A Proof of differential privacy

Proof of Theorem 3. Define the ℓ_2 **sensitivity** of any function g to be $\Delta g = \sup_{S,S'} \|g(S) - g(S')\|_2$ where the supreme is over all neighboring (S, S'). Then the **Gaussian mechanism** $\hat{g}(S) = g(S) + \sigma \Delta g \cdot \mathcal{N}(0, \mathbf{I})$.

σ denotes the "Noise multiplier", which corresponds to the noise-level when a Gaussian mechanism is applied to a query with sensitivity 1.

Observe that automatic clipping (AUTO-V and AUTO-S (4.1)) ensures the bounded global-sensitivity of the stochastic gradient as in Abadi's clipping. Aligning the noise-multiplier (rather than the noise-level itself) ensures that the the noise-to-sensitivity ratio $\frac{\sigma \Delta g}{\Delta g} = \sigma$ is fixed regardless of Δg. The Gaussian mechanism's privacy guarantees are equivalent. Thus from the privacy accountant perspective, DP-SGD with both Abadi's clipping and our autoclipping method can be equivalently represented as the adaptive composition of T Poisson sampled Gaussian Mechanism with sampling probability B/n and noise multiplier σ. □

B Proof of automaticity

B.1 Non-adaptive DP optimizers

Proof of Theorem 1. We prove Theorem 1 by showing that, DP-SGD using R-dependent AUTO-S with learning rate η and weight decay λ is equivalent to R-independent AUTO-S with learning rate ηR and weight decay λ/R. We claim other non-adaptive optimizers such as HeavyBall and NAG can be easily shown in a similar manner.

Recall the standard SGD with weight decay is

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \left(\sum_{i \in B_t} \frac{\partial l_i}{\partial \boldsymbol{w}_t} + \lambda \boldsymbol{w}_t \right)$$

Replacing the standard gradient $\sum_i \frac{\partial l_i}{\partial \boldsymbol{w}_t}$ with the private gradient, we write the R-dependent case as

$$\begin{aligned} \boldsymbol{w}_{t+1} &= \boldsymbol{w}_t - \eta \left(\sum_{i \in B_t} \frac{\partial l_i}{\partial \boldsymbol{w}_t} \cdot R / \| \frac{\partial l_i}{\partial \boldsymbol{w}_t} \|_2 + \sigma R \cdot \mathcal{N}(0, \mathbf{I}) + \lambda \boldsymbol{w}_t \right) \\ &= \boldsymbol{w}_t - \eta R \left(\sum_{i \in B_t} \frac{\partial l_i}{\partial \boldsymbol{w}_t} / \| \frac{\partial l_i}{\partial \boldsymbol{w}_t} \|_2 + \sigma \cdot \mathcal{N}(0, \mathbf{I}) \right) - \eta \lambda \boldsymbol{w}_t \end{aligned}$$

which is clearly equivalent to the R-independent case:

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta' \left(\sum_{i \in B_t} \frac{\partial l_i}{\partial \boldsymbol{w}_t} / \| \frac{\partial l_i}{\partial \boldsymbol{w}_t} \|_2 + \sigma \cdot \mathcal{N}(0, \mathbf{I}) + \lambda' \boldsymbol{w}_t \right)$$

if we use $\eta' = \eta R$ and $\lambda' = \lambda / R$. □

B.2 Adaptive DP optimizers

Proof of Theorem 2. We prove Theorem 2 by showing that, DP-AdamW using R-dependent AUTO-S with learning rate η and weight decay λ is equivalent to R-independent AUTO-S with the same learning rate η and weight decay λ/R. This is the most complicated case. We claim other adaptive optimizers such as AdaDelta, Adam with weight decay (not AdamW), and NAdam can be easily shown in a similar manner.

Recall the standard AdamW is

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \left(\frac{\mathbf{m}_t / (1 - \beta_1)}{\sqrt{\boldsymbol{v}_t / (1 - \beta_2)}} + \lambda \boldsymbol{w}_t \right)$$

where β_1, β_2 are constants, $\boldsymbol{g}_t := \sum_i \frac{\partial l_i}{\partial \boldsymbol{w}_t}$ is the standard gradient,

$$\mathbf{m}_t = \beta_1 \mathbf{m}_{t-1} + (1-\beta_1)\boldsymbol{g}_t \longrightarrow \mathbf{m}_t = \sum_\tau \beta_1^{t-\tau}(1-\beta_1)\boldsymbol{g}_\tau,$$

$$\boldsymbol{v}_t = \beta_2 \boldsymbol{v}_{t-1} + (1-\beta_2)\boldsymbol{g}_t^2 \longrightarrow \boldsymbol{v}_t = \sum_\tau \beta_2^{t-\tau}(1-\beta_2)\boldsymbol{g}_\tau^2.$$

