

APPENDIX

The appendix is structured as follows: Appendix A presents details on the choice of the sigmoid function for clipping. The complexity analysis for DP-SGD-AdaSig is given in Appendix B. An approximation of the sample loss derivative is provided in Appendix C. The proof of Lemma 5.1 is presented in Appendix D. The proof of Theorem 7.5 and the preliminary lemmas and theorems related to the convergence analysis are included in Appendix E. The detailed experimental setup is presented in Appendix F. A numerical analysis of the direction and magnitude deviations is provided in Appendix G. The ablation study of AdaSig is presented in Appendix H. Finally, Appendix I and Appendix J report the variation of α_t across training iterations and the numerical convergence results, respectively.

A Details on the Choice of Sigmoid Function for Clipping

The sigmoid function is simple yet well-suited to our goal of adaptively balancing the trade-off between direction and magnitude deviations. In particular, it enables control over the range of the linear span in Figure 1 through the parameter α . Although PSAC [36] and Auto-S [4] could be extended to variants that adjust their linear region span by varying the constant parameter r throughout training, the sigmoid function provides greater flexibility and control.

Specifically, in the PSAC clipping function, varying r alters the span of its linear region; however, the change in this span is confined to a limited range. This behavior can be illustrated by analyzing the PSAC clipping function, $\tilde{\mathbf{g}}_{i,t} = \frac{C\mathbf{g}_{i,t}}{\|\mathbf{g}_{i,t}\| + \frac{r}{\|\mathbf{g}_{i,t}\| + r}}$, in extreme cases. Due to the term $\frac{r}{\|\mathbf{g}_{i,t}\| + r}$ in the denominator, varying r from 0 to ∞ causes the clipped gradient norm curve $\|\tilde{\mathbf{g}}_{i,t}\|$ (shown in Figure 1) to transition only between two limiting curves: C and $\frac{C\|\mathbf{g}_{i,t}\|}{\|\mathbf{g}_{i,t}\| + 1}$. Consequently, the span of the linear region in PSAC is restricted to lie between these two curves, and adaptively updating r cannot effectively balance the trade-off between direction and magnitude deviations. In contrast, in the AdaSig clipping function of (6), the clipped gradient norm curve $\|\tilde{\mathbf{g}}_{i,t}\|$ (illustrated in Figure 1) can vary from C (as $\alpha \rightarrow \infty$) down to values arbitrarily close to 0 (as $\alpha \rightarrow 0$), thereby providing greater flexibility in adjusting the span of the linear region.

While the Auto-S clipping function, $\tilde{\mathbf{g}}_{i,t} = \frac{C\mathbf{g}_{i,t}}{\|\mathbf{g}_{i,t}\| + r}$, does not have the restricted variation of the linear region span seen in PSAC, it is still less flexible than the sigmoid function. Its linear region span is less responsive to changes in r than the sigmoid function is to α , so larger adjustments of r are required to produce a noticeable effect. In contrast, the sigmoid function, with its exponential dependence on α , offers more flexible control over the linear region and thus the direction–magnitude mismatch trade-off.

B Complexity Analysis

Compared with DP-SGD (with vanilla clipping), one can observe that DP-SGD-AdaSig (Algorithm 1) introduces a negligible increase in computational and memory costs. Specifically, with respect to computation, the algorithm performs lines 8, 12, and 14 in addition to the standard operations in vanilla clipping. Line 8 is carried out efficiently using the per-sample gradient norms computed in line 7, the computational cost of step 12 is identical to that of step 11, and line 14 incurs only a constant additional cost per iteration. Consequently, the overall computational cost introduced by these additional steps is dominated by other operations and the cost of per-sample gradient computation, making the additional overall computational cost minimal.

With respect to memory, DP-SGD-AdaSig requires extra memory to store $\hat{r}(\boldsymbol{\theta}_t, \alpha_t; \mathcal{B}_t)$ (line 12), which is required for the slope update in the next iteration (line 14). This additional memory cost is negligible as compared with the memory required for storing the model parameters and per-sample gradients.

C Approximation of the Sample Loss Derivative with Respect to α

Since the loss on sample i in iteration t , denoted by $h^i(\theta_t)$, does not directly depend on α_t , to derive an update rule for α , we unroll the expression for $h^i(\theta_t)$ and write it in terms of θ_{t-1} and α_{t-1} . In particular, using (10), we obtain

$$h^i(\theta_t) = h^i\left(\theta_{t-1} - \frac{\lambda}{B} \hat{s}(\theta_{t-1}, \alpha_{t-1}; \mathcal{B}_{t-1})\right). \quad (21)$$

For sufficiently small learning rate for α (i.e., λ_α in (16)), α_{t-1} and α_t differ only slightly, and hence we adopt the approximation $\alpha_{t-1} \approx \alpha_t$. Consequently, the loss derivative at α_t , $\frac{\partial h^i(\theta_t)}{\partial \alpha_t}$, can be approximated by the loss derivative at α_{t-1} , $\frac{\partial h^i(\theta_t)}{\partial \alpha_{t-1}}$, i.e.,

$$\frac{\partial h^i(\theta_t)}{\partial \alpha_t} \approx \frac{\partial h^i(\theta_t)}{\partial \alpha_{t-1}}. \quad (22)$$

We now apply the chain rule to (21) to compute $\frac{\partial h^i(\theta_t)}{\partial \alpha_{t-1}}$ as

$$\frac{\partial h^i(\theta_t)}{\partial \alpha_{t-1}} = (\nabla_{\theta_t} h^i(\theta_t))^\top \frac{\partial \theta_t}{\partial \alpha_{t-1}}, \quad (23)$$

where the first factor in the RHS of (23) is the sample loss gradient w.r.t. model parameters, i.e., $\mathbf{g}_{i,t}$. To compute the second factor, we write

$$\begin{aligned} \frac{\partial \theta_t}{\partial \alpha_{t-1}} &= \frac{\partial(\theta_{t-1} - \frac{\lambda}{B} \hat{s}(\theta_{t-1}, \alpha_{t-1}; \mathcal{B}_{t-1}))}{\partial \alpha_{t-1}} \\ &\stackrel{(a)}{=} \frac{-\lambda}{B} \sum_{i \in \mathcal{B}_{t-1}} \frac{\partial \psi_{\alpha_{t-1}}(\|\mathbf{g}_{i,t-1}\|)}{\partial \alpha_{t-1}} \frac{\mathbf{g}_{i,t-1}}{\|\mathbf{g}_{i,t-1}\|} \\ &\stackrel{(b)}{=} \frac{-\lambda}{B} \sum_{i \in \mathcal{B}_{t-1}} \frac{2e^{-\alpha_{t-1}\|\mathbf{g}_{i,t-1}\|} \mathbf{g}_{i,t-1}}{(1 + e^{-\alpha_{t-1}\|\mathbf{g}_{i,t-1}\|})^2}, \end{aligned} \quad (24)$$

where (a) follows from the definition of $\hat{s}(\theta_{t-1}, \alpha_{t-1}; \mathcal{B}_{t-1})$ in (9b), and (b) is obtained by computing the derivative of the AdaSig function $\psi_{\alpha_{t-1}}(\cdot)$. Using the definition of $r(\theta_t, \alpha_t; \mathcal{B}_t)$ in (12), we rewrite (24) as

$$\frac{\partial \theta_t}{\partial \alpha_{t-1}} = \frac{-\lambda}{B} r(\theta_{t-1}, \alpha_{t-1}; \mathcal{B}_{t-1}). \quad (25)$$

Substituting (25) in (23), we finally get

$$\frac{\partial h^i(\theta_t)}{\partial \alpha_{t-1}} = -\frac{\lambda}{B} \mathbf{g}_{i,t}^\top r(\theta_{t-1}, \alpha_{t-1}; \mathcal{B}_{t-1}). \quad (26)$$

Using (22) together with (26), we have

$$\frac{\partial h^i(\theta_t)}{\partial \alpha_t} \approx -\frac{\lambda}{B} \mathbf{g}_{i,t}^\top r(\theta_{t-1}, \alpha_{t-1}; \mathcal{B}_{t-1}). \quad (27)$$

D Proof of Lemma 5.1: Sensitivity Bound

Proof. The ℓ_2 -sensitivity of $r(\theta_t, \alpha_t; \mathcal{B}_t)$ w.r.t. the training samples in iteration t is defined as

$$\Delta r = \max_{\mathcal{B}_t, \mathcal{B}'_t} \left\| r(\theta_t, \alpha_t; \mathcal{B}_t) - r(\theta_t, \alpha_t; \mathcal{B}'_t) \right\|_2, \quad (28)$$

where \mathcal{B}_t and \mathcal{B}'_t are two adjacent batches of data that differ by exactly one data point. Without loss of generality, we assume that \mathcal{B}'_t contains all elements of \mathcal{B}_t together with an additional data point (x', y') , where x' is the input feature and y' is its corresponding target output. That is, $\mathcal{B}'_t = \mathcal{B}_t \cup \{(x', y')\}$. Moreover, let \mathbf{g}' denote the gradient of sample loss at (x', y') , i.e., $\mathbf{g}' = \nabla_{\theta} \ell(f_{\theta}(x'), y')$.

