt + \frac{\eta dL}{2}dt + \mathcal{O}(\text{Noise}). \tag{158}$$

Then, by integrating and taking the expectation

$$\mathbb{E}\int_0^\tau \|\nabla f(X_t)\|_2^2 dt \leq K_3\sqrt{T}\left(f(X_0)+\frac{\eta dL\tau}{2}\right)\varepsilon^{-1} \tag{159}$$

$$\Longrightarrow \mathbb{E}\int_0^\tau \frac{1}{\tau}\|\nabla f(X_t)\|_2^2 dt \leq \frac{K_3}{\eta\sqrt{T}}\left(f(X_0)+\frac{\eta dL\tau}{2}\right)\varepsilon^{-1} \tag{160}$$

$$\Longrightarrow \mathbb{E}\left[\|\nabla f(X_{\tilde{t}})\|_2^2\right] \leq K_3\left(\frac{f(X_0)}{\eta\sqrt{T}}+\frac{\eta dL\sqrt{T}}{2}\right)\frac{1}{\varepsilon} \tag{161}$$

where in the last step we used the Law of the Unconscious Statistician and $\tilde{t} \sim \text{Unif}(0,\tau)$. Finally, by suppressing problem-independent constants, we get the thesis. □

B.2.1 MIXED-PHASE GRADIENT BOUND

Analogously to Section B.1.1, we extend the two-phase SDE derivation to a single mixed setting. Recall that, at any point during training, some per-example gradients may lie above the clipping threshold while others remain below it. The next result shows that, even in this more realistic mixed regime, we recover the same upper bound on the gradient norm as in Theorem B.13.

Theorem B.14 *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth. Then, we can write the SDE of* `DP-SignSGD` *as*

$$dX_t = b_{\text{mix}}(X_t)\,dt + \sqrt{\eta}\,\Sigma_{\text{mix}}(X_t)^{1/2} dW_t, \tag{162}$$

where

$$b_{\text{mix}}(x) = -\mathbb{E}\left[\text{Erf}\left(\frac{B}{C\sigma_{DP}\sqrt{2}}G(x)\right)\right], \tag{163}$$

$$\bar{\Sigma}_{\text{mix}}(x) = I_d - \mathbb{E}\left[\text{Erf}\left(\frac{B}{C\sigma_{DP}\sqrt{2}}G(x)\right)\right]^2, \tag{164}$$

and

$$G(x) = \frac{1}{B}\sum_{i\in\gamma_k} \mathcal{C}(\nabla f_i(x)) \tag{165}$$

Define

$$K_4 = \max\left\{\sqrt{\frac{\pi d}{2}}\frac{\sigma_\gamma q\sqrt{\log(1/\delta)}}{BK(\nu)}, \sqrt{\frac{\pi}{2}}\frac{Cq\sqrt{\log(1/\delta)}}{B}\right\}. \tag{166}$$

Then

$$\mathbb{E}\left[\|\nabla f(X_{\tilde{t}})\|_2^2\right] \le K_4\left(\frac{f(X_0)}{\eta\sqrt{T}} + \frac{\eta dL\sqrt{T}}{2}\right)\frac{1}{\varepsilon}, \tag{167}$$

where $\tilde{t} \sim \text{Unif}(0,\tau)$.

Remark B.3 *By construction we have $K_4 \le K_3$, so Theorem B.14 provides a formally tighter upper bound than Theorem B.13. However, note that the first term in the definitions of K_3 and K_4 (Equations 155 and 166, respectively) scales as $\sqrt{d}$. Since d is assumed to be large, this term typically dominates the maximum in both constants. As a consequence, in the high-dimensional regime of interest we effectively have $K_3 = K_4$, and the improvement from the mixed-phase analysis is negligible in practice.*

Proof: We divide the proof into two steps: i) SDE derivation, ii) gradient bound.

• Step 1: SDE derivation. Using the same notation as in the proof of Theorem B.8, we write the update of `DP-SignSGD` as

$$x_{k+1} = x_k - \eta\left(G_k + \frac{1}{B}Z_{\text{DP}}\right), \tag{168}$$

where

$$G_k := p_k\frac{1}{B_k}\sum_{i=1}^{B_k}\mathcal{C}(\nabla f_i(x_k)) + (1-p_k)\frac{1}{B-B_k}\sum_{i=1}^{B-B_k}\nabla f_i(x_k). \tag{169}$$

Since $Z_{DP} \sim \mathcal{N}$ we have

$$\mathbb{E}[x_{k+1} - x_k] = -\eta\mathbb{E}\left[\text{sign}\left(G_k + \frac{1}{B}Z_{DP}\right)\right] \tag{170}$$

$$= -\eta\mathbb{E}\left[\text{Erf}\left(\frac{B}{C\sigma_{DP}\sqrt{2}}G_k\right)\right] \tag{171}$$

and

$$\mathrm{Cov}(x_{k+1}-x_k) =\eta^2\mathbb{E}\left[\left(\mathrm{sign}\left(G_k+\frac{1}{B}Z_{DP}\right)-\mathbb{E}\left[\mathrm{Erf}\left(\frac{B}{C\sigma_{DP}\sqrt{2}}G_k\right)\right]\right)\right. \tag{172}$$

$$\left.\left(\mathrm{sign}\left(G_k+\frac{1}{B}Z_{DP}\right)-\mathbb{E}\left[\mathrm{Erf}\left(\frac{B}{C\sigma_{DP}\sqrt{2}}G_k\right)\right]\right)^\top\right] \tag{173}$$

$$=\eta^2\left(I_d-\mathbb{E}\left[\mathrm{Erf}\left(\frac{B}{C\sigma_{DP}\sqrt{2}}G_k\right)\right]^2\right). \tag{174}$$