Replacing the standard gradient with the private gradient $R\tilde{\boldsymbol{g}}_t := R(\sum_i \frac{\partial l_i}{\partial \boldsymbol{w}_t}/\|\frac{\partial l_i}{\partial \boldsymbol{w}_t}\|_2 + \sigma \cdot \mathcal{N}(0, I))$, we write the R-dependent DP-AdamW as

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \left(\frac{\tilde{\mathbf{m}}_t/(1-\beta_1)}{\sqrt{\tilde{\boldsymbol{v}}_t/(1-\beta_2)}} + \lambda \boldsymbol{w}_t \right)$$

where

$$\tilde{\mathbf{m}}_t = \beta_1 \tilde{\mathbf{m}}_{t-1} + (1-\beta_1)R\tilde{\boldsymbol{g}}_t \longrightarrow \tilde{\mathbf{m}}_t = \sum_\tau \beta_1^{t-\tau}(1-\beta_1)R\tilde{\boldsymbol{g}}_\tau,$$

$$\tilde{\boldsymbol{v}}_t = \beta_2 \tilde{\boldsymbol{v}}_{t-1} + (1-\beta_2)R^2\tilde{\boldsymbol{g}}_t^2 \longrightarrow \tilde{\boldsymbol{v}}_t = \sum_\tau \beta_2^{t-\tau}(1-\beta_2)R^2\tilde{\boldsymbol{g}}_\tau^2.$$

Clearly, the R factor in the numerator and denominator of $\frac{\tilde{\mathbf{m}}_t/(1-\beta_1)}{\sqrt{\tilde{\boldsymbol{v}}_t/(1-\beta_2)}}$ cancel each other. Therefore we claim that the R-dependent DP-AdamW is in fact completely independent of R. □

B.3 Automatic per-layer clipping

In some cases, the per-layer clipping is desired, where we use a clipping threshold vector $\boldsymbol{R} = [R_1, \cdots, R_L]$ and each layer uses a different clipping threshold. We claim that DP optimizers under automatic clipping works with the per-layer clipping when $\boldsymbol{R}$ is tuned proportionally, e.g. $\boldsymbol{R} = R \cdot [a_1, \cdots, a_L]$, but not entry-wise (see counter-example in Fact B.1). One special case is the *uniform per-layer clipping* when $R_1 = \cdots = R_L = R/\sqrt{L}$. This is widely applied as only one norm R requires tuning, instead of L norms in $\boldsymbol{R}$, particularly in the case of deep models with hundreds of layers. The corresponding DP-SGD with AUTO-S in (3.3) gives

$$\boldsymbol{w}_{t+1}^{(l)} = \boldsymbol{w}_t^{(l)} - \eta \left(\sum_{i \in B_t} \frac{R}{\sqrt{L}} \frac{\boldsymbol{g}_{t,i}^{(l)}}{||\boldsymbol{g}_{t,i}^{(l)}|| + \gamma} + \sigma R \cdot \mathcal{N}(0, \mathbf{I}) \right)$$

Here the superscript (l) is the layer index. Clearly R couples with the learning rate η and the same analysis as in Theorem 1 follows. The adaptive optimizers can be similarly analyzed from Theorem 2.

Fact B.1. Changing one clipping threshold in the clipping threshold vector $\boldsymbol{R}$ (i.e. not proportionally) can break the coupling with learning rate.

Proof of Fact B.1. We prove by a counter-example of $\boldsymbol{R}$ in $\mathbb{R}^2$. Consider DP-SGD with per-layer clipping thresholds $(R_1, R_2) = (9, 12)$:

$$\boldsymbol{w}_{t+1}^{(l)} = \boldsymbol{w}_t^{(l)} - \eta \left(\sum_{i \in B} \frac{R_l \boldsymbol{g}_{t,i,l}}{||\boldsymbol{g}_{t,i,l}||} + \sigma \sqrt{R_1^2 + R_2^2} \cdot \mathcal{N}(0, \mathbf{I}) \right)$$

Increasing R_1 from 9 to 16 changes the update for the first layer

$$\eta \left(\sum_{i \in B} \frac{9\boldsymbol{g}_{t,i,l}}{||\boldsymbol{g}_{t,i,l}||} + 15\sigma \cdot \mathcal{N}(0, 1) \right) \to \eta \left(\sum_{i \in B} \frac{16\boldsymbol{g}_{t,i,l}}{||\boldsymbol{g}_{t,i,l}||} + 20\sigma \cdot \mathcal{N}(0, \mathbf{I}) \right)$$

The noise-to-signal ratio decreases from 5/3 to 5/4 for this layer, and increases from 5/4 to 5/3 for the second layer. This breaks the coupling with learning rate, since the coupling does not change the noise-to-signal ratio. □

C Main results of convergence for DP-SGD with automatic clipping

C.1 Main proof of convergence for DP-SGD (the envelope version)

Proof of Theorem 4. In this section, we prove two parts of Theorem 4.