Considering the definition of $r(\boldsymbol{\theta}_t, \alpha_t; \mathcal{B}_t)$ in (12), we can conclude that the terms under summation in $r(\boldsymbol{\theta}_t, \alpha_t; \mathcal{B}_t)$ and $r(\boldsymbol{\theta}_t, \alpha_t; \mathcal{B}'_t)$ are identical, except a single term that arises due to the extra element in \mathcal{B}'_t . The ℓ_2 -sensitivity in (28) can hence be upper bounded by the norm of this extra term, i.e.,

$$\Delta r \leq \max_{\mathbf{g}'} \left\| \frac{2e^{-\alpha_t} \|\mathbf{g}'\| \mathbf{g}'}{(1 + e^{-\alpha_t} \|\mathbf{g}'\|)^2} \right\|_2 \quad (29a)$$

$$= \max_{\mathbf{g}'} \frac{2e^{-\alpha_t} \|\mathbf{g}'\| \|\mathbf{g}'\|}{(1 + e^{-\alpha_t} \|\mathbf{g}'\|)^2} \quad (29b)$$

$$\stackrel{(a)}{=} \frac{1}{\alpha_t} \max_{z \geq 0} \frac{2e^{-z} z}{(1 + e^{-z})^2}, \quad (29c)$$

where (a) follows from the variable exchange $z \triangleq \alpha_t \|\mathbf{g}'\|$. We next find the solution to the maximization in (29c) by setting the derivative of the objective to zero, i.e., $\frac{\partial}{\partial z} \frac{2e^{-z} z}{(1 + e^{-z})^2} = 0$, which results in the following fixed-point equation as

$$z = \ln \frac{z + 1}{z - 1}. \quad (30)$$

Let's denote the solution for (30) by z_* . Substituting the solution, i.e., z_* , into objective (29c), we get the upper bound for Δr as

$$\Delta r \leq \frac{1}{\alpha_t} \frac{2e^{-z_*} z_*}{(1 + e^{-z_*})^2} = \frac{z_*^2 - 1}{2z_* \alpha_t}. \quad (31)$$

Solving (30) numerically for z and substituting z_* in (31) results in $\Delta r \leq 0.448/\alpha_t$, which completes the proof. \square

E Theoretical Convergence Analysis

We first introduce the preliminary lemmas and theorems required for proving the main theorem.

E.1 Preliminaries

Lemma E.1. *Suppose that Assumptions 7.2 and 7.4 hold. Under Algorithm 1 the expectation of the population loss difference in two consecutive iterations is upper-bounded by*

$$\mathbb{E} \left[L(\boldsymbol{\theta}_{t+1}) - L(\boldsymbol{\theta}_t) \middle| \boldsymbol{\theta}_t, \alpha_t \right] \leq -\lambda \mathbf{g}_t^\top \mathbb{E} \left[\left(\frac{2}{1 + e^{-\alpha_t} \|\mathbf{v}_t\|} - 1 \right) \frac{\mathbf{v}_t}{\|\mathbf{v}_t\|} \middle| \boldsymbol{\theta}_t, \alpha_t \right] + \beta \lambda^2 \left(1 + \frac{d\sigma_s^2}{2B^2} \right), \quad (32)$$

where $\mathbf{g}_t = \nabla L(\boldsymbol{\theta}_t)$, $\mathbb{E}[\cdot | \boldsymbol{\theta}_t, \alpha_t]$ denote the expectation over the randomness in iteration t for given $\boldsymbol{\theta}_t$ and α_t , and \mathbf{v}_t is the random process from which $\mathbf{g}_{i,t}$ is sampled, i.e., $\mathbf{g}_{i,t} \sim \mathbf{v}_t$ for $i \in \mathcal{B}_t$.

Proof. Based on Assumption 7.2 the population loss is β -smooth. Thus, we have

$$L(\boldsymbol{\theta}_{t+1}) - L(\boldsymbol{\theta}_t) \leq \mathbf{g}_t^\top (\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t) + \frac{\beta}{2} \|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t\|^2. \quad (33)$$

Taking expectation from both sides of (33) for given θ_t and α_t , we have

$$\begin{aligned} & \mathbb{E}[L(\theta_{t+1}) - L(\theta_t) | \theta_t, \alpha_t] \\ & \leq \mathbf{g}_t^\top \mathbb{E}[(\theta_{t+1} - \theta_t) | \theta_t, \alpha_t] + \frac{\beta}{2} \mathbb{E}[\|\theta_{t+1} - \theta_t\|^2 | \theta_t, \alpha_t] \end{aligned} \quad (34)$$

$$\begin{aligned} & \stackrel{(a)}{=} -\frac{\lambda}{B} \mathbf{g}_t^\top \mathbb{E}\left[\sum_{i \in \mathcal{B}_t} \psi_{\alpha_t}(\|\mathbf{g}_{i,t}\|) \frac{\mathbf{g}_{i,t}}{\|\mathbf{g}_{i,t}\|} + \mathbf{n}_t^s | \theta_t, \alpha_t\right] \\ & \quad + \frac{\beta\lambda^2}{2B^2} \mathbb{E}\left[\left\|\sum_{i \in \mathcal{B}_t} \psi_{\alpha_t}(\|\mathbf{g}_{i,t}\|) \frac{\mathbf{g}_{i,t}}{\|\mathbf{g}_{i,t}\|} + \mathbf{n}_t^s\right\|^2 | \theta_t, \alpha_t\right] \end{aligned} \quad (35)$$

$$\begin{aligned} & \stackrel{(b)}{=} -\frac{\lambda}{B} \mathbf{g}_t^\top \mathbb{E}\left[\sum_{i \in \mathcal{B}_t} \psi_{\alpha_t}(\|\mathbf{g}_{i,t}\|) \frac{\mathbf{g}_{i,t}}{\|\mathbf{g}_{i,t}\|} | \theta_t, \alpha_t\right] \\ & \quad + \frac{\beta\lambda^2}{2B^2} \mathbb{E}\left[\left\|\sum_{i \in \mathcal{B}_t} \psi_{\alpha_t}(\|\mathbf{g}_{i,t}\|) \frac{\mathbf{g}_{i,t}}{\|\mathbf{g}_{i,t}\|}\right\|^2 | \theta_t, \alpha_t\right] + \frac{\beta\lambda^2}{2B^2} \mathbb{E}[\|\mathbf{n}_t^s\|^2 | \theta_t, \alpha_t] \end{aligned} \quad (36)$$

$$\begin{aligned} & \stackrel{(c)}{\leq} -\frac{\lambda}{B} \mathbf{g}_t^\top \mathbb{E}\left[\sum_{i \in \mathcal{B}_t} \psi_{\alpha_t}(\|\mathbf{g}_{i,t}\|) \frac{\mathbf{g}_{i,t}}{\|\mathbf{g}_{i,t}\|} | \theta_t, \alpha_t\right] \\ & \quad + \frac{\beta\lambda^2}{2B^2} \mathbb{E}\left[\left(\sum_{i \in \mathcal{B}_t} \left\|\psi_{\alpha_t}(\|\mathbf{g}_{i,t}\|) \frac{\mathbf{g}_{i,t}}{\|\mathbf{g}_{i,t}\|}\right\|\right)^2 | \theta_t, \alpha_t\right] + \frac{\beta\lambda^2}{2B^2} d\sigma_s^2 \end{aligned} \quad (37)$$

$$\stackrel{(d)}{\leq} -\frac{\lambda}{B} \mathbf{g}_t^\top \mathbb{E}\left[\sum_{i \in \mathcal{B}_t} \psi_{\alpha_t}(\|\mathbf{g}_{i,t}\|) \frac{\mathbf{g}_{i,t}}{\|\mathbf{g}_{i,t}\|} | \theta_t, \alpha_t\right] + \beta\lambda^2 \left(\frac{\mathbb{E}|\mathcal{B}_t|^2}{2B^2} + \frac{d\sigma_s^2}{2B^2}\right) \quad (38)$$

$$\stackrel{(e)}{\leq} -\frac{\lambda}{B} \mathbf{g}_t^\top \mathbb{E}\left[\sum_{i \in \mathcal{B}_t} \psi_{\alpha_t}(\|\mathbf{g}_{i,t}\|) \frac{\mathbf{g}_{i,t}}{\|\mathbf{g}_{i,t}\|} | \theta_t, \alpha_t\right] + \beta\lambda^2 \left(1 + \frac{d\sigma_s^2}{2B^2}\right), \quad (39)$$

where (a) follows from (10), (b) comes from the fact that $\mathbf{g}_{i,t}$ for $i \in \mathcal{B}_t$ are independent of the process \mathbf{n}_t^s , and the fact that \mathbf{n}_t^s is zero-mean, (c) is obtained by applying the Triangle inequality to the second term in (36) and substituting the variance of \mathbf{n}_t^s , (d) results from the upper bound

$$\sum_{i \in \mathcal{B}_t} \left\|\psi_{\alpha_t}(\|\mathbf{g}_{i,t}\|) \frac{\mathbf{g}_{i,t}}{\|\mathbf{g}_{i,t}\|}\right\| \leq |\mathcal{B}_t|, \quad (40)$$

since $\psi_{\alpha_t}(x) \leq 1$ for any $x \in \mathbb{R}$, and (e) follows from the fact that, under Poisson sampling with rate B/N , we have

$$\frac{\mathbb{E}[|\mathcal{B}_t|^2]}{B^2} = \frac{B+1}{B} - \frac{1}{N} \leq 1 + \frac{1}{B} \leq 2, \quad (41)$$

where the last inequality holds since the expected batch size $B \geq 1$.