If we define

$$b_{\mathrm{mix}}(x)=-\mathbb{E}\left[\mathrm{Erf}\left(\frac{B}{C\sigma_{DP}\sqrt{2}}G(x)\right)\right], \tag{175}$$

$$\bar{\Sigma}_{\mathrm{mix}}(x)=I_d-\mathbb{E}\left[\mathrm{Erf}\left(\frac{B}{C\sigma_{DP}\sqrt{2}}G(x)\right)\right]^2, \tag{176}$$

Then, the following SDE is an order-1 approximation of the update step of `DP-SignSGD`

$$dX_t=b_{\mathrm{mix}}(X_t)+\sqrt{\eta}\sqrt{\bar{\Sigma}_{\mathrm{mix}}(X_t)}dW_t. \tag{177}$$

• Step 2: Gradient bound. As argued in the proof of Corollary B.11, we assume $\left|\frac{B}{C\sigma_{DP}\sqrt{2}}G(x)\right|\ll 1$ without loss of generality. Therefore, applying the Itô formula, we have

$$df_t\le-\sqrt{\frac{2}{\pi}}\frac{B}{C\sigma_{DP}}\nabla f(X_t)^\top\mathbb{E}[G_k]dt+\frac{\eta}{2}\mathrm{Tr}\left(\nabla^2 f(X_t)\bar{\Sigma}(X_t)\right)dt+\mathcal{O}(\text{Noise}) \tag{178}$$

$$\le-(p_ka_1+(1-p_k)a_2)\nabla f(X_t)dt+\frac{\eta dL}{2}dt+\mathcal{O}(\text{Noise}), \tag{179}$$

where $a_1=\sqrt{\frac{2}{\pi}}\frac{B}{\sigma_{DP}}\frac{K(\nu)}{\sigma_\gamma\sqrt{d}}$ and $a_2=\sqrt{\frac{2}{\pi}}\frac{B}{C\sigma_{DP}}$. Expand σ_{DP} and define

$$K_4=\max\left\{\sqrt{\frac{\pi d}{2}}\frac{\sigma_\gamma q\sqrt{\log(1/\delta)}}{BK(\nu)},\sqrt{\frac{\pi}{2}}\frac{Cq\sqrt{\log(1/\delta)}}{B}\right\}. \tag{180}$$

Then

$$\varepsilon\sqrt{T^{-1}}K_4^{-1}\le p_ka_1+(1-p_k)a_2,\qquad\forall p_k\in[0,1] \tag{181}$$

Then

$$df_t\le-\varepsilon\sqrt{T^{-1}}K_4^{-1}\nabla f(X_t)dt+\frac{\eta dL}{2}dt+\mathcal{O}(\text{Noise}). \tag{182}$$

Then, by taking the expectation and integrating over $[0,\tau]$

$$\mathbb{E}\int_0^\tau\|\nabla f(X_t)\|_2^2dt\le\varepsilon^{-1}\sqrt{T}K_4\left(f(X_0)+\frac{\eta dL\tau}{2}\right) \tag{183}$$

$$\Longrightarrow\mathbb{E}\int_0^\tau\frac{1}{\tau}\|\nabla f(X_t)\|_2^2dt\le\frac{K_4}{\eta\sqrt{T}}\left(f(X_0)+\frac{\eta dL\tau}{2}\right)\varepsilon^{-1} \tag{184}$$

$$\Longrightarrow\mathbb{E}\left[\|\nabla f(X_{\tilde{t}})\|_2^2\right]\le K_4\left(\frac{f(X_0)}{\eta\sqrt{T}}+\frac{\eta dL\sqrt{T}}{2}\right)\frac{1}{\varepsilon}, \tag{185}$$

where in the last step we used the Law of the Unconscious Statistician and $\tilde{t}\sim\mathrm{Unif}(0,\tau)$.

□

Finally, we derive the stationary distribution of `DP-SignSGD`: We empirically validate it in Fig. C.3.

Theorem B.15 *Let $f(x) = \frac{1}{2}x^\top Hx$ where $H = \mathrm{diag}(\lambda_1, \dots, \lambda_d)$. The stationary distribution of Phase 2 is*

$$\mathbb{E}[X_T] = X_0 e^{-KH\tau}; \tag{186}$$

$$\mathrm{Cov}(X_T) = X_0^2 e^{-2KH\tau}\left(e^{-\eta K^2 H\tau} - 1\right) \tag{187}$$

$$+ \eta\left(2KH + \eta H^2 K^2\right)^{-1}\left(1 - e^{-(2KH+\eta K^2 H^2)\tau}\right) \tag{188}$$

where $K = \sqrt{\frac{2}{\pi}}\frac{1}{\sqrt{\varepsilon^2 B^{-1}\sigma_\gamma^2 + B^{-2}C^2q^2\log(1/\delta)T}}\varepsilon$.