The first part of Theorem 4 is the upper bound on $\min_t \mathbb{E}(\|\boldsymbol{g}_t\|)$, which is a direct result following from Theorem 6, and we prove it in Appendix C.2.

Theorem 6. *Under Assumption 5.1, 5.2, 5.3, running DP-SGD with automatic clipping for T iterations gives*

$$\min_t \mathbb{E}(\|\boldsymbol{g}_t\|) \leq \frac{\xi}{r} + \mathcal{F}\left(\frac{4}{\sqrt{T}}\sqrt{(\mathcal{L}_0 - \mathcal{L}_*)L\left(1 + \frac{\sigma^2 d}{B^2}\right)}; r, \xi, \gamma\right) \tag{C.1}$$

where

- *for* $r < 1, \gamma = 0$ *and* $\eta \propto 1/\sqrt{T}$, $\mathcal{F}(x) = \frac{x}{\min_{0<c<1} f(c,r)}$ *and* $f(c,r) := \frac{(1+rc)}{\sqrt{r^2+2rc+1}} + \frac{(1-rc)}{\sqrt{r^2-2rc+1}}$*; for* $r \geq 1, \gamma = 0$ *and* $\eta \propto 1/\sqrt{T}$, $\mathcal{F}(x) = \infty$*;*
- *for* $r \geq 1, \gamma > 0$ *and* $\eta \propto 1/\sqrt{T}$, $\mathcal{F}$ *is the convex envelope of* (C.8)*, and is strictly increasing.*

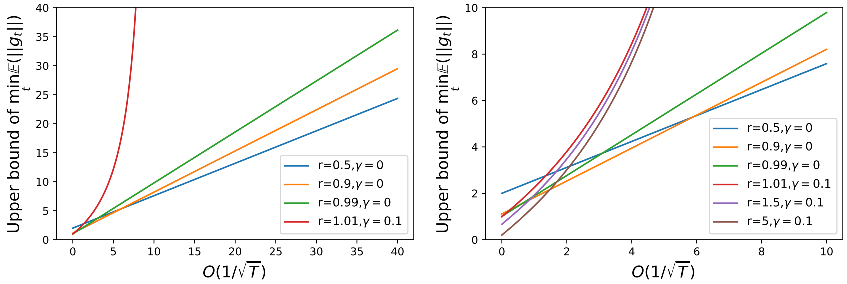

Figure 6: Visualization of upper bound $\frac{\xi}{r} + \mathcal{F}\left(O(1/\sqrt{T}); r, \xi, \gamma\right)$ for gradient norm, with $O(1/\sqrt{T})$ in (C.1). Here $\xi = 1$. The right plot is a zoom-in (with additional lines) of the left one.

Notice that, (C.1) holds for any $r > 0$. However, we have to consider an envelope curve over r in (C.1) to reduce the upper bound: with AUTO-V clipping ($\gamma = 0$), the upper bound in (C.1) is always larger than ξ as $r < 1$; we must use AUTO-S clipping ($\gamma > 0$) to reduce the upper bound to zero, as can be seen from Figure 6. In fact, larger T needs larger r to reduce the upper bound.

All in all, we specifically focus on $r \geq 1$ and $\gamma > 0$, which is the only scenario that (C.1) can converge to zero. This scenario is also where we prove the second part of Theorem 4.

The second part of Theorem 4 is the asymptotic convergence rate $O(T^{-1/4})$ of DP-SGD, only possible under $r \geq 1$ and $\gamma > 0$.

By (C.1) in Theorem 6, our upper bound $\mathcal{G}$ from Theorem 4 can be simplified to

$$\min_{r>0} \frac{\xi}{r} + (\mathcal{M}^{-1})_{ccv}\left(\frac{4}{\sqrt{T}}\sqrt{(\mathcal{L}_0 - \mathcal{L}_*)L\left(1 + \frac{\sigma^2 d}{B^2}\right)}; r, \xi, \gamma\right)$$

where the function $\mathcal{M}^{-1}$ is explicitly defined in (C.8) and the subscript ccv means the upper concave envelope. Clearly, as $T \to \infty$, $\mathcal{M}^{-1}(\frac{1}{\sqrt{T}}) \to 0$. We will next show that the convergence rate of $\mathcal{M}^{-1}$ is indeed $O(\frac{1}{\sqrt{T}})$ and the minimization over r makes the overall convergence rate $O(T^{-1/4})$.