Based on Assumption 7.4, the sample gradients are identically distributed, i.e., $\mathbf{g}_{i,t} \sim \mathbf{v}_t, \forall i$. We can hence write

$$\frac{1}{B} \mathbb{E}\left[\sum_{i \in \mathcal{B}_t} \psi_{\alpha_t}(\|\mathbf{g}_{i,t}\|) \frac{\mathbf{g}_{i,t}}{\|\mathbf{g}_{i,t}\|} | \theta_t, \alpha_t\right] = \mathbb{E}\left[\psi_{\alpha_t}(\|\mathbf{v}_t\|) \frac{\mathbf{v}_t}{\|\mathbf{v}_t\|} | \theta_t, \alpha_t\right]. \quad (42)$$

Substituting into (39) and replacing $\psi_{\alpha_t}(\|\mathbf{v}_t\|)$ with its definition, (32) is concluded. \square

Lemma E.2. Under Assumption 7.4 i.e., $\mathbf{v}_t = \mathbf{g}_t + \Delta_t$, the following equality holds:

$$\mathbf{g}_t^\top \mathbb{E}\left[\left(\frac{2}{1 + e^{-\alpha_t \|\mathbf{v}_t\|}} - 1\right) \frac{\mathbf{v}_t}{\|\mathbf{v}_t\|} | \theta_t, \alpha_t\right] = \frac{\|\mathbf{g}_t\|}{2} \mathbb{E}_{s,c} \left[f(s, c, \gamma_t) | 0 \leq c \leq 1\right], \quad (43)$$

where $\gamma_t \triangleq \alpha_t \|\mathbf{g}_t\|$, $s \triangleq \frac{\|\Delta_t\|}{\|\mathbf{g}_t\|}$, $c \triangleq \frac{\mathbf{g}_t^\top \Delta_t}{\|\Delta_t\| \|\mathbf{g}_t\|}$, and

$$\begin{aligned} f(s, c, \gamma_t) & \triangleq \frac{1 + sc}{\sqrt{1 + s^2 + 2sc}} \left(\frac{2}{1 + e^{-\gamma_t \sqrt{1 + s^2 + 2sc}}} - 1\right) \\ & \quad + \frac{1 - sc}{\sqrt{1 + s^2 - 2sc}} \left(\frac{2}{1 + e^{-\gamma_t \sqrt{1 + s^2 - 2sc}}} - 1\right). \end{aligned} \quad (44)$$

Proof. We adopt an approach similar to that proposed in [4]. Let \mathcal{H}^+ and \mathcal{H}^- denote the following two halfspaces:

$$\mathcal{H}^+ \triangleq \{\mathbf{o} \in \mathbb{R}^d | \mathbf{g}_t^\top \mathbf{o} \geq 0\}, \quad (45)$$

$$\mathcal{H}^- \triangleq \{\mathbf{o} \in \mathbb{R}^d | \mathbf{g}_t^\top \mathbf{o} \leq 0\}. \quad (46)$$

Using $\mathbf{v}_t = \mathbf{g}_t + \Delta_t$, we can write

$$\mathbf{g}_t^\top \mathbb{E} \left[\left(\frac{2}{1 + e^{-\alpha_t \|\mathbf{v}_t\|}} - 1 \right) \frac{\mathbf{v}_t}{\|\mathbf{v}_t\|} \middle| \boldsymbol{\theta}_t, \alpha_t \right] \quad (47)$$

$$= \mathbb{E}_{\Delta_t} \left[\left(\frac{2}{1 + e^{-\alpha_t \|\mathbf{g}_t + \Delta_t\|}} - 1 \right) \frac{\mathbf{g}_t^\top (\mathbf{g}_t + \Delta_t)}{\|\mathbf{g}_t + \Delta_t\|} \right] \quad (48)$$

$$\stackrel{(a)}{=} \frac{1}{2} \mathbb{E}_{\Delta_t} \left[\left(\frac{2}{1 + e^{-\alpha_t \sqrt{\|\mathbf{g}_t\|^2 + \|\Delta_t\|^2 + 2\mathbf{g}_t^\top \Delta_t}}} - 1 \right) \frac{\mathbf{g}_t^\top \mathbf{g}_t + \mathbf{g}_t^\top \Delta_t}{\sqrt{\|\mathbf{g}_t\|^2 + \|\Delta_t\|^2 + 2\mathbf{g}_t^\top \Delta_t}} \middle| \Delta_t \in \mathcal{H}^+ \right] \\ + \frac{1}{2} \mathbb{E}_{\Delta_t} \left[\left(\frac{2}{1 + e^{-\alpha_t \sqrt{\|\mathbf{g}_t\|^2 + \|\Delta_t\|^2 + 2\mathbf{g}_t^\top \Delta_t}}} - 1 \right) \frac{\mathbf{g}_t^\top \mathbf{g}_t + \mathbf{g}_t^\top \Delta_t}{\sqrt{\|\mathbf{g}_t\|^2 + \|\Delta_t\|^2 + 2\mathbf{g}_t^\top \Delta_t}} \middle| \Delta_t \in \mathcal{H}^- \right] \quad (49)$$

$$\stackrel{(b)}{=} \frac{1}{2} \mathbb{E}_{\Delta_t} \left[\left(\frac{2}{1 + e^{-\alpha_t \sqrt{\|\mathbf{g}_t\|^2 + \|\Delta_t\|^2 + 2\mathbf{g}_t^\top \Delta_t}}} - 1 \right) \frac{\mathbf{g}_t^\top \mathbf{g}_t + \mathbf{g}_t^\top \Delta_t}{\sqrt{\|\mathbf{g}_t\|^2 + \|\Delta_t\|^2 + 2\mathbf{g}_t^\top \Delta_t}} \middle| \Delta_t \in \mathcal{H}^+ \right] \\ + \frac{1}{2} \mathbb{E}_{\Delta_t} \left[\left(\frac{2}{1 + e^{-\alpha_t \sqrt{\|\mathbf{g}_t\|^2 + \|\Delta_t\|^2 - 2\mathbf{g}_t^\top \Delta_t}}} - 1 \right) \frac{\mathbf{g}_t^\top \mathbf{g}_t - \mathbf{g}_t^\top \Delta_t}{\sqrt{\|\mathbf{g}_t\|^2 + \|\Delta_t\|^2 - 2\mathbf{g}_t^\top \Delta_t}} \middle| \Delta_t \in \mathcal{H}^+ \right], \quad (50)$$

where (a) follows from conditioning the expectation w.r.t. Δ_t on two halfspaces, and using the fact that under the symmetric distribution assumption on Δ_t (i.e., $\Delta_t \stackrel{\mathcal{D}}{=} -\Delta_t$ in Assumption 7.4), we have

$$\Pr\{\Delta_t \in \mathcal{H}^+\} = \Pr\{\Delta_t \in \mathcal{H}^-\} = \frac{1}{2}, \quad (51)$$

and (b) follows again from symmetric distribution assumption on Δ_t and along with the variable exchange $\Delta_t = -\Delta_t$ in the second term of (49).

We now rewrite (50) in terms of the random variables s and $c \in [-1, 1]$ introduced in the lemma. Noting that the event $\Delta_t \in \mathcal{H}^+$ corresponds to $c \geq 0$, we can write

$$\frac{1}{2} \mathbb{E}_{\Delta_t} \left[\left(\frac{2}{1 + e^{-\alpha_t \sqrt{\|\mathbf{g}_t\|^2 + \|\Delta_t\|^2 + 2\mathbf{g}_t^\top \Delta_t}}} - 1 \right) \frac{\mathbf{g}_t^\top \mathbf{g}_t + \mathbf{g}_t^\top \Delta_t}{\sqrt{\|\mathbf{g}_t\|^2 + \|\Delta_t\|^2 + 2\mathbf{g}_t^\top \Delta_t}} \middle| \Delta_t \in \mathcal{H}^+ \right] \\ + \frac{1}{2} \mathbb{E}_{\Delta_t} \left[\left(\frac{2}{1 + e^{-\alpha_t \sqrt{\|\mathbf{g}_t\|^2 + \|\Delta_t\|^2 - 2\mathbf{g}_t^\top \Delta_t}}} - 1 \right) \frac{\mathbf{g}_t^\top \mathbf{g}_t - \mathbf{g}_t^\top \Delta_t}{\sqrt{\|\mathbf{g}_t\|^2 + \|\Delta_t\|^2 - 2\mathbf{g}_t^\top \Delta_t}} \middle| \Delta_t \in \mathcal{H}^+ \right] \quad (52)$$

$$= \frac{\|\mathbf{g}_t\|}{2} \mathbb{E}_{s,c} \left[\frac{1+sc}{\sqrt{1+s^2+2sc}} \left(\frac{2}{1 + e^{-\gamma_t \sqrt{1+s^2+2sc}}} - 1 \right) \middle| 0 \leq c \leq 1 \right] \\ + \frac{\|\mathbf{g}_t\|}{2} \mathbb{E}_{s,c} \left[\frac{1-sc}{\sqrt{1+s^2-2sc}} \left(\frac{2}{1 + e^{-\gamma_t \sqrt{1+s^2-2sc}}} - 1 \right) \middle| 0 \leq c \leq 1 \right], \quad (53)$$

where the equality follows directly from the definitions of s and c . Expressing (53) in terms of $f(s, c, \gamma_t)$ yields the result. \square

Lemma E.3. The function $\frac{\psi_\alpha(x)}{x}$ with $\psi_\alpha(x)$ defined in (5) is a non-increasing function in x for $\alpha \geq 0$, $x \geq 0$.