Proof: Since H is diagonal, we can work component-wise. Let us remember the SDE:

$$dX_{t,i} = -\sqrt{\frac{2}{\pi}}\frac{1}{\sqrt{\frac{C^2\sigma_{DP}^2}{B^2} + \frac{\sigma_\gamma^2}{B}}}\lambda_i X_{t,i} + \sqrt{\eta}\sqrt{1 - \frac{2}{\pi\left(\frac{C^2\sigma_{DP}^2}{B^2} + \frac{\sigma_\gamma^2}{B}\right)}\lambda_i^2 X_{t,i}^2}dW_t. \tag{189}$$

To ease the notation, we write $K = \sqrt{\frac{2}{\pi}}\frac{1}{\sqrt{\varepsilon^2 B^{-1}\sigma_\gamma^2 + B^{-2}C^2q^2\log(1/\delta)T}}\varepsilon$. Hence, we can write $X_{t,i}$ in closed form as

$$X_{t,i} = x_{0,i}e^{-K\lambda_i t} + \sqrt{\eta}\int_0^t e^{-K\lambda_i(t-s)}\sqrt{1 - K^2\lambda_i^2 X_{t,i}^2}dW_s. \tag{190}$$

Due to the properties of the stochastic integral, we immediately have

$$\mathbb{E}[X_{t,i}] = X_{0,i}e^{-\sqrt{\frac{2}{\pi}}\frac{1}{\sqrt{\varepsilon^2 B^{-1}\sigma_\gamma^2 + B^{-2}C^2q^2\log(1/\delta)T}}\varepsilon\lambda_i t}. \tag{191}$$

Using the Itô formula on $g(x) = x^2$, we have

$$d\left(X_{t,i}^2\right) = -2K\lambda_i X_{t,i}^2 dt + \frac{\eta}{2}2dt - \frac{\eta}{2}2K^2\lambda_i^2 X_{t,i}^2 dt + \mathcal{O}(\text{Noise}) \tag{192}$$

$$\Longrightarrow \mathbb{E}[X_{t,i}^2] = X_{0,i}^2 e^{-(2K\lambda_i + \eta K^2\lambda_i^2)t} + \frac{\eta}{2K\lambda_i + \eta\lambda_i^2 K^2}\left(1 - e^{-(2K\lambda_i+\eta K^2\lambda_i^2)t}\right), \tag{193}$$

therefore

$$\mathrm{Cov}(X_{t,i}) = \mathbb{E}[X_{t,i}^2] - \mathbb{E}[X_{t,i}]^2 \tag{194}$$

$$= X_{0,i}^2 e^{-(2K\lambda_i+\eta K^2\lambda_i^2)t} + \frac{\eta}{2K\lambda_i + \eta\lambda_i^2 K^2}\left(1 - e^{-(2K\lambda_i+\eta K^2\lambda_i^2)t}\right) - X_{0,i}^2 e^{-2K\lambda_i t}$$

$$= X_{0,i}^2 e^{-2K\lambda_i t}\left(e^{-\eta K^2\lambda_i^2 t} - 1\right) + \frac{\eta}{2K\lambda_i + \eta\lambda_i^2 K^2}\left(1 - e^{-(2K\lambda_i+\eta K^2\lambda_i^2)t}\right). \tag{195}$$

□

Finally, we present a result that allows us to determine which of `DP-SignSGD` and `DP-SignSGD` is more advantageous depending on the training setting.

Corollary B.16 *If $\frac{\sigma_\gamma^2}{B} \geq 1$, then `DP-SignSGD` always achieves a better privacy-utility trade-off than `DP-SGD`, though its convergence is slower. If $\frac{\sigma_\gamma^2}{B} < 1$, there exists a critical privacy level*

$$\varepsilon^\star = \sqrt{\frac{C^2TB}{n^2\left(B - \sigma_\gamma^2\right)}\log\left(\tfrac{1}{\delta}\right)}, \tag{196}$$

such that `DP-SignSGD` outperforms `DP-SGD` in utility whenever $\varepsilon < \varepsilon^\star$, but still converges more slowly than `DP-SGD`.

Proof: The Phase 2 asymptotic terms at $t = T$ are

$$A_{\text{SGD}} = \frac{T\eta dL}{\mu}\Big(\tfrac{\varepsilon^2\sigma_\gamma^2}{TB} + C^2\tfrac{\Phi^2}{B^2}\Big)\tfrac{1}{\varepsilon^2}, \qquad A_{\text{Sign}} = \frac{\sqrt{T}\eta dL}{\mu}\sqrt{\tfrac{\varepsilon^2\sigma_\gamma^2}{TB} + C^2\tfrac{\Phi^2}{B^2}}\ \tfrac{1}{\varepsilon}. \tag{197}$$

We compare $A_{\text{Sign}} < A_{\text{SGD}}$. Cancelling the common factor $\frac{\eta dL}{\mu}$ gives

$$\frac{\sqrt{T}}{\varepsilon}\sqrt{\tfrac{\varepsilon^2\sigma_\gamma^2}{TB} + C^2\tfrac{\Phi^2}{B^2}} \ < \ \frac{T}{\varepsilon^2}\Big(\tfrac{\varepsilon^2\sigma_\gamma^2}{TB} + C^2\tfrac{\Phi^2}{B^2}\Big). \tag{198}$$

Multiplying by ε^2 and dividing by the positive square root yields

$$\varepsilon\sqrt{T} \ < \ T\sqrt{\tfrac{\varepsilon^2\sigma_\gamma^2}{TB} + C^2\tfrac{\Phi^2}{B^2}}. \tag{199}$$

All quantities are non-negative, so squaring preserves the inequality:

$$\varepsilon^2 T \ < \ T^2\Big(\tfrac{\varepsilon^2\sigma_\gamma^2}{TB} + C^2\tfrac{\Phi^2}{B^2}\Big) \iff \Big(1 - \tfrac{\sigma_\gamma^2}{B}\Big)\varepsilon^2 \ < \ C^2\tfrac{\Phi^2}{B^2}T. \tag{200}$$

Using $\frac{\Phi}{B} = \frac{1}{n}\sqrt{\log(1/\delta)}$ gives

$$\Big(1 - \tfrac{\sigma_\gamma^2}{B}\Big)\varepsilon^2 \ < \ \frac{C^2}{n^2}T\log\Big(\tfrac{1}{\delta}\Big). \tag{201}$$