Starting from (C.8), we denote $x = \frac{4}{\sqrt{T}}\sqrt{(\mathcal{L}_0 - \mathcal{L}_*)L\left(1+\frac{\sigma^2 d}{B^2}\right)}$ and write

$$
\begin{aligned}
\mathcal{M}^{-1}(x; r, \xi, \gamma) &= \frac{-\frac{\xi}{r}\gamma + (r^2-1)\frac{\xi}{r}x + r\gamma x + \gamma\sqrt{(\frac{\xi}{r})^2 + 2\xi x + 2\gamma x + x^2}}{2\gamma - (r^2-1)x} \\
&= \left(-\frac{\gamma\xi}{r} + (r^2-1)\frac{\xi}{r}x + r\gamma x + \gamma\sqrt{(\frac{\xi}{r})^2 + 2\xi x + 2\gamma x + x^2}\right) \\
&\quad \cdot \frac{1 + \frac{r^2-1}{2\gamma}x + O(x^2)}{2\gamma} \\
&= \frac{1}{2\gamma}\left(-\frac{\gamma\xi}{r} + (r^2-1)\frac{\xi}{r}x + r\gamma x + \frac{\gamma\xi}{r}\sqrt{1 + \frac{2(\xi+\gamma)r^2 x}{\xi^2} + O(x^2)}\right) \\
&\quad \cdot (1 + \frac{r^2-1}{2\gamma}x + O(x^2)) \\
&= \frac{1}{2\gamma}\left(-\frac{\gamma\xi}{r} + (r^2-1)\frac{\xi}{r}x + r\gamma x + \frac{\gamma\xi}{r}\left(1 + \frac{(\xi+\gamma)r^2 x}{\xi^2} + O(x^2)\right)\right) \\
&\quad \cdot (1 + \frac{r^2-1}{2\gamma}x + O(x^2)) \\
&= \frac{1}{2\gamma}\left((r^2-1)\frac{\xi}{r}x + r\gamma x + \frac{\gamma(\xi+\gamma)rx}{\xi} + O(x^2)\right) \cdot (1 + \frac{r^2-1}{2\gamma}x + O(x^2)) \\
&= \frac{1}{2\gamma}\left((r^2-1)\frac{\xi}{r} + r\gamma + \frac{\gamma(\xi+\gamma)r}{\xi}\right) \cdot x + O(x^2) \\
&= \frac{1}{2\gamma}\left(\frac{(\xi+\gamma)^2}{\xi}r - \frac{\xi}{r}\right) \cdot x + O(x^2)
\end{aligned}
$$

Since $\mathcal{M}^{-1}$ is asymptotically linear as $x \to 0$, we instead study

$$
\min_{r>0} \frac{\xi}{r} + \mathcal{M}^{-1}(x; r, \xi, \gamma) \equiv \min_{r>0} \frac{\xi}{r} + \frac{1}{2\gamma}\left(\frac{(\xi+\gamma)^2}{\xi}r - \frac{\xi}{r}\right) \cdot x + O(x^2).
$$

That is, ignoring the higher order term for the asymptotic analysis, the $\mathcal{M}^{-1}$ part converges as $O(x) = O(1/\sqrt{T})$, and we visualize this in Figure 8.

Although DP-SGD converges faster than SGD, the former converges to ξ/r and the latter converges to 0. Thus, taking ξ/r into consideration, the objective reduces to a hyperbola

$$
\frac{\left(\xi(1 - \frac{x}{2\gamma})\right)}{r} + \frac{x(\xi+\gamma)^2}{2\gamma\xi} \cdot r
$$

whose minimum over r is obviously $2\sqrt{\xi(1 - \frac{x}{2\gamma})\frac{x(\xi+\gamma)^2}{2\gamma\xi}} = O(\sqrt{x}) = O(T^{-1/4})$. □

To give more details about the upper bound in (5.2), we demonstrate its dependence on ξ and γ in Figure 7.

C.2 Main proof of convergence for DP-SGD (the non-envelope version)

Proof of Theorem 6. Consider DP-SGD with AUTO-S clipping

$$
\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta\left(\sum_i \frac{\tilde{\boldsymbol{g}}_{t,i}}{\|\tilde{\boldsymbol{g}}_{t,i}\| + \gamma} + \sigma\mathcal{N}(0, \mathbf{I})\right)
$$

where $\tilde{\boldsymbol{g}}_{t,i}$ is i.i.d. samples of $\tilde{\boldsymbol{g}}_t$, an unbiased estimate of $\boldsymbol{g}_t$, with a bounded variance as described in Assumption 5.3.

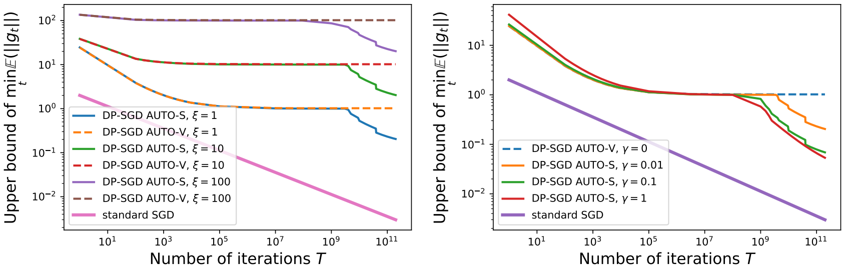

Figure 7: Dependence of the upper bound $\mathcal{G}$ on ξ (left) and γ (right). Here the $O(1/\sqrt{T})$ term is set to 10 and either $\gamma = 0.01$ (left) or $\xi = 1$ (right).