Proof. Computing the derivative of the function w.r.t. x and simplifying the terms yields the following:

$$\frac{\partial(\psi_\alpha(x)/x)}{\partial x} = \frac{\partial \psi_\alpha(x)}{\partial x} \frac{1}{x} - \frac{1}{x^2} \psi_\alpha(x) \\ = \frac{2e^{-\alpha x}(\alpha x - \sinh \alpha x)}{x^2(1 + e^{\alpha x})^2}, \quad (54)$$

where the RHS is a non-positive term, since $\alpha x \leq \sinh \alpha x$ for $\alpha \geq 0, x \geq 0$. Thus, the derivative is non-positive, which shows that the function is non-increasing. \square

Lemma E.4. *The function $\frac{\partial \psi_\alpha(x)}{\partial x}$ with $\psi_\alpha(x)$ defined in (5) is a non-increasing function in x for $\alpha \geq 0, x \geq 0$.*

Proof. We compute the derivative of the function as

$$\begin{aligned} \frac{\partial(\partial \psi_\alpha(x)/\partial x)}{\partial x} &= \partial \left(\frac{2\alpha e^{-\alpha x}}{(1 + e^{-\alpha x})^2} \right) / \partial x \\ &= \frac{-2\alpha^2 e^{-\alpha x} (1 - e^{-\alpha x})}{(1 + e^{-\alpha x})^3}, \end{aligned} \quad (55)$$

which is non-positive for $\alpha \geq 0$ and $x \geq 0$, which leads to the desired conclusion. \square

Lemma E.5. *For the function $\psi_\alpha(x)$ defined in (5), the following holds:*

$$\frac{\partial \psi_\alpha(x)}{\partial x} \leq \frac{\psi_\alpha(x)}{x}, \quad \alpha \geq 0, x \geq 0. \quad (56)$$

Proof. We begin with the inequality $\alpha x \leq \sinh \alpha x$ for $x \geq 0$ and $\alpha \geq 0$. We have

$$\begin{aligned} \alpha x \leq \sinh \alpha x &\stackrel{(a)}{\Rightarrow} 2\alpha x e^{-\alpha x} \leq 1 - e^{-2\alpha x} \\ &\stackrel{(b)}{\Rightarrow} \frac{2\alpha e^{-\alpha x}}{(1 + e^{-\alpha x})^2} \leq \frac{1}{x} \left(\frac{2}{1 + e^{-\alpha x}} - 1 \right) \\ &\stackrel{(c)}{\Rightarrow} \frac{\partial \psi_\alpha(x)}{\partial x} \leq \frac{\psi_\alpha(x)}{x}, \end{aligned} \quad (57)$$

where (a) follows from multiplying both sides of the previous inequality by $2e^{-\alpha x}$, (b) follows from dividing both sides by $x(1 + e^{-\alpha x})^2$, and (c) results from writing both sides in terms of $\psi_\alpha(x)$ and $\frac{\partial \psi_\alpha(x)}{\partial x}$. \square

Theorem E.6. *Let $f(s, c, \gamma_t)$ be the function defined in Lemma E.2. Then, the following properties hold:*

1. $f(s, c, \gamma_t)$ is non-increasing in s for $0 \leq c \leq 1$ and $\gamma_t \geq 0$.
2. $f(s, c, \gamma_t) \geq 0$ for $0 \leq c \leq 1$ and $\gamma_t \geq 0$.
3. $f(s, c, \gamma_t)$ is non-increasing in c on the interval $0 \leq c \leq 1$, for $s \geq 1$ and $\gamma_t \geq 0$.

Proof. We first prove the first property, showing that $f(s, c, \gamma_t)$ is non-increasing in s . Based on (44) in Lemma E.2, we rewrite $f(s, c, \gamma_t)$ as

$$f(s, c, \gamma_t) = f_1(s, c) \psi_{\gamma_t}(x_1) + f_2(s, c) \psi_{\gamma_t}(x_2), \quad (58)$$

where we define $x_1 \triangleq \sqrt{1 + s^2 + 2sc}$, $x_2 \triangleq \sqrt{1 + s^2 - 2sc}$, and

$$f_1(s, c) \triangleq \frac{1 + sc}{\sqrt{1 + s^2 + 2sc}}, \quad (59)$$

$$f_2(s, c) \triangleq \frac{1 - sc}{\sqrt{1 + s^2 - 2sc}}. \quad (60)$$

Taking the derivative of $f(s, c, \gamma_t)$ w.r.t. s , we have

$$\begin{aligned} \frac{\partial f(s, c, \gamma_t)}{\partial s} &= \frac{\partial f_1(s, c)}{\partial s} \psi_{\gamma_t}(x_1) + f_1(s, c) \frac{\partial \psi_{\gamma_t}(x_1)}{\partial x_1} \frac{\partial x_1}{\partial s} \\ &\quad + \frac{\partial f_2(s, c)}{\partial s} \psi_{\gamma_t}(x_2) + f_2(s, c) \frac{\partial \psi_{\gamma_t}(x_2)}{\partial x_2} \frac{\partial x_2}{\partial s}. \end{aligned} \quad (61)$$

Substituting the derivatives into the above expression results in

$$\begin{aligned} \frac{\partial f(s, c, \gamma_t)}{\partial s} = & \frac{-s(1-c^2)}{(1+s^2+2sc)^{\frac{3}{2}}} \psi_{\gamma_t}(x_1) + \frac{\partial \psi_{\gamma_t}(x_1)}{\partial x_1} \left(c + \frac{s(1-c^2)}{1+s^2+2sc} \right) \\ & + \frac{-s(1-c^2)}{(1+s^2-2sc)^{\frac{3}{2}}} \psi_{\gamma_t}(x_2) + \frac{\partial \psi_{\gamma_t}(x_2)}{\partial x_2} \left(-c + \frac{s(1-c^2)}{1+s^2-2sc} \right). \end{aligned} \quad (62)$$

Rearranging the terms, we have

$$\begin{aligned} \frac{\partial f(s, c, \gamma_t)}{\partial s} = & -c \left(\frac{\partial \psi_{\gamma_t}(x_2)}{\partial x_2} - \frac{\partial \psi_{\gamma_t}(x_1)}{\partial x_1} \right) \\ & - \frac{s(1-c^2)}{1+s^2+2sc} \left(\frac{\psi_{\gamma_t}(x_1)}{x_1} - \frac{\partial \psi_{\gamma_t}(x_1)}{\partial x_1} \right) \\ & - \frac{s(1-c^2)}{1+s^2-2sc} \left(\frac{\psi_{\gamma_t}(x_2)}{x_2} - \frac{\partial \psi_{\gamma_t}(x_2)}{\partial x_2} \right). \end{aligned} \quad (63)$$

The first term in the RHS of (63) is non-positive since, by Lemma E.4, the derivative of $\psi_{\alpha}(x)$ is a non-increasing function of x , and we have $x_2 \leq x_1$ for $c \geq 0, s \geq 0$. The second and third terms are also non-positive because $c \leq 1$ and, by Lemma E.5, it holds that

$$\frac{\psi_{\gamma_t}(x)}{x} \geq \frac{\partial \psi_{\gamma_t}(x)}{\partial x}, \quad \forall x \geq 0, \forall \gamma_t \geq 0. \quad (64)$$

Thus, we conclude that $\frac{\partial f(s, c, \gamma_t)}{\partial s} \leq 0$, implying that $f(s, c, \gamma_t)$ is non-increasing in s .

To prove the second property, we first evaluate the limit of the function as s approaches infinity:

$$\lim_{s \rightarrow \infty} f(s, c, \gamma_t) = 0, \quad 1 \geq c \geq 0, \gamma_t \geq 0. \quad (65)$$

By the first property, $f(s, c, \gamma_t)$ is non-increasing in s , which implies that its minimum value is zero.

To establish the third property, we compute the derivative of $f(s, c, \gamma_t)$ w.r.t. c . Based on (58), we have

$$\begin{aligned} \frac{\partial f(s, c, \gamma_t)}{\partial c} = & \frac{\partial f_1(s, c)}{\partial c} \psi_{\gamma_t}(x_1) + f_1(s, c) \frac{\partial \psi_{\gamma_t}(x_1)}{\partial x_1} \frac{\partial x_1}{\partial c} + \frac{\partial f_2(s, c)}{\partial c} \psi_{\gamma_t}(x_2) \\ & + f_2(s, c) \frac{\partial \psi_{\gamma_t}(x_2)}{\partial x_2} \frac{\partial x_2}{\partial c}. \end{aligned} \quad (66)$$

Plugging the derivatives into the above expression and rearranging the terms, we obtain

$$\begin{aligned} \frac{\partial f(s, c, \gamma_t)}{\partial c} = & \frac{\psi_{\gamma_t}(x_1)}{x_1} \frac{s^2(s+c)}{1+s^2+2sc} - \frac{\psi_{\gamma_t}(x_2)}{x_2} \frac{s^2(s-c)}{1+s^2-2sc} \\ & + \frac{\partial \psi_{\gamma_t}(x_1)}{\partial x_1} \frac{s(1+sc)}{1+s^2+2sc} - \frac{\partial \psi_{\gamma_t}(x_2)}{\partial x_2} \frac{s(1-sc)}{1+s^2-2sc}. \end{aligned} \quad (67)$$

Next, we find an upper bound for $\frac{\partial f(s, c, \gamma_t)}{\partial c}$. Based on Lemma E.3, $\frac{\psi_{\gamma_t}(x)}{x}$ is a non-increasing function in x . Thus, we have

$$\frac{\psi_{\gamma_t}(x_1)}{x_1} \leq \frac{\psi_{\gamma_t}(x_2)}{x_2}, \quad (68)$$

which follows from the fact that $x_2 \leq x_1$ for $s \geq 0$ and $1 \geq c \geq 0$. Furthermore, by Lemma E.4, the function $\frac{\partial \psi_{\gamma_t}(x)}{\partial x}$ is non-increasing in x . Thus, since $x_2 \leq x_1$, it follows that

$$\frac{\partial \psi_{\gamma_t}(x_1)}{\partial x_1} \leq \frac{\partial \psi_{\gamma_t}(x_2)}{\partial x_2}. \quad (69)$$

Substituting the upper bounds from (68) and (69) into (67), we obtain

$$\begin{aligned} \frac{\partial f(s, c, \gamma_t)}{\partial c} \leq & \frac{\psi_{\gamma_t}(x_2)}{x_2} \left(\frac{s^2(s+c)}{1+s^2+2sc} - \frac{s^2(s-c)}{1+s^2-2sc} \right) + \frac{\partial \psi_{\gamma_t}(x_2)}{\partial x_2} \left(\frac{s(1+sc)}{1+s^2+2sc} - \frac{s(1-sc)}{1+s^2-2sc} \right) \\ = & \frac{\psi_{\gamma_t}(x_2)}{x_2} \frac{2cs^2(1-s^2)}{(1+s^2-2sc)(1+s^2+2sc)} - \frac{\partial \psi_{\gamma_t}(x_2)}{\partial x_2} \frac{2cs^2(1-s^2)}{(1+s^2-2sc)(1+s^2+2sc)}. \end{aligned} \quad (70)$$