If $\frac{\sigma_\gamma^2}{B} \geq 1$, the left coefficient is non-positive and the inequality holds for all $\varepsilon > 0$. If $\frac{\sigma_\gamma^2}{B} < 1$, solving for ε yields

$$\varepsilon < \sqrt{\frac{C^2TB}{n^2\left(B - \sigma_\gamma^2\right)}\log\left(\tfrac{1}{\delta}\right)} = \ \varepsilon^\star, \tag{202}$$

which proves the claim. □

Interestingly, by keeping η and C depend on the optimizer, we get

$$\sqrt{T}\eta_{\text{sign}}\sqrt{\frac{\sigma_\gamma^2}{BT} + \frac{C_{\text{sign}}^2\Phi^2}{B^2\varepsilon^2}} \ < \ T\eta_{\text{sgd}}\left(\frac{\sigma_\gamma^2}{BT} + \frac{C_{\text{sgd}}^2\Phi^2}{B^2\varepsilon^2}\right). \tag{203}$$

We observe that if $\sigma_\gamma \to \infty$, DP-SignSGD is always better than DP-SGD, while if $\sigma_\gamma \to 0$, there is always a threshold $\varepsilon^\star$. Since the algebraic expressions are complex, we believe this is enough to show that our insight is much more general than the case derived here and presented in the main paper.

C Experimental Details and Additional Results

Our empirical analysis is based on the official GitHub repository https://github.com/kenziyuliu/DP2 released with the Google paper (Li et al., 2023). In particular we consider the two following classification problems:

IMDB (Maas et al., 2011) is a sentiment analysis dataset for movie reviews, posed as a binary classification task. It contains 25,000 training samples and 25,000 test samples, with each review represented using a vocabulary of 10,000 words. We train a logistic regression model with 10,001 parameters.

StackOverflow (Kaggle, 2022), (TensorFlow Federated, 2022) is a large-scale text dataset derived from Stack Overflow questions and answers. Following the setup in (TensorFlow Federated, 2022), we consider the task of predicting the tag(s) associated with a given sentence, but we restrict our experiments to the standard centralized training setting rather than the federated one. We randomly select 246,092 sentences for training and 61,719 for testing, each represented with 10,000 features. The task is cast as a 500-class classification problem, yielding a model with approximately 5 million parameters.

Optimizers. We train both classification problems using DP-SGD, DP-SignSGD and DP-Adam. For $k \geq 0$, learning rate η, variance σ_{DP}^2, and batches γ_k of size B modeled as i.i.d. uniform random variables taking values in $\{1, \ldots, n\}$. Let g_k be the private gradient, defined as

$$g_k \coloneqq \frac{1}{B} \sum_{i \in \gamma_k} \mathcal{C}(\nabla f_i(x_k)) + \frac{1}{B} \mathcal{N}(0, C^2 \sigma_{\text{DP}}^2 I_d) \tag{204}$$

and $\mathcal{C}[\cdot]$ be the clipping function

$$\mathcal{C}(x) = \begin{cases} C \frac{x}{\|x\|_2} & \text{if } \|x\|_2 \geq C \\ x & \text{if } \|x\|_2 < C \end{cases}. \tag{205}$$

The iterates of DP-SGD are defined as

$$x_{k+1} = x_k - \eta g_k, \tag{206}$$

while those of DP-SignSGD are defined as

$$x_{k+1} = x_k - \eta \operatorname{sign}[g_k], \tag{207}$$

where $\operatorname{sign}[\cdot]$ is applied component-wise. The update rule of DP-Adam is defined as follows:

$$\begin{aligned} m_{k+1} &= \beta_1 m_k + (1-\beta_1) g_k, \qquad \hat{m}_{k+1} = \frac{m_{k+1}}{1-\beta_1^{k+1}}, \\ v_{k+1} &= \beta_2 v_k + (1-\beta_2) g_k^2, \qquad \hat{v}_{k+1} = \frac{v_{k+1}}{1-\beta_2^{k+1}}, \\ x_{k+1} &= x_k - \eta \frac{\hat{m}_{k+1}}{\sqrt{\hat{v}_{k+1}} + \epsilon}, \end{aligned} \tag{208}$$

where g_k is the privatized stochastic gradient and is defined in Equation 204.

Hyper-parameters. Unless stated otherwise, we fix the following hyperparameters in our experiments: for IMDB and StackOverflow respectively, we train for $100, 50$ epochs with batch size $B = 64$. The choice of batch size follows the setting in (Li et al., 2023). We also aimed to avoid introducing unnecessary variability, keeping the focus on the direction suggested by our theoretical results. Finally, we set $\delta = 10^{-5}, 10^{-6}$, corresponding to the rule $\delta = 10^{-k}$, where k is the smallest integer such that $10^{-k} \leq 1/n$ for the training dataset size n.

Protocol A. we perform a grid search on *learning rate* $\eta = \{0.001, 0.01, 0.1, 1, 3, 5, 10\}$ and *clipping threshold* $C = \{0.1, 0.25, 0.5, 1, 5\}$ for DP-SGD, DP-SignSGD and DP-Adam on both datasets, using $\sigma_{DP} = 1$: this gives $\varepsilon = 2.712$ and $\varepsilon = 0.424$ for IMDB and StackOverflow respectively. We summarize the best set of hyperparameters for each method on both datasets in Table C.1.