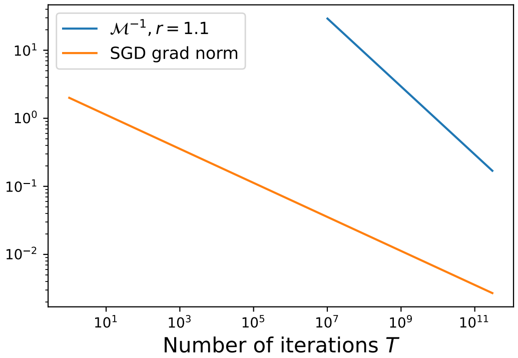

Figure 8: Convergence with respect to T. Same setting as Figure 5.

By Lipschitz smoothness in Assumption 5.2, and denoting $Z = \mathcal{N}(0, \mathbf{I})$, we have

$$\begin{aligned}
\mathcal{L}_{t+1} - \mathcal{L}_t &\leq \boldsymbol{g}_t^\top(\boldsymbol{w}_{t+1} - \boldsymbol{w}_t) + \frac{L}{2}\|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\|^2 \\
&= -\eta \boldsymbol{g}_t^\top \left(\sum_i \frac{\tilde{\boldsymbol{g}}_{t,i}}{\|\tilde{\boldsymbol{g}}_{t,i}\| + \gamma} + \sigma Z\right) + \frac{L\eta^2}{2}\left\|\sum_i \frac{\tilde{\boldsymbol{g}}_{t,i}}{\|\tilde{\boldsymbol{g}}_{t,i}\| + \gamma} + \sigma Z\right\|^2 \\
&\leq -\eta \boldsymbol{g}_t^\top \left(\sum_i \frac{\tilde{\boldsymbol{g}}_{t,i}}{\|\tilde{\boldsymbol{g}}_{t,i}\| + \gamma} + \sigma Z\right) \\
&\quad + L\eta^2\left(\left\|\sum_i \frac{\tilde{\boldsymbol{g}}_{t,i}}{\|\tilde{\boldsymbol{g}}_{t,i}\| + \gamma}\right\|^2 + \sigma^2\|Z\|^2\right)
\end{aligned}$$

where the last inequality follows from Cauchy Schwartz.

Given the fact that $\|\tilde{\boldsymbol{g}}_{t,i}/(\|\tilde{\boldsymbol{g}}_{t,i}\| + \gamma)\| \leq 1$, the expected improvement at one iteration is

$$\begin{aligned}
\mathbb{E}(\mathcal{L}_{t+1} - \mathcal{L}_t | \boldsymbol{w}_t) &\leq -\eta \boldsymbol{g}_t^\top \mathbb{E}\left(\sum_i \frac{\tilde{\boldsymbol{g}}_{t,i}}{\|\tilde{\boldsymbol{g}}_{t,i}\| + \gamma}\right) + L\eta^2\left(B^2 + \sigma^2 d\right) \\
&= -\eta B \boldsymbol{g}_t^\top \mathbb{E}\left(\frac{\tilde{\boldsymbol{g}}_t}{\|\tilde{\boldsymbol{g}}_t\| + \gamma}\right) + L\eta^2\left(B^2 + \sigma^2 d\right)
\end{aligned} \tag{C.2}$$

Now we want to lower bound $\boldsymbol{g}_t^\top \mathbb{E}\left(\frac{\tilde{\boldsymbol{g}}_t}{\|\tilde{\boldsymbol{g}}_t\| + \gamma}\right)$ in (C.2).

Write $\tilde{\boldsymbol{g}}_t = \boldsymbol{g}_t + \Delta_t$ where the gradient noise Δ_t follows $\mathbb{E}\Delta_t = 0, \mathbb{E}\|\Delta_t\| < \xi$ by Assumption 5.3. Then

