Adaptive Methods Are Preferable in High Privacy Settings: An SDE Perspective

University of Basel

Enea Monzio Compagnoni, Alessandro Stanghellini, Rustem Islamov, Aurelien Lucchi, Anastasiia Koloskova

Universität Zürich[UZH]

Core Question and Setup

Question. Differential privacy adds noise to the training process. Does this hurt adaptive and non-adaptive optimizers equally?

Two practical protocols.

- ▷ **Protocol A:** tune once, then vary the privacy budget ε without re-tuning;
- ▷ **Protocol B:** re-tune hyperparameters for every target ε.

Takeaway. The best private optimizer depends on the privacy budget, the batch noise, and if re-tuning is feasible.

Definitions

Let g_k denote the clipped stochastic (per-example) gradient after privacy perturbation:

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

and $\mathcal{C}[\cdot]$ be the clipping function $\mathcal{C}[x] = \min\left\{\frac{C}{\|x\|_2}, 1\right\} x$.

Differentially Private SGD is

$$x_{k+1} = x_k - \eta g_k. \tag{2}$$

Differentially Private SignSGD is the adaptive method s.t.

$$x_{k+1} = x_k - \eta \mathrm{sign}(g_k). \tag{3}$$

Main Scientific Message

1. Under **fixed hyperparameters,** DP noise affects adaptive and non-adaptive methods in **structurally different** ways.
2. Under **best tuning**, the methods reach comparable asymptotic neighborhoods, but the optimal learning rate of **DP-SGD** scales with ε, while that of **DP-SignSGD** is essentially ε-independent.
3. Empirically, the same qualitative story transfers from training to **test loss** and from **DP-SignSGD** to **DP-Adam**.

Bottom line. SDEs make this analysis possible: they provide a clean and powerful continuous-time lens to understand how DP noise shapes optimization.

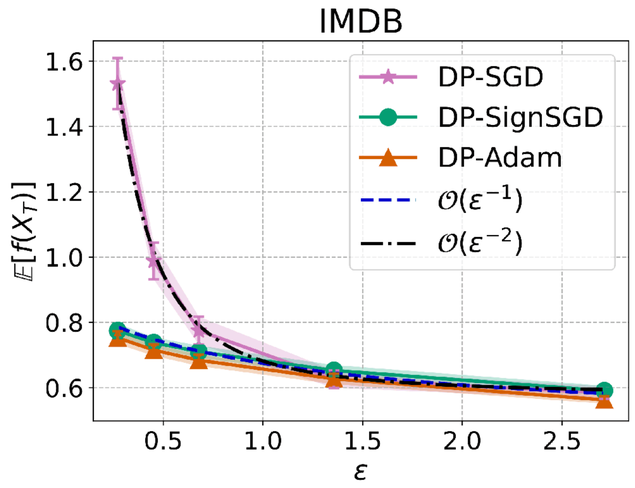
Protocol A: Fixed Hyperparameters

Hyperparameters are tuned once and then transferred across privacy budgets without re-tuning.

Method	Convergence speed	Privacy term
DP-SGD	indep. of ε	$\mathcal{O}(\varepsilon^{-2})$
DP-SignSGD	proport. to ε	$\mathcal{O}(\varepsilon^{-1})$

Takeaway. Without re-tuning, adaptive methods are preferable in high-privacy regimes as their privacy-utility degradation is milder.

Empirical Scaling Matches the Theory



The scalings predicted by our theory are found empirically, and the insights extend empirically from DP-SignSGD to **DP-Adam**.

Main Practical Message

If batch noise is large, **DP-SignSGD** has better privacy-utility trade-off.

If batch noise is smaller, there exists a critical privacy level $\varepsilon^\star$ such that adaptive methods win for $\varepsilon < \varepsilon^\star$, while **DP-SGD** can be better for looser privacy.

Bottom line. DP-SignSGD is best if the batch noise is too large or if the privacy requirements are sufficiently tight.

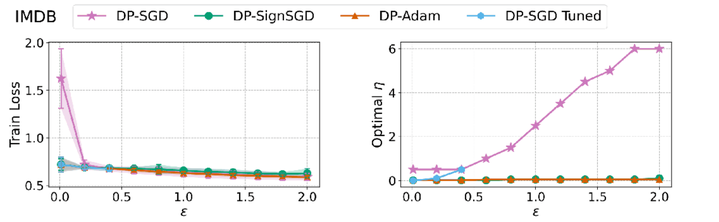
Protocol B: Best-Tuned Hyperparameters

Hyperparams are tuned for each privacy budget. We find that:

1. The optimal learning rate of DP-SGD is $\eta^\star_{\text{DP-SGD}} \propto \varepsilon$
2. The optimal learning rate of DP-SignSGD is **independent** of ε.
3. Under optimal $\eta^\star$, their asymptotic performance is $\widetilde{\mathcal{O}}(\varepsilon^{-1})$

Takeaway. With careful re-tuning, **DP-SGD** can catch up asymptotically, but adaptive methods remain substantially easier to tune across privacy budgets.

Empirical Tuning Results



1. **Left**: Under optimal learning rate, both algorithms achieve comparable performance in line with the theory. However, if the grid swept is not large enough, **DP-SGD** severely underperforms.
2. **Right**: The optimal tuned learning rates scale in accordance with our theory.

Takeaway. Adaptive methods remain much easier to tune across privacy budgets.

Main Practical Message

- ▷ The key difference is not the final asymptotic order, but the **tuning burden**: the optimal learning rate of **DP-SGD** scales with ε, whereas for **DP-SignSGD** it remains approximately constant.
- ▷ **DP-SGD** requires an ε-aware search grid. Adaptive methods transfer η much more easily across privacy budgets.
- ▷ In differential privacy, this matters twice: extra sweeps cost compute and also consume additional privacy budget.

Bottom line. Adaptive methods remain substantially easier to tune in practice.