The RHS of the upper bound in (70) is non-positive for $s \geq 1$ and $1 \geq c \geq 0$, since $\frac{\partial \psi_{\gamma_t}(x_2)}{\partial x_2} \leq \frac{\psi_{\gamma_t}(x_2)}{x_2}$ by Lemma E.5. \square

Lemma E.7. For any $r \geq \|\mathbf{g}_t\|$, the following inequality holds:

$$\mathbb{E}_{s,c} \left[f(s, c, \gamma_t) \middle| 0 \leq c \leq 1 \right] \geq f\left(\frac{r}{\|\mathbf{g}_t\|}, c = 1, \gamma_t\right) \Pr \left\{ s \leq \frac{r}{\|\mathbf{g}_t\|} \right\}. \quad (71)$$

Proof. For any $r \in \mathbb{R}$, we have

$$\begin{aligned} \mathbb{E}_{s,c} \left[f(s, c, \gamma_t) \middle| 0 \leq c \leq 1 \right] &\stackrel{(a)}{\geq} \mathbb{E}_{s,c} \left[f(s, c, \gamma_t) \middle| 0 \leq c \leq 1, s \leq \frac{r}{\|\mathbf{g}_t\|} \right] \Pr \left\{ s \leq \frac{r}{\|\mathbf{g}_t\|} \right\} \\ &\quad + \mathbb{E}_{s,c} \left[f(s, c, \gamma_t) \middle| 0 \leq c \leq 1, s \geq \frac{r}{\|\mathbf{g}_t\|} \right] \Pr \left\{ s \geq \frac{r}{\|\mathbf{g}_t\|} \right\} \end{aligned} \quad (72)$$

$$\stackrel{(b)}{\geq} \mathbb{E}_{s,c} \left[f(s, c, \gamma_t) \middle| 0 \leq c \leq 1, s \leq \frac{r}{\|\mathbf{g}_t\|} \right] \Pr \left\{ s \leq \frac{r}{\|\mathbf{g}_t\|} \right\} \quad (73)$$

$$\stackrel{(c)}{\geq} \mathbb{E}_c \left[f\left(\frac{r}{\|\mathbf{g}_t\|}, c, \gamma_t\right) \middle| 0 \leq c \leq 1 \right] \Pr \left\{ s \leq \frac{r}{\|\mathbf{g}_t\|} \right\}, \quad (74)$$

where (a) follows from conditioning the expectation on the events $s \geq \frac{r}{\|\mathbf{g}_t\|}$ and $s \leq \frac{r}{\|\mathbf{g}_t\|}$, (b) is concluded by dropping the second term in (72) and noting that $f(s, c, \gamma_t) \geq 0$ based on Theorem E.6, and (c) follows from the fact that $f(s, c, \gamma_t)$ is non-increasing in s for any $1 \geq c \geq 0$ and $\gamma_t \geq 0$ due to Theorem E.6.

We now restrict $r \geq \|\mathbf{g}_t\|$ and use the fact from Theorem E.6 that $f(s, c, \gamma_t)$ is non-increasing in c for any $s \geq 1$ to conclude that for any $r \geq \|\mathbf{g}_t\|$,

$$\mathbb{E}_c \left[f\left(\frac{r}{\|\mathbf{g}_t\|}, c, \gamma_t\right) \middle| 0 \leq c \leq 1 \right] \Pr \left\{ s \leq \frac{r}{\|\mathbf{g}_t\|} \right\} \geq f\left(\frac{r}{\|\mathbf{g}_t\|}, c = 1, \gamma_t\right) \Pr \left\{ s \leq \frac{r}{\|\mathbf{g}_t\|} \right\}, \quad (75)$$

which completes the proof. \square

Lemma E.8. For $s \geq 1$, $f(s, c = 1, \gamma_t)$ is lower bounded as

$$f(s, c = 1, \gamma_t) \geq \frac{\gamma_t}{\cosh(s\gamma_t)}. \quad (76)$$

Proof. By substituting $c = 1$ into the definition of $f(s, c, \gamma_t)$ given in Lemma E.2 and simplifying the expression, we obtain

$$\begin{aligned} f(s, c = 1, \gamma_t) &= \frac{1+s}{\sqrt{1+s^2+2s}} \left(\frac{2}{1+e^{-\gamma_t\sqrt{1+s^2+2s}}} - 1 \right) + \frac{1-s}{\sqrt{1+s^2-2s}} \left(\frac{2}{1+e^{-\gamma_t\sqrt{1+s^2-2s}}} - 1 \right) \end{aligned} \quad (77)$$

$$= \frac{2}{1+e^{-\gamma_t(s+1)}} - \frac{2}{1+e^{-\gamma_t(s-1)}} \quad (78)$$

$$= \frac{2\sinh(\gamma_t)}{\cosh(\gamma_t) + \cosh(s\gamma_t)}. \quad (79)$$

We next note that $s \geq 1$. Using the fact that $\cosh(x)$ is an increasing function on $x \geq 0$, we can conclude that $\cosh(s\gamma_t) \geq \cosh(\gamma_t)$. We then use the lower bound $\sinh(x) \geq x$ for $x \geq 0$ to write

$$\frac{2\sinh(\gamma_t)}{\cosh(\gamma_t) + \cosh(s\gamma_t)} \geq \frac{\gamma_t}{\cosh(s\gamma_t)}, \quad (80)$$

which completes the proof. \square

Lemma E.9. Let $\lambda_\alpha = k_1/T$ and $\alpha_0 = k_2/r$ for some $k_1 > 0$ and $k_2 > 0$. Under Algorithm I the saturation slope is bounded on both sides as

$$\frac{\kappa_1}{r} \leq \alpha_t \leq \frac{\kappa_2}{r}, \quad 0 \leq t \leq T-1, \quad (81)$$

where $\kappa_1 \triangleq k_2 e^{-k_1}$ and $\kappa_2 \triangleq k_2 e^{k_1}$.

Proof. First, observe that for $t = 0$ we have $\alpha_0 = \frac{k_2}{r}$, which satisfies

$$\frac{\kappa_1}{r} \leq \alpha_0 \leq \frac{\kappa_2}{r},$$

since $k_1 > 0$. We next show that the same bound also holds for $1 \leq t \leq T-1$. Substituting $\lambda_\alpha = \frac{k_1}{T}$ and $\alpha_0 = \frac{k_2}{r}$ into the update rule (I6), we obtain for any $t \geq 1$,

$$\alpha_t = \frac{k_2}{r} e^{\frac{k_1 q_t}{T}}, \quad (82)$$

where q_t is defined as

$$q_t \triangleq \sum_{\tau=0}^{t-1} \text{sign}(\hat{s}(\boldsymbol{\theta}_\tau, \alpha_\tau; \mathcal{B}_\tau)^\top \hat{r}(\boldsymbol{\theta}_{\tau-1}, \alpha_{\tau-1}; \mathcal{B}_{\tau-1})), \quad (83)$$

with $\hat{r}(\boldsymbol{\theta}_{-1}, \alpha_{-1}; \mathcal{B}_{-1}) = \mathbf{0}$. Since $-t \leq q_t \leq t$ and $1 \leq t \leq T-1$, we have

$$\alpha_t \geq \frac{k_2}{r} e^{-\frac{k_1(T-1)}{T}} \geq \frac{k_2}{r} e^{-k_1} = \frac{\kappa_1}{r} \quad (84)$$

$$\alpha_t \leq \frac{k_2}{r} e^{\frac{k_1(T-1)}{T}} \leq \frac{k_2}{r} e^{k_1} = \frac{\kappa_2}{r}, \quad (85)$$

which completes the proof. \square

Theorem E.10. Let Assumptions 7.1–7.4 hold, and suppose that $\lambda_\alpha \propto 1/T$ and $\alpha_0 \propto 1/r$ for some constant $r \geq G$. Then, under the DP-SGD-AdaSig algorithm described in Algorithm 7 the following inequality holds:

$$\frac{1}{T} \sum_{t=0}^{T-1} \Pr \left\{ \|\Delta_t\| \leq r \right\} \mathbb{E} \|\mathbf{g}_t\|^2 \leq r \tilde{G} \left(\frac{L(\boldsymbol{\theta}_0) - L^*}{\lambda T} + \beta \lambda \left(1 + \frac{d\sigma_s^2}{2B^2} \right) \right), \quad (86)$$

where $\tilde{G} > 0$ is a constant.

Proof. Using the results of Lemmas E.1, E.2, E.7, E.8 and E.9, we have

$$\begin{aligned} & \frac{1}{\lambda} \mathbb{E} \left[L(\boldsymbol{\theta}_t) - L(\boldsymbol{\theta}_{t+1}) \middle| \boldsymbol{\theta}_t, \alpha_t \right] + \beta \lambda \left(1 + \frac{d\sigma_s^2}{2B^2} \right) \\ & \stackrel{\text{Lemma E.1}}{\geq} \mathbf{g}_t^\top \mathbb{E} \left[\left(\frac{2}{1 + e^{-\alpha_t \|\mathbf{v}_t\|}} - 1 \right) \frac{\mathbf{v}_t}{\|\mathbf{v}_t\|} \middle| \boldsymbol{\theta}_t, \alpha_t \right] \end{aligned} \quad (87)$$

$$\stackrel{\text{Lemma E.2}}{=} \frac{\|\mathbf{g}_t\|}{2} \mathbb{E}_{s,c} \left[f(s, c, \gamma_t) \middle| 0 \leq c \leq 1 \right] \quad (88)$$

$$\stackrel{(a)}{\geq} \frac{\|\mathbf{g}_t\|}{2} f\left(\frac{r}{\|\mathbf{g}_t\|}, c = 1, \gamma_t\right) \Pr \left\{ \|\Delta_t\| \leq r \right\} \quad (89)$$

$$\stackrel{(b)}{\geq} \frac{\alpha_t \|\mathbf{g}_t\|^2}{2 \cosh(r\alpha_t)} \Pr \left\{ \|\Delta_t\| \leq r \right\} \quad (90)$$

$$\stackrel{\text{Lemma E.9}}{\geq} \frac{\kappa_1 \|\mathbf{g}_t\|^2}{2r \cosh(\kappa_2)} \Pr \left\{ \|\Delta_t\| \leq r \right\}, \quad (91)$$

where (a) is obtained by applying Lemma E.7 which requires the bound $r \geq \|\mathbf{g}_t\|, \forall t$. This bound follows directly from Assumption 7.3 (i.e., $\|\mathbf{g}_t\| \leq G, \forall t$) together with the assumption $r \geq G$. Inequality (b) is obtained by applying Lemma E.8, using the bound $r \geq \|\mathbf{g}_t\|, \forall t$. The last inequality follows from Lemma E.9 since under the assumptions $\lambda_\alpha \propto 1/T$ and $\alpha_0 \propto 1/r$, we can equivalently write $\lambda_\alpha = k_1/T$ and $\alpha_0 = k_2/r$ for some constants $k_1 > 0$ and $k_2 > 0$. Substituting these expressions into the lemma, the inequality holds with $\kappa_1 = k_2 e^{-k_1}$, $\kappa_2 = k_2 e^{k_1}$.