$$\begin{aligned}
\boldsymbol{g}_t^\top \mathbb{E}\left(\frac{\tilde{\boldsymbol{g}}_t}{\|\tilde{\boldsymbol{g}}_t\| + \gamma}\right) &= \mathbb{E}\left(\frac{\|\boldsymbol{g}_t\|^2 + \boldsymbol{g}_t^\top \Delta_t}{\|\boldsymbol{g}_t + \Delta_t\| + \gamma}\right) \\
&= \frac{1}{2}\mathbb{E}\left(\frac{\|\boldsymbol{g}_t\|^2 + \boldsymbol{g}_t^\top \Delta_t}{\|\boldsymbol{g}_t + \Delta_t\| + \gamma}\Big|\Delta_t \in H_+\right) + \frac{1}{2}\mathbb{E}\left(\frac{\|\boldsymbol{g}_t\|^2 + \boldsymbol{g}_t^\top \Delta_t}{\|\boldsymbol{g}_t + \Delta_t\| + \gamma}\Big|\Delta_t \in H_-\right) \\
&= \frac{1}{2}\mathbb{E}\left(\frac{\|\boldsymbol{g}_t\|^2 + \boldsymbol{g}_t^\top \Delta_t}{\|\boldsymbol{g}_t + \Delta_t\| + \gamma}\Big|\Delta_t \in H_+\right) + \frac{1}{2}\mathbb{E}\left(\frac{\|\boldsymbol{g}_t\|^2 - \boldsymbol{g}_t^\top \Delta_t}{\|\boldsymbol{g}_t - \Delta_t\| + \gamma}\Big|\Delta_t \in H_+\right)
\end{aligned}$$

where we use the hyperplane perpendicular to $\boldsymbol{g}_t$ to divide the support of Δ_t into two half-spaces:

$$H_+ := \{\boldsymbol{v} : \boldsymbol{g}_t^\top \boldsymbol{v} > 0\}, \quad H_- := \{\boldsymbol{v} : \boldsymbol{g}_t^\top \boldsymbol{v} < 0\}.$$

We use the symmetry assumption in Assumption 5.3 to get

$$\mathbb{P}(\Delta_t \in H_+) = \mathbb{P}(\Delta_t \in H_-) = \frac{1}{2}$$

and notice that $\Delta_t \overset{D}{=} -\Delta_t$, i.e., if $\Delta_t \in H_+$, then $-\Delta_t \in H_-$ with the same distribution.

The next result further gives a lower bound for $\boldsymbol{g}_t^\top \mathbb{E}\left(\frac{\tilde{\boldsymbol{g}}_t}{\|\tilde{\boldsymbol{g}}_t\|+\gamma}\right)$ using $\|\boldsymbol{g}_t\|$.

Lemma C.1.

$$\mathbb{E}\left(\frac{\|\boldsymbol{g}_t\|^2 + \boldsymbol{g}_t^\top \Delta_t}{\|\boldsymbol{g}_t + \Delta_t\| + \gamma} + \frac{\|\boldsymbol{g}_t\|^2 - \boldsymbol{g}_t^\top \Delta_t}{\|\boldsymbol{g}_t - \Delta_t\| + \gamma}\Big|\Delta_t \in H_+\right) \geq \min_{0<c\leq 1} f(c, r; \frac{\gamma}{\|\boldsymbol{g}_t\|}) \cdot (\|\boldsymbol{g}_t\| - \xi/r)$$

for any $r > 0$ *and* $f(c, r; \Gamma) = \frac{(1+rc)}{\sqrt{r^2+2rc+1}+\Gamma} + \frac{(1-rc)}{\sqrt{r^2-2rc+1}+\Gamma}$.

For the simplicity of notation, we denote the distance measure

$$\mathcal{M}(\|\boldsymbol{g}_t\| - \xi/r; r, \xi, \gamma) = \min_{0<c\leq 1} f\left(c, r; \frac{\gamma}{\|\boldsymbol{g}_t\|}\right) \cdot (\|\boldsymbol{g}_t\| - \xi/r) \tag{C.3}$$

and leave the fine-grained analysis (e.g. its explicit form in some scenarios) at the end of this section.

Using the lower bound from Lemma C.1, the expected improvement (C.2) becomes

$$\mathbb{E}(\mathcal{L}_{t+1} - \mathcal{L}_t | \boldsymbol{w}_t) \leq -\frac{\eta B}{2}\mathcal{M}(\|\boldsymbol{g}_t\| - \xi/r) + L\eta^2 B^2\left(1 + \frac{\sigma^2 d}{B^2}\right)$$

Now extend the expectation over randomness in the trajectory, and perform a telescoping sum over the iterations

$$\begin{aligned}
\mathcal{L}_0 - \mathcal{L}_* \geq \mathcal{L}_0 - \mathbb{E}\mathcal{L}_T &= \sum_t \mathbb{E}(\mathcal{L}_t - \mathcal{L}_{t+1}) \\
&\geq \frac{\eta B}{2}\mathbb{E}\left(\sum_t \mathcal{M}(\|\boldsymbol{g}_t\| - \xi/r)\right) - TL\eta^2 B^2\left(1 + \frac{\sigma^2 d}{B^2}\right)
\end{aligned}$$