Dataset	DP-SGD	DP-SignSGD	DP-Adam
IMDB	(5, 0.5)	(0.1, 0.5)	(0.1, 0.5)
StackOverflow	(3, 0.25)	(0.01, 0.5)	(0.01, 0.5)

Table C.1: Tuned hyperparameters for different methods across the two datasets. The values refer to (learning rate, clipping parameter); For DP-Adam we also used $\beta_1 = 0.9, \beta_2 = 0.999$ and adaptivity $\epsilon = 10^{-8}$ in both cases.

Protocol B. For each noise multiplier, we tune a new pair of learning rate and clipping parameter by performing a grid search. **IMDB**: For DP-SignSGD and DP-Adam, we consider the following learning rates $\eta = \{0.01, 0.05, 0.10, 0.15, 0.22, 0.27, 0.33, 0.38, 0.44, 0.50\}$ and clipping thresholds $C = \{0.05, 0.1, 0.25, 0.5\}$, while for DP-SGD we consider a different range of learning rates $\eta = \{0.5, 0.7, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0, 5.5, 6.0\}$ and $C = \{0.1, 0.25, 0.5\}$. This tuning is designed to identify the best hyperparameters across a broad range of privacy budgets $\varepsilon = \{0.01, 0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6, 1.8, 2.0\}$, which

correspond to the following noise multipliers: $\{271.23, 13.56, 6.78, 4.52, 3.39, 2.71, 2.26, 1.94, 1.70, 1.51, 1.36\}$. **StackOverflow**: For DP−Adam we consider the following learning rates $\{0.001, 0.003, 0.005, 0.01, 0.03, 0.05, 0.1, 0.5\}$, for DP−SignSGD we add $\{0.008, 0.015, 0.02, 0.03, 0.04\}$ to the list, while for DP−SGD we consider a different range of learning rates $\eta = \{0.1, 0.5, 1.0, 2.0, 2.5, 3.0, 3.5, 4.0, 5.0\}$. For the clipping thresholds we consider $C = \{0.05, 0.1, 0.25, 0.35, 0.5, 1.0\}$ for every method. This tuning is designed to identify the best hyperparameters across a broad range of privacy budgets $\varepsilon = \{0.01, 0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4\}$, which correspond to the following noise multipliers: $\{42.384, 2.119, 1.060, 0.706, 0.530, 0.424, 0.353, 0.303\}$.

C.1 DP−SGD AND DP−SIGNSGD: SDE VALIDATION (FIGURE C.1).

In this section, we describe how we validated the SDE models derived in Theorem B.5 and Theorem B.10 (Figure C.1). In line with works in the literature Compagnoni et al. (2025c;a), we optimize a quadratic and a quartic function. We run both DP−SGD and DP−SignSGD, calculating the full gradient and injecting noise as described in Assumption B.2. Similarly, we integrate our SDEs using the Euler-Maruyama algorithm (See, e.g., (Compagnoni et al., 2025c), Algorithm 1) with $\Delta t = \eta$. Results are averaged over 200 repetitions. For each of the two functions, the details are presented in the following paragraphs.

Quadratic function: We consider the quadratic function $f(x) = \frac{1}{2}x^\top Hx$, with $H = 0.1\,\mathrm{diag}(2, 1, \ldots, 1)$, in dimension $d = 1024$. The clipping parameter is set to $C = 5$, and each algorithm is run for $T = 10000$ iterations. The gradient noise scale is $\sigma_\gamma = 1/\sqrt{d}$. The learning rate is $\eta = 0.1$ for DP−SGD and $\eta = 0.01$ for DP−SignSGD. The differential privacy parameters are $(\varepsilon, \delta, q) = (5, 10^{-4}, 10^{-4})$, corresponding to a noise multiplier of $\sigma_{DP} = 0.03$. The initial point is sampled as $x_0 = \frac{50}{\sqrt{d}}\mathcal{N}(0, I_d)$, using an independent seed for each method.

Quartic function: We also test on the quartic function $f(x) = \frac{1}{2}\sum_{i=0}^{d-1} H_{ii}x_i^2 + \frac{\lambda}{4}\sum_{i=0}^{d-1} x_i^4 - \frac{\xi}{3}\sum_{i=0}^{d-1} x_i^3$, where $H = \mathrm{diag}(-2, 1, \ldots, 1)$, $\lambda = 0.5$, and $\xi = 0.1$. The problem dimension, clipping, and number of iterations are the same: $d = 1024$, $C = 5$, $T = 10000$, with gradient noise $\sigma_\gamma = 1/\sqrt{d}$. Both methods use a learning rate of $\eta = 0.01$. The differential privacy parameters are $(\varepsilon, \delta, q) = (5, 10^{-4}, 10^{-4})$ for DP−SGD and $(5, 10^{-4}, 2 \times 10^{-4})$ for DP−SignSGD, corresponding to noise multipliers $\sigma_{DP} = 0.03$ and $\sigma_{DP} = 0.06$, respectively. Initialization is $x_0 = \frac{50}{\sqrt{d}}\mathcal{N}(0, I_d)$ for DP−SGD and $y_0 = -x_0$ for DP−SignSGD.