Let us now define $\tilde{G} \triangleq 2 \cosh(\kappa_2)/\kappa_1$. After taking expectation from both sides of the last inequality, summing over iterations from 0 to $T-1$, and dividing by T , we have

$$\begin{aligned} \frac{1}{r\tilde{G}T} \sum_{t=0}^{T-1} \Pr\{\|\Delta_t\| \leq r\} \mathbb{E}\|\mathbf{g}_t\|^2 \\ \leq \frac{1}{\lambda T} \sum_{t=0}^{T-1} \mathbb{E}[L(\boldsymbol{\theta}_t) - L(\boldsymbol{\theta}_{t+1})] + \frac{1}{T} \sum_{t=0}^{T-1} \beta\lambda \left(1 + \frac{d\sigma_s^2}{2B^2}\right) \end{aligned} \quad (92)$$

$$\stackrel{(a)}{\leq} \frac{L(\boldsymbol{\theta}_0) - L^*}{\lambda T} + \beta\lambda \left(1 + \frac{d\sigma_s^2}{2B^2}\right), \quad (93)$$

where (a) follows from Assumption 7.1. Rearranging the terms yields the inequality in (86). \square

E.2 Proof of Theorem 7.5

Proof. We use the result in Theorem E.10 and set the noise multiplier σ_s such that privacy is guaranteed. To ensure the privacy guarantee, we use the result in Theorem 1 in [1]. For clarity, we first restate Theorem 1 in [1].

Theorem E.11 (Theorem 1 of [1]). *There exist constants u and ν such that, given the sampling probability $q = B/N$, for any $\epsilon \leq uq^2T$, the composition of T Gaussian mechanisms, each with noise multiplier σ , satisfies (ϵ, δ) -DP if*

$$\sigma^2 \geq \frac{\nu^2 q^2 T \log(1/\delta)}{\epsilon^2}. \quad (94)$$

According to Theorem E.11, the composition of T Gaussian mechanisms ensures (ϵ, δ) -DP provided that the noise multiplier σ is set to

$$\sigma^2 = \frac{\nu^2 B^2 T \log(1/\delta)}{N^2 \epsilon^2}. \quad (95)$$

Note that our proposed algorithm (DP-SGD-AdaSig, described in Algorithm 1) requires two noise multipliers, σ_s and σ_r , since each iteration involves two Gaussian mechanisms, $\hat{s}(\cdot)$ and $\hat{r}(\cdot)$. According to Proposition 6.1, these two parallel Gaussian mechanisms are equivalent to a single Gaussian mechanism with noise multiplier σ , where

$$\frac{1}{\sigma^2} = \frac{1}{\sigma_s^2} + \frac{1}{\sigma_r^2}. \quad (96)$$

To ensure (ϵ, δ) -DP for the overall algorithm after T iterations, we set σ_s and σ_r such that the equivalent noise multiplier σ in (96) satisfies (95). This can be achieved by setting σ_s and σ_r as follows:

$$\sigma_s^2 = \frac{\nu_s^2 B^2 T \log(1/\delta)}{N^2 \epsilon^2}, \quad (97)$$

$$\sigma_r^2 = \frac{\nu_r^2 B^2 T \log(1/\delta)}{N^2 \epsilon^2}, \quad (98)$$

with constants ν_s and ν_r satisfying

$$\frac{1}{\nu^2} = \frac{1}{\nu_s^2} + \frac{1}{\nu_r^2}. \quad (99)$$

The choices of σ_s and σ_r in (97) and (98), together with constants ν_s and ν_r that satisfy the equality in (99), ensure that (96) holds and thereby guarantee (ϵ, δ) -DP for the DP-SGD-AdaSig algorithm.

Substituting σ_s^2 from (97) into (86) in Theorem E.10, we then optimize its LHS w.r.t. the learning rate λ , which yields

$$\lambda = \sqrt{\frac{2N^2 \epsilon^2 (L(\boldsymbol{\theta}_0) - L^*)}{\beta T (2N^2 \epsilon^2 + d\nu_s^2 T \log(\frac{1}{\delta}))}}. \quad (100)$$

Substituting this learning rate into the LHS concludes the result. \square

E.3 Proof of Corollary 7.7

Proof. The probability $\Pr \left\{ \|\Delta_t\| \leq r \right\}$ can be lower bounded as

$$\Pr \left\{ \|\Delta_t\| \leq r \right\} \stackrel{(a)}{\geq} 1 - \frac{\mathbb{E} \|\Delta_t\|}{r} \quad (101)$$

$$\stackrel{(b)}{\geq} 1 - \frac{\sqrt{\mathbb{E} \|\Delta_t\|^2}}{r} \quad (102)$$

$$\stackrel{(c)}{\geq} 1 - \frac{\zeta}{r}, \quad (103)$$

where (a) follows from Markov's inequality, (b) follows from Jensen's inequality, and (c) is due to the bounded variance assumption.

Using the derived probability lower bound in (103) on the LHS of (18) in Theorem 7.5, and considering $r \geq \max\{\zeta, G\}$, we divide both sides by $1 - \frac{\zeta}{r}$ to obtain:

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|\mathbf{g}_t\|^2 \leq \frac{2\tilde{G}r^2}{r - \zeta} \sqrt{\frac{\beta(L(\theta_0) - L^*)}{T} \left(1 + \frac{d\nu_s^2 T \log(\frac{1}{\delta})}{2N^2\epsilon^2}\right)}. \quad (104)$$

We next set r such that the RHS of the above inequality is minimized. Note that when $r \geq \zeta$, the function $\frac{r^2}{r - \zeta}$ achieves its minimum value at $r = 2\zeta$, and for $r \geq 2\zeta$, the function is increasing. Noting that in Theorem 7.5 $r \geq G$, two cases may occur: namely, $G \leq 2\zeta$ or $G \geq 2\zeta$. By considering these two cases, separately, we can conclude that

$$r_* = \underset{r \geq \max\{\zeta, G\}}{\operatorname{argmin}} \frac{r^2}{r - \zeta} = \begin{cases} 2\zeta, & \text{if } G \leq 2\zeta, \\ G, & \text{if } G \geq 2\zeta \end{cases} = \max\{2\zeta, G\}. \quad (105)$$

Substituting r_* into (104) completes the proof. \square

F Experiments Settings

This section provides further information on the experimental settings considered in the paper. We present more details on the datasets and architectures used for the numerical experiments presented in the paper.

F.1 General Settings

Privacy Settings In all experiments, we fix the DP budget (ϵ, δ) , and compute the noise multiplier, i.e., σ , numerically using the Opacus library [40], such that the DP budget spent after T iterations (or equivalently Tq epochs) equals to the fixed budget (ϵ, δ) . Considering Proposition 6.1, for a given σ , different pairs of σ_s and σ_r can be obtained, where decreasing one increases the other. To simplify the parameter space, we set $\sigma_s = 1.01\sigma$ in all AdaSig experiments, making σ_s smaller than σ_r . This choice prioritizes a more accurate update of the model over the saturation slope, as roughly updating α in the descent direction is sufficient.

Hyperparameters To ensure fair comparison, we use the same batch size and number of epochs across all methods. For AdaSig, we tune the learning rates λ , λ_α , and the initial saturation slope α_0 . For vanilla clipping, we tune C and λ . For Auto-S and PSAC, we use the tuned value of C for vanilla clipping following their original papers [4, 36], and tune hyperparameters r and learning rate λ . For the method in [3], we tune the initial clipping threshold C^0 , clipping threshold learning rate η_C , learning rate λ , σ_b , and γ .

F.2 Settings for Image Classification

MNIST and FashionMNIST We use the 4-layer CNN with tanh activation proposed in [27] and described in Table 6 of [34], with cross-entropy loss and the DP-SGD optimizer. The DP budget is set to $(\epsilon, \delta) = (3, 10^{-5})$. For batch size, number of epochs, and momentum, which are common across different methods, we use the values reported in [34]. These values are summarized in Table 2. Table 3 presents the best values of other hyperparameters for each method.

Table 2: Common hyperparameters across different methods used for training CNN on MNIST, FashionMNIST, and CIFAR-10.

Parameter	MNIST	FashionMNIST	CIFAR-10
Batch size (B)	512	2048	4096
Number of epochs (Tq)	40	40	60
Momentum	0.9	0.9	0.9

Table 3: Hyperparameters selected for each method for training CNN on MNIST, FashionMNIST, and CIFAR-10.