Substituting $\eta B = \eta_0/\sqrt{T}$ where η_0 is a base learning rate, we have

$$2(\mathcal{L}_0 - \mathcal{L}_*) \geq \sqrt{T}\eta_0 \mathbb{E}\left(\frac{1}{T}\sum_t \mathcal{M}(\|\boldsymbol{g}_t\| - \xi/r)\right) - 2L\eta_0^2\left(1 + \frac{\sigma^2 d}{B^2}\right)$$

and finally

$$\mathbb{E}\left(\frac{1}{T}\sum_t \mathcal{M}(\|\boldsymbol{g}_t\| - \xi/r)\right) \leq \frac{1}{\sqrt{T}}\left[\frac{2(\mathcal{L}_0 - \mathcal{L}_*)}{\eta_0} + 2L\eta_0\left(1 + \frac{\sigma^2 d}{B^2}\right)\right] \tag{C.4}$$

With η_0 chosen properly at $\eta_0 = \sqrt{\frac{\mathcal{L}_0-\mathcal{L}_*}{L\left(1+\frac{\sigma^2 d}{B^2}\right)}}$, the hyperbola on the right hand side in (C.4) is minimized to $4\sqrt{(\mathcal{L}_0-\mathcal{L}_*)L\left(1+\frac{\sigma^2 d}{B^2}\right)}$, and we obtain

$$\mathbb{E}\left(\frac{1}{T}\sum_t \mathcal{M}(\|\boldsymbol{g}_t\|-\xi/r)\right) \leq \frac{4}{\sqrt{T}}\sqrt{(\mathcal{L}_0-\mathcal{L}_*)L\left(1+\frac{\sigma^2 d}{B^2}\right)}$$

Since the minimum of a sequence is smaller than the average, we have

$$\min_t \mathbb{E}(\mathcal{M}(\|\boldsymbol{g}_t\|-\xi/r)) \leq \frac{1}{T}\sum_t \mathbb{E}\left(\mathcal{M}(\|\boldsymbol{g}_t\|-\xi/r)\right) \leq \frac{4}{\sqrt{T}}\sqrt{(\mathcal{L}_0-\mathcal{L}_*)L\left(1+\frac{\sigma^2 d}{B^2}\right)} \tag{C.5}$$

We claim that $\mathcal{M}$ may not be concave or convex. Therefore we use $\mathcal{M}_{cvx}$ to denote its lower convex envelope, i.e. the largest convex function that is smaller than $\mathcal{M}$. Then by Jensen's inequality (C.5) becomes

$$\min_t \mathcal{M}_{cvx}(\mathbb{E}(\|\boldsymbol{g}_t\|-\xi/r)) \leq \min_t \mathbb{E}(\mathcal{M}_{cvx}(\|\boldsymbol{g}_t\|-\xi/r)) \leq \frac{4}{\sqrt{T}}\sqrt{(\mathcal{L}_0-\mathcal{L}_*)L\left(1+\frac{\sigma^2 d}{B^2}\right)} \tag{C.6}$$

It is obvious that $\mathcal{M}_{cvx}$ is increasing as $\mathcal{M}$ is increasing by Theorem 8. Hence, $(\mathcal{M}_{cvx})^{-1}$ is also increasing, as the inverse of $\mathcal{M}_{cvx}$. We write (C.6) as

$$\min_t \mathbb{E}(\|\boldsymbol{g}_t\|-\xi/r) \leq (\mathcal{M}_{cvx})^{-1}\left(\frac{4}{\sqrt{T}}\sqrt{(\mathcal{L}_0-\mathcal{L}_*)L\left(1+\frac{\sigma^2 d}{B^2}\right)}\right)$$

and equivalently

$$\min_t \mathbb{E}(\|\boldsymbol{g}_t\|) \leq \frac{\xi}{r} + (\mathcal{M}_{cvx})^{-1}\left(\frac{4}{\sqrt{T}}\sqrt{(\mathcal{L}_0-\mathcal{L}_*)L\left(1+\frac{\sigma^2 d}{B^2}\right)}\right) \tag{C.7}$$

Finally, we derive the explicit properties of $\mathcal{M}(\|\boldsymbol{g}_t\|-\xi/r)$ in Theorem 8. These properties allow us to further analyze on the convergence of $\mathcal{M}(\|\boldsymbol{g}_t\|-\xi/r)$, based on AUTO-V and AUTO-S, respectively.

1. DP-SGD with AUTO-V clipping. By Theorem 8, we write

$$\mathcal{M}(x;r) = \min_{c\in(0,1]} f(c,r;0)\cdot x$$

This is a linear function and thus $\mathcal{M}_{cvx} = \mathcal{M} = 1/\mathcal{M}_{cvx}^{-1}$. As a result, we have

$$\min_t \mathbb{E}(\|\boldsymbol{g}_t\|) \leq \frac{\xi}{r} + \frac{1}{\min_{c\in(0,1]} f(c,r;0)}\cdot\frac{4}{\sqrt{T}}\sqrt{(\mathcal{L}_0-\mathcal{L}_*)L\left(1+\frac{\sigma^2 d}{B^2}\right)}$$

We note here r plays an important role under AUTO-V clipping: when $r<1$, we spend more iterations to converge to better and smaller gradient norm ξ/r; when $r\geq 1$, $\min_c f(c,r;0) = f(1,r;0) = 0$ and it takes forever to converge. This is demonstrated in the left plot of Figure 5.