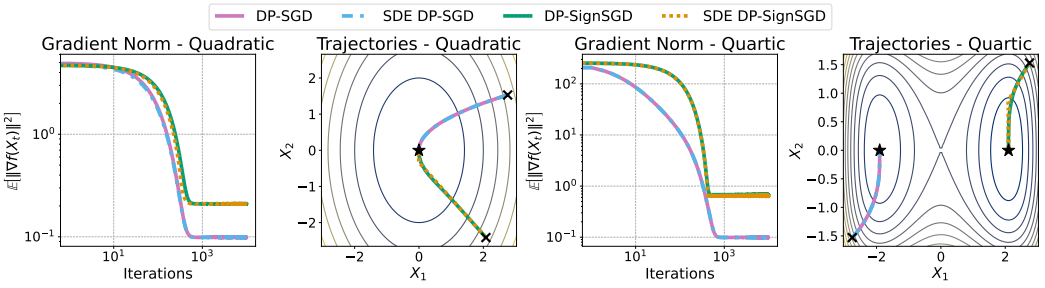

Figure C.1: Consistent with Theorem B.5 and Theorem B.10, we empirically validate that the SDEs of DP−SGD and DP−SignSGD model their respective optimizers. For a convex quadratic function (**left two panels**) and a nonconvex quartic function (**right two panels**), the SDEs accurately track both the trajectories and the gradient norm of the corresponding algorithms, averaged over 200 runs.

C.2 ASYMPTOTIC LOSS BOUND (FIGURES 1 AND C.4)

This section refers to Figure 1 and Figure C.4. We consider three different scenarios: A quadratic function, IMDB, and StackOverflow. Each setup is optimized using DP−SGD, DP−SignSGD, and DP−Adam, and we plot the final averaged training loss across a range of privacy levels. In the left panel, we include the exact bounds from Theorem 4.1 and Theorem 4.3 to show agreement with

theory; in the central and right panels, we compare the final losses with the trends in ε predicted by the same theorems. Experimental details are as follows.

Quadratic: $f(x) = \frac{1}{2}x^\top Hx$, $H = 10I_d$; $d = 1024$, $C = 5$, $T = 50000$, $\sigma_\gamma = 0.01$; learning rate $\eta = 0.01 \cdot \eta_t$ with $\eta_t = (1 + \eta t)^{-0.6}$; Adam parameters: $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$. We used 8 noise multipliers, linearly spaced from 0 to 2, which with $q = 10^{-4}, \delta = 10^{-4}$ correspond to $\varepsilon \in \{\infty, 6.78, 2.38, 1.19, 0.79, 0.59, 0.48, 0.40, 0.34\}$.

IMDB: Hyperparameters are given in Table C.1. We performed 10 runs for each noise multiplier $\{0.5, 1.0, 2.0, 4.0, 6.0, 8.0, 10.0, 12.0\}$, yielding the following values for ε $\{5.425, 2.712, 1.356, 0.678, 0.452, 0.339, 0.271, 0.226\}$, respectively. We report the average training and test loss of the final epoch with confidence bounds (Figure 1 and Figure C.4).

StackOverflow: Hyperparameters are given in Table C.1. We performed 3 runs using for each multiplier $\{0.1, 0.3, 0.5, 1.0, 2.0, 4.0, 6.0, 8.0\}$, yielding the following values for ε $\{4.238, 1.413, 0.848, 0.424, 0.212, 0.106, 0.071, 0.053\}$, respectively. We report the average training and test loss of the final epoch with confidence bounds (Figure 1 and Figure C.4).

C.3 Convergence Speed Analysis (Figure 2)

This section refers to Figure 2. We consider two different scenarios: IMDB and StackOverflow. Each setup is optimized using `DP-SGD`, `DP-SignSGD`, and `DP-Adam` and six different privacy levels: We plot the average trajectories of the training losses and observe that, when it converges, the convergence speed of `DP-SGD` does not depend on the level of privacy, while the two adaptive method are more resilient to the demands of high levels of privacy, but their convergence speed changes for every ε, as predicted in Theorem 4.3.

IMDB: Hyperparameters are given in Table C.1. We performed 10 runs for each noise multiplier $\{0.8, 1.0, 1.2, 1.6, 4.0, 6.0\}$ and corresponding epsilons $\{3.390, 2.712, 2.260, 1.695, 0.678, 0.452\}$. We report the average trajectories of the training loss with confidence bounds (Figure 2).

StackOverflow: Hyperparameters are given in Table C.1. We performed 3 runs for each noise multiplier $\{0.37, 0.5, 0.64, 1.19, 1.46, 1.73\}$ and corresponding epsilons $\{1.146, 0.848, 0.662, 0.356, 0.290, 0.245\}$. We report the average trajectories of the training loss with confidence bounds (Figure 2).

C.4 When Adaptivity Really Matters (Figure 3)

This section refers to Figure 3 and Figure C.2. Each setup is optimized using `DP-SGD`, `DP-SignSGD`, and `DP-Adam`. We consider different batch sizes and for each we plot the final loss values for different privacy levels, similarly to Section C.2. We highlight the possible range of $\varepsilon^\star$ and a dash-dotted line to mark its approximate value, suggested by each graph. As predicted by Theorem 4.5, the empirical value of $\varepsilon^\star$ shifts left as we increase the batch size. Experimental details are as follows.

IMDB: Hyperparameters are given in Table C.1. We select a wide range of noise multipliers: $\{0.5, 1.0, 1.2, 1.5, 1.8, 2.0, 2.2, 2.5, 2.8, 3.0, 3.2, 3.5, 3.8, 4.0, 4.5, 5.0, 6.0, 8.0, 10.0, 12.0\}$ and increasing batch sizes $B = \{48, 56, 64, 72, 80\}$. The corresponding epsilons are

$B = 48$: $\{4.698, 2.349, 1.879, 1.566, 1.342, 1.174, 1.044, 0.940, 0.854, 0.783, 0.723, 0.671, 0.626, 0.587, 0.522, 0.470, 0.391, 0.294, 0.235, 0.196\}$;

$B = 56$: $\{5.070, 2.535, 2.028, 1.690, 1.449, 1.268, 1.127, 1.014, 0.922, 0.845, 0.780, 0.724, 0.676, 0.634, 0.563, 0.507, 0.423, 0.317, 0.254, 0.211\}$;