Method	Parameter	MNIST	FashionMNIST	CIFAR-10
Vanilla	C	0.1	0.1	0.1
	λ	0.5	4.0	3.0
Auto-S	r	0.01	0.01	0.01
	C	0.1	0.1	0.1
	λ	0.5	4.0	3.0
PSAC	r	0.1	0.1	0.1
	C	0.1	0.1	0.1
	λ	0.5	4.0	3.0
Method in [3]	C^0	0.1	0.1	0.1
	λ	0.5	4.0	3.0
	η_C	0.05	0.01	0.01
	σ_b	40.0	30.0	25.0
	γ	0.5	0.5	1.0
AdaSig	α_0	5.0	1.0	1.0
	λ	0.05	0.4	0.25
	λ_α	0.01	0.01	0.01

CIFAR-10 We use the 8-layer CNN with tanh activation from [27], as detailed in Table 7 of [34], with cross-entropy loss and DP-SGD optimizer. The DP budget is set to $(\epsilon, \delta) = (3, 10^{-5})$. We use the same batch size, number of epochs, and momentum across different methods as reported in Table 2⁴. The best values of other hyperparameters for different methods are reported in Table 3.

ImageNette We use ImageNette, a 10-class subset of ImageNet [9], with an image size of 160×160 . We consider ResNet-9 architecture (about 2.5 million parameters), with the Mish activation function [25] and cross-entropy loss. For training, the DP-Nesterov-accelerated Adam (DP-NAdam) optimizer is utilized.⁵ We consider a DP budget of $(\epsilon, \delta) = (8, 10^{-4})$. We follow [17] by using group normalization instead of batch normalization without scale normalization. The only difference is that we do not apply the learning rate decay schedule. The ResNet-9 architecture can be found at [17].⁶ All methods use the same batch size and number of epochs, which are set according to the values reported in [4, 36], as shown in Table 4. Table 5 lists the best values of the remaining hyperparameters for each method.

CelebA We use the same ResNet-9 architecture as for the ImageNette dataset, with the DP-Adam optimizer. The CelebA dataset contains 40 labels per image and we use it for both single-label and multi-label classification tasks.

⁴Note that the batch size and the number of epochs differ from those reported in [34] and used in [4, 36]. Our experiments showed that these hyperparameter changes yield better performance across all baselines compared with those reported in the references.

⁵For ImageNette and CelebA, we use more advanced optimizers than DP-SGD, namely, DP-NAdam and DP-Adam, respectively, to achieve improved performance for these more challenging datasets. These experiments further demonstrate the wide applicability of the AdaSig approach with different optimizers.

⁶Check also <https://gist.github.com/gkaissis/6db6b7271f93d3459263b6978cfd4146>.

Table 4: Common hyperparameters across different methods used for training Resnet-9 on ImageNette and CelebA.

Parameter	ImageNette	CelebA (Multi-label/Smiling)
Batch size (B)	1024	512
Number of epochs (Tq)	50	10

Table 5: Hyperparameters selected for each method for training ResNet-9 on ImageNette, and CelebA.

Method	Parameter	ImageNette	CelebA(Multi-label)	CelebA (Smiling)
Vanilla	C	1.5	0.1	0.1
	λ	5×10^{-4}	10^{-3}	10^{-3}
Auto-S	r	0.01	0.01	0.01
	C	1.5	0.1	0.1
	λ	5×10^{-4}	10^{-3}	10^{-3}
PSAC	r	0.1	0.1	0.1
	C	1.5	0.1	0.1
	λ	5×10^{-4}	10^{-3}	10^{-3}
Method in [3]	C^0	1.5	0.1	0.1
	λ	5×10^{-4}	5×10^{-4}	2×10^{-3}
	η_C	0.01	0.01	0.05
	σ_b	30.0	20.0	25.0
	γ	3.0	1.0	1.0
AdaSig	α_0	5.0	1.0	1.0
	λ	10^{-3}	5×10^{-4}	5×10^{-4}
	λ_α	0.02	0.001	0.02

- For the single-label task, we perform binary classification considering the label 'Smiling', where we use the binary cross-entropy loss for training. In this scenario, the output layer of ResNet-9 consists of a single neuron.
- For the multi-label classification task, all available 40 labels are considered for prediction. Here, the output layer contains 40 neurons, and we use a scalar loss function that averages the 40 binary cross-entropy losses from each label.

In both cases, we set the DP budget to $(\epsilon, \delta) = (8, 5 \times 10^{-6})$.

The batch size and the number of epochs are set the same across all methods, following the settings in [4, 36], and are reported in Table 4. The best values of other hyperparameters for each method are presented in Table 5.

F.3 Setting for Sentence Classification

SST-2 and QNLI We use a pre-trained RoBERTa-base model (about 125 million parameters) [22] and perform full parameter fine-tuning. We use the cross-entropy loss function. The AdamW optimizer [24] without weight decay is applied. A learning rate scheduler is used to linearly reduce the learning rate from its initial value to zero throughout the training process. The DP budget is set to $(\epsilon, \delta) = (3, \frac{1}{2N})$, where N is the number of training samples for each dataset. It is worth mentioning that SST-2 and QNLI contain 67,349 and 104,743 training samples, respectively.

The batch size, number of epochs, and maximum sequence length for all methods are set according to [21, 4], and are given in Table 6. Table 7 lists the best value of other hyperparameters for each method.

Table 6: Common hyperparameters across different methods used for RoBERTa-base full parameter fine-tuning on SST-2 and QNLI.

Parameter	SST-2	QNLI
Batch size (B)	1000	2000
Number of epochs (Tq)	3	6
Max sequence length	256	256

Table 7: Hyperparameters selected for each method for fine-tuning RoBERTAa-base on SST-2 and QNLI.

Method	Parameter	SST-2	QNLI
Vanilla	C	0.1	0.1
	λ	5×10^{-4}	5×10^{-4}
Auto-S	r	0.01	0.01
	C	0.1	0.1
	λ	5×10^{-4}	5×10^{-4}
PSAC	r	0.1	0.1
	C	0.1	0.1
	λ	5×10^{-4}	5×10^{-4}
Method in [3]	C^0	0.1	0.1
	λ	5×10^{-4}	5×10^{-4}
	η_C	0.01	0.01
	σ_b	35.00	25.0
	γ	1.0	1.0
AdaSig	α_0	1.0	1.0
	λ	5×10^{-4}	5×10^{-4}
	λ_α	0.005	0.01

F.4 Hardware and Software Information

All experiments are performed on a server equipped with Intel Xeon E5-2683 v4 CPUs, NVIDIA V100 GPUs, and 251 GiB of memory. The operating system used is AlmaLinux 9.3, and the CUDA Toolkit version is 12.2. The implementation of all training procedures is based on PyTorch 2.3.0 and Opacus 1.4.1.

G Numerical Analysis of Direction Deviation and Magnitude Deviation

In this section, we provide additional figures to further assess the direction deviation and magnitude deviation achieved by the proposed AdaSig clipping method and the baselines.

G.1 Defining Metrics

Cosine Similarity To evaluate the direction deviation, we compute the cosine similarity between the aggregation of the clipped gradients, i.e., $\tilde{\mathbf{g}}_t \triangleq \sum_{i \in \mathcal{B}_t} \tilde{\mathbf{g}}_{i,t}$, and the true batch gradient, i.e., $\mathbf{g}_t \triangleq \sum_{i \in \mathcal{B}_t} \mathbf{g}_{i,t}$, as $\cos \phi_t = \frac{\langle \tilde{\mathbf{g}}_t, \mathbf{g}_t \rangle}{\|\tilde{\mathbf{g}}_t\| \cdot \|\mathbf{g}_t\|}$, which measures the cosine of the angle between $\tilde{\mathbf{g}}_t$ and \mathbf{g}_t .

SNR To characterize the magnitude deviation, we define the signal-to-noise-ratio (SNR) as a normalized magnitude of the aggregation of the clipped gradients. Let us denote by $\|\tilde{\mathbf{g}}_t\|$ the magnitude of the aggregation of the clipped gradients after clipping, i.e., $\|\tilde{\mathbf{g}}_t\| = \left\| \sum_{i \in \mathcal{B}_t} \tilde{\mathbf{g}}_{i,t} \right\|$. We define the SNR to be the ratio of this magnitude to the standard deviation of the added privacy noise denoted by σ_n , i.e., $\text{SNR} = \frac{\|\tilde{\mathbf{g}}_t\|}{\sigma_n}$. For AdaSig, $\sigma_n = \sigma_s$. For vanilla clipping, Auto-S, and PSAC,

$\sigma_n = C\sigma$. For the method in [3], $\sigma_n = C^t z_\Delta$. This metric can describe the magnitude deviation against various approaches with potentially different privacy noise variances: as the SNR decreases, the performance degradation caused by the privacy noise becomes more severe.

G.2 Performance Comparison

ImageNette Figure 4 shows the histogram of the SNR during training on the ImageNette dataset. The results depict that using AdaSig, the SNR is concentrated around larger values. We further recall the earlier observations reported in Figure 3 in the main paper. There, we observe that using AdaSig, direction deviation is reduced compared with the baselines. Considering these two results, i.e., Figure 3 and Figure 4, we conclude that using AdaSig, the deteriorating impact of clipping is reduced with respect to *both* direction deviation and magnitude deviation. This demonstrates that, for ImageNette, AdaSig provides highly effective clipping that preserves closeness to the true batch gradient.

CIFAR-10 We next present the results of similar experiments on CIFAR-10. Figures 5 and 6 show the histogram of cosine similarity and SNR while training the 8-layer CNN on CIFAR-10, respectively. As observed in these figures, the cosine similarity distribution of AdaSig is skewed toward higher values relative to the baselines. This means that AdaSig incurs less direction deviation. Nevertheless, the histogram of the SNR achieved by AdaSig is close to those of the baselines. Thus, for CIFAR-10, AdaSig clipping is still able to improve the trade-off between the direction deviation and the magnitude deviation.