$B = 64$: $\{5.425, 2.712, 2.170, 1.808, 1.550, 1.356, 1.205, 1.085, 0.986, 0.904, 0.835, 0.775, 0.723, 0.678, 0.603, 0.542, 0.452, 0.339, 0.271, 0.226\}$;

$B = 72$: $\{5.740, 2.870, 2.296, 1.913, 1.640, 1.435, 1.276, 1.148, 1.044, 0.957, 0.883, 0.820, 0.765, 0.717, 0.638, 0.574, 0.478, 0.359, 0.287, 0.239\}$;

$B = 80$: $\{6.070, 3.035, 2.428, 2.023, 1.734, 1.517, 1.349, 1.214, 1.104, 1.012, 0.934, 0.867, 0.809, 0.759, 0.674, 0.607, 0.506, 0.379, 0.303, 0.253\}$.

For each batch size, we performed 10 runs and plotted the average final value of the Train Loss and the empirical $\varepsilon^\star$: these observed values follow the direction indicated in Thm. 4.5. For visualization purposes, we show only a smaller window of ε values satisfying $0.75 \leq \varepsilon \leq 1.25$.

StackOverflow: Due to the higher computational cost required, with our limited resources we managed to select only a restricted range of noise multipliers: $\{0.1, 0.3, 0.5, 1.0, 2.0, 4.0, 6.0, 8.0\}$ and batch sizes: $\{48, 56, 64\}$. The corresponding epsilons are

$B = 48$: $\{1.223, 0.734, 0.367, 0.184, 0.092, 0.061, 0.046\}$;

$B = 56$: $\{1.322, 0.793, 0.396, 0.198, 0.099, 0.066, 0.050\}$;

$B = 64$: $\{1.413, 0.848, 0.424, 0.212, 0.106, 0.071, 0.053\}$.

For each batch size, we performed 3 runs and plotted the average final value of the Train Loss and the empirical $\varepsilon^\star$: these observed values follow the direction indicated in Thm. 4.5. For visualization purposes, we show only a smaller window of ε values satisfying $0.08 \leq \varepsilon \leq 1.1$.

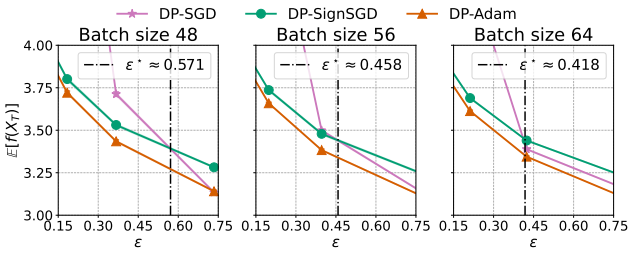

Figure C.2: StackOverflow: From left to right, we decrease the batch noise, i.e., increase the batch size, taking values $B = 48, 56, 64$: As per Theorem 4.5, the privacy threshold $\varepsilon^\star$ that determines when `DP-SignSGD` is more advantageous than `DP-SGD` shifts to the left. This confirms that if there is more noise due to the batch size, less privacy noise is needed for `DP-SignSGD` to be preferable over `DP-SGD`.

C.5 Best-Tuned Hyperparameters (Figures 4)

This section refers to Figure 4. On top of the hyperparameter sweep performed described in Section C, we additionally tune `DP-SGD` for the smaller values of ε. As predicted by Theorem 4.6, the optimal learning rate for `DP-SGD` scales with ε, while those of the adaptive methods are almost constant. Furthermore, we observe that once we reach the limits of the hyperparameter grid, `DP-SGD` loses performance drastically.

IMDB: We additionally tune `DP-SGD` using $\eta = \{0.001, 0.005, 0.01, 0.05, 0.1, 0.5\}$ and $C = \{0.1, 0.25, 0.5\}$ and add the corresponding values using the cyan line. On the left, we plot the average of the final 5 train loss values and confidence bound for each method against the privacy budget ε; On the right, we focus on the scaling of the optimal learning rate with respect to ε.

StackOverflow: We additionally tune `DP-SGD` using $\eta = \{0.001, 0.01, 0.05\}$ and add the corresponding values using the cyan line. As above, on the left, we plot the average of the final 5 training loss values and confidence bounds for each method against the privacy budget ε; on the right, we focus on the scaling of the optimal learning rate with respect to ε.

C.6 Stationary Distributions

In this paragraph, we describe how we validated the convergence behavior predicted in Theorem B.9 and Theorem B.15. To produce Figure C.3, we run both `DP-SGD` and `DP-SignSGD` on $f(x) = \frac{1}{2}x^\top Hx$, where $H = \mathrm{diag}(2, 1)$, $x_0 = (0.01, 0.005)$, $\eta = 0.001$, $\sigma_\gamma = \sigma_{DP} = 0.1$, $C = 5$. We average over 20000 runs and plot the evolution of the moments compared to the theoretical prediction provided in Theorem B.9 and Theorem B.15.