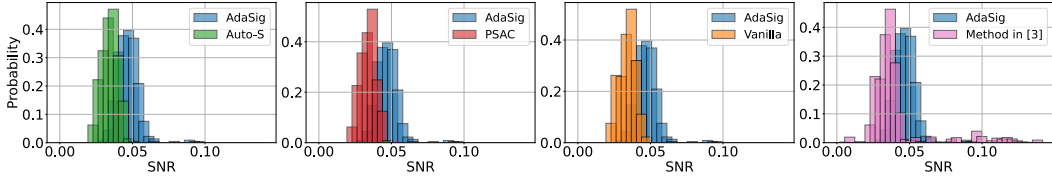


Figure 4: Comparison of normalized magnitude for different approaches during training ResNet-9 on ImageNette.

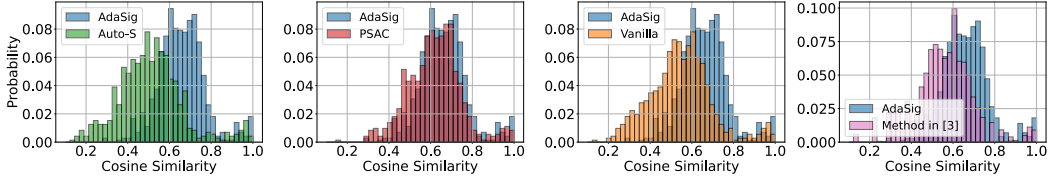


Figure 5: Comparison of cosine similarity for different approaches during training 8-layer CNN on CIFAR-10.

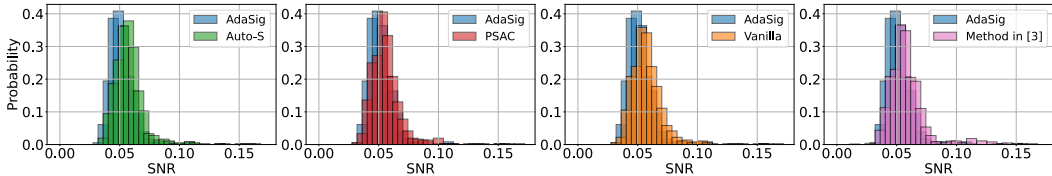


Figure 6: Comparison of normalized magnitude for different approaches during training 8-layer CNN on CIFAR-10.

H Ablation Study: Sigmoid Clipping with Constant α

An ablation method for the AdaSig approach is to use a fixed α across all iterations, i.e., $\alpha_t = \alpha, \forall t$. We examine sigmoid clipping with various fixed α values for training an 8-layer CNN on the CIFAR-10 dataset. The experimental setup is based on Appendix F.2.

H.1 Impact of α on Direction Deviation and Magnitude Deviation

In this section, we study the impact of α on the *trade-off* between direction deviation and magnitude deviation. Figure 7 shows the histogram of cosine similarity and SNR during the training for three fixed values of α . As seen, decreasing the α value results in an increase in cosine similarity, while reducing the SNR. This occurs because decreasing α expands the linear region of the sigmoid function, providing equal scaling for different gradient samples during clipping (see Figure 1 in the main paper), which reduces the direction deviation. This observation also aligns with the numerical example presented in Section 4 of the main paper.

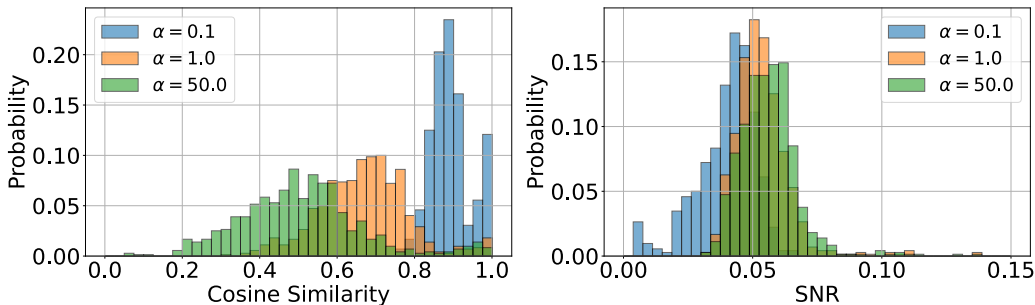


Figure 7: Trade-off between direction deviation and magnitude deviation for several fixed α values during the training of an 8-layer CNN on CIFAR-10. left: Cosine similarity histogram, right: SNR histogram.

H.2 Impact of α on Test Accuracy

Next, we present the average final test accuracies for sigmoid clipping with several fixed α values in Table 8. As observed, very low values of α (less than 0.1) result in low accuracy due to a low SNR, while very large α values (greater than 5) also degrade accuracy due to direction deviation. However, for α in the middle range, we achieve the highest accuracy, resulting from a balance between direction deviation and magnitude deviation. It is worth noting that the highest accuracy achieved by $\alpha = 1$ is still lower than the accuracy attained by the AdaSig method, as shown in Table 1 in the main paper, which highlights that adaptively adjusting α over iterations further improves accuracy.

Table 8: Average test accuracies with 95% confidence intervals over five runs for sigmoid clipping with various fixed α values.

	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 1.0$	$\alpha = 5.0$	$\alpha = 50.0$	$\alpha = 100.0$
Average test accuracy (%)	45.27 ± 0.67	61.74 ± 0.21	62.13 ± 0.12	61.64 ± 0.27	60.62 ± 0.38	60.39 ± 0.31

I Variation of α_t During Training

In this section, we illustrate how α_t evolves during training on CIFAR-10 for three different random seeds. As shown in Figure 8, the general trend across seeds is that α_t initially increases from its starting value 1.0 up to a peak in the middle of training, and then gradually decreases toward the end. The exact trajectory varies across seeds since each seed corresponds to a different batch sampling process, leading to distinct gradient statistics in each run.

When examining α_t over training iterations across different datasets and models, we do not observe a consistent pattern. This behavior arises from the dependence of α_t on individual sample gradients, which vary substantially across datasets and model architectures.

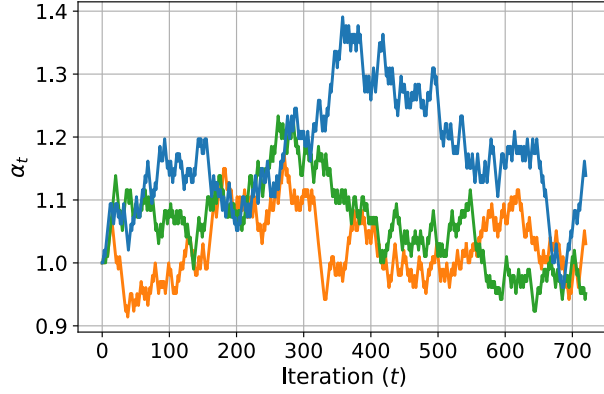


Figure 8: Evolution of α_t across training iterations on CIFAR-10 for three random seeds.

J Convergence Plots

In this section, we present convergence plots for several experiments. Figure 9 illustrates the average test accuracy and test loss over five runs versus epochs during training on the CIFAR-10 dataset. The shaded regions around the curves represent the 95% confidence intervals. As shown, the increasing test accuracy in the left plot and the decreasing test loss in the right plot demonstrate convergence for all methods, including AdaSig. Additionally, the figure highlights that AdaSig achieves a lower final test loss and higher final test accuracy compared with the baselines.

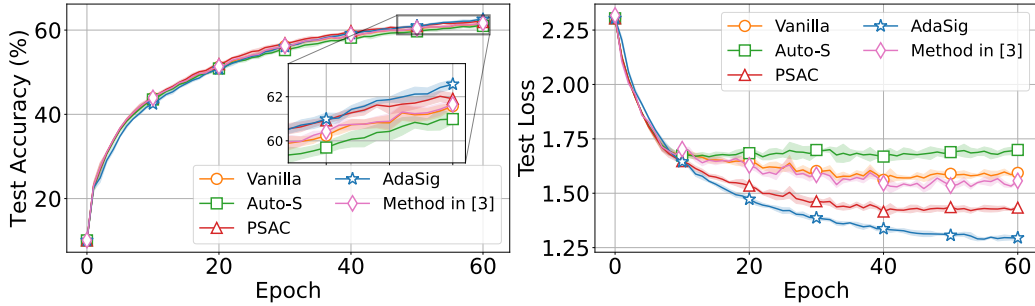


Figure 9: Convergence plot for CIFAR-10 dataset, left: Test accuracy vs. epoch, right: Test loss vs. epoch.

Additionally, Figure 10 shows the average test loss and accuracy throughout the training on the SST-2 dataset. As observed, all methods progressively reduce the test loss while increasing accuracy, with AdaSig outperforming the other methods.

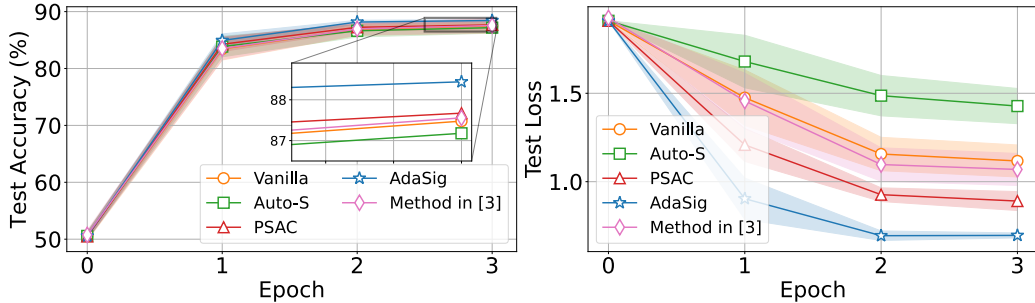


Figure 10: Convergence plot for SST-2 dataset, left: Test accuracy vs. epoch, right: Test loss vs. epoch.