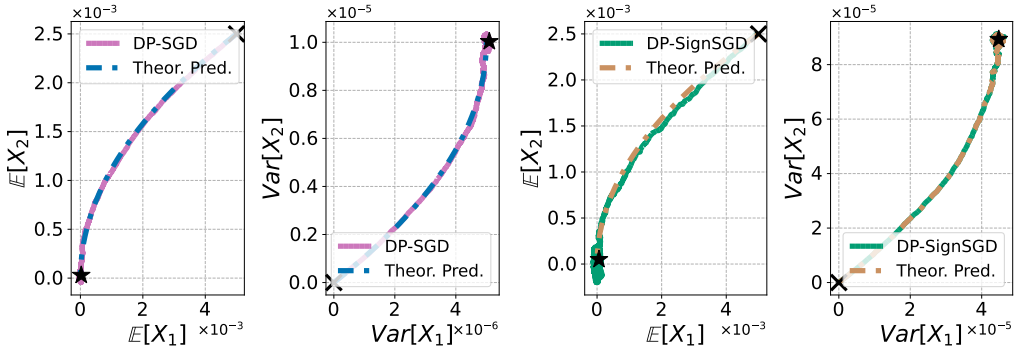

Figure C.3: The empirical dynamics of the first and second moments of the iterates X_t of DP−SGD (left two panels) and of DP−SignSGD (right two panels) match that prescribed in Theorem B.9 and Theorem B.15, respectively.

C.7 Additional Results — Test Loss

Interestingly, the insights provided in Theorem 4.1 and Theorem 4.3 regarding both the asymptotic bound and the convergence speed extend, in practice, also to the test loss. In the same set-up of Section C.2, we plot the asymptotic values of the Test Loss and interpolate with $\mathcal{O}(1/\varepsilon)$ and $\mathcal{O}(1/\varepsilon^2)$ to show that they match the predicted scaling.

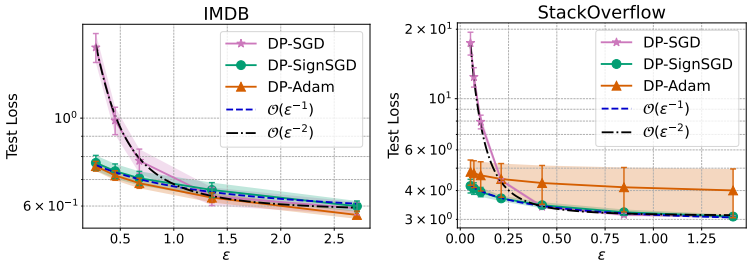

Figure C.4: Privacy-utility trade-off on the *test loss*, comparing DP−SGD, DP−SignSGD, and DP−Adam. **Left:** Logistic regression on the IMDB dataset. **Right:** Logistic regression on the StackOverflow dataset. In both cases, the empirical scalings predicted by Thm. 4.1 and Thm. 4.3 carry over from training to test: DP−SGD follows the $\frac{1}{\varepsilon^2}$ trend, while adaptive methods follow the $\frac{1}{\varepsilon}$ trend. This demonstrates that not only do our theoretical insights generalize to the widely used DP−Adam, but also extend from *training* to *test* loss.

Similarly, in the same set-up as Section C.3, we plot the trajectories of the Test Loss (Fig. C.5): we observe that once again the convergence speed of DP−SGD is not affected by the choice of ε, while adaptive methods clearly present different ε-dependent rates.

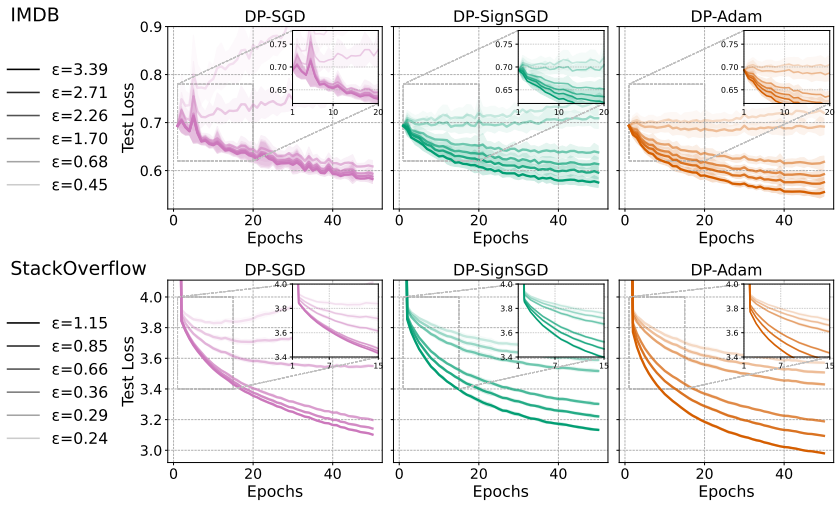

Figure C.5: We compare the Test Loss of `DP-SGD`, `DP-SignSGD`, and `DP-Adam` as we train a logistic regression on the IMDB dataset (**Top Row**) and on the StackOverflow dataset (**Bottom Row**).

D LIMITATIONS

As highlighted by Li et al. (2021b), the approximation capability of SDEs can break down when the learning rate η is large or when certain regularity assumptions on ∇f and the noise covariance matrix are not fulfilled. Although such limitations can, in principle, be alleviated by employing higher-order weak approximations, our position is that the essential function of SDEs is to provide a simplified yet faithful description of the discrete dynamics that offers practical insight. We do not anticipate that raising the approximation order beyond what is required to capture curvature-dependent effects would deliver substantial additional benefits.

We stress that our SDE formulations have been thoroughly validated empirically: the derived SDEs closely track their corresponding optimizers across a wide range of architectures, including MLPs, CNNs, ResNets, and ViTs (Paquette et al., 2021; Malladi et al., 2022; Compagnoni et al., 2024; 2025c;a; Xiao et al., 2025; Marshall et al., 2025).

Acknowledgments. We acknowledge the use of OpenAI's ChatGPT as a writing assistant to help us rephrase and refine parts of the manuscript. All technical content, derivations, and scientific contributions remain the sole responsibility of the authors.