2. DP-SGD with AUTO-S clipping. By Theorem 8 and for $r>1$, we write

$$\mathcal{M}(x;r,\xi,\gamma) = \left(\frac{\gamma}{(r-1)(x+\xi/r)+\gamma} - \frac{\gamma}{(r+1)(x+\xi/r)+\gamma}\right)\cdot x.$$

Notice that the inverse of a lower convex envelope is equivalent to the upper concave envelope (denoted by the subscript ccv) of an inverse. Therefore we can derive $(\mathcal{M}_{cvx})^{-1} = (\mathcal{M}^{-1})_{ccv}$ with the explicit form

$$\mathcal{M}^{-1}(x;r,\xi,\gamma) = \frac{-\frac{\xi}{r}\gamma + (r^2-1)\frac{\xi}{r}x + r\gamma x + \gamma\sqrt{(\frac{\xi}{r})^2 + 2\xi x + 2\gamma x + x^2}}{2\gamma-(r^2-1)x}. \tag{C.8}$$

we can derive it based on r, ξ, γ and substitute back to (C.7).

Note that the domain of $\mathcal{M}^{-1}$ (or the image of $\mathcal{M}$) is $[0, \frac{\gamma}{r-1} - \frac{\gamma}{r+1})$.

In comparison to the AUTO-V clipping, $\mathcal{M}^{-1}$ takes a much more complicated form, as depicted in the middle plot of Figure 5, where $r > 1$ plays an important role for the gradient norm to converge to zero. □

C.3 Proof of Lemma C.1

Proof of Lemma C.1. We want to lower bound

$$\mathbb{E}\left(\frac{\|\boldsymbol{g}_t\|^2 + \boldsymbol{g}_t^\top \Delta_t}{\|\boldsymbol{g}_t + \Delta_t\| + \gamma} + \frac{\|\boldsymbol{g}_t\|^2 - \boldsymbol{g}_t^\top \Delta_t}{\|\boldsymbol{g}_t - \Delta_t\| + \gamma}\Big|\Delta_t \in H_+\right) \tag{C.9}$$

To simplify the notation, we denote noise-to-signal ratio $S := \frac{\|\Delta_t\|}{\|\boldsymbol{g}_t\|}$ and $c := \cos\theta = \frac{\boldsymbol{g}_t^\top \Delta_t}{\|\boldsymbol{g}_t\|\|\Delta_t\|}$, with θ be the random angle between $\boldsymbol{g}_t$ and Δ_t. Note that $0 < c \leq 1$ when $\Delta_t \in H_+$.

The term inside the conditional expectation in (C.9) can be written as

$$\begin{aligned}
&\frac{(1+Sc)\|\boldsymbol{g}_t\|^2}{\sqrt{S^2+2Sc+1}\|\boldsymbol{g}_t\| + \gamma} + \frac{(1-Sc)\|\boldsymbol{g}_t\|^2}{\sqrt{S^2-2Sc+1}\|\boldsymbol{g}_t\| + \gamma} \\
=&\|\boldsymbol{g}_t\|\left(\frac{(1+Sc)}{\sqrt{S^2+2Sc+1} + \gamma/\|\boldsymbol{g}_t\|} + \frac{(1-Sc)}{\sqrt{S^2-2Sc+1} + \gamma/\|\boldsymbol{g}_t\|}\right)
\end{aligned}$$

Defining $\Gamma = \gamma/\|\boldsymbol{g}_t\|$ and

$$f(c, S; \Gamma) := \frac{(1+Sc)}{\sqrt{S^2+2Sc+1} + \Gamma} + \frac{(1-Sc)}{\sqrt{S^2-2Sc+1} + \Gamma}, \tag{C.10}$$

we turn the conditional expectation in (C.9) into

$$\mathbb{E}\left(\frac{\|\boldsymbol{g}_t\|^2 + \boldsymbol{g}_t^\top \Delta_t}{\|\boldsymbol{g}_t + \Delta_t\| + \gamma} + \frac{\|\boldsymbol{g}_t\|^2 - \boldsymbol{g}_t^\top \Delta_t}{\|\boldsymbol{g}_t - \Delta_t\| + \gamma}\Big|\Delta_t \in H_+\right) = \|\boldsymbol{g}_t\|\mathbb{E}(f(c, S; \Gamma)|\Delta_t \in H_+) \tag{C.11}$$

for which we want to lower bound $f(c, S; \Gamma)$ over $0 < c \leq 1, S > 0, \Gamma > 0$. We use the next theorem to prepare some helpful properties. The proof can be found in Appendix E.1.

Theorem 7. *For f defined in (C.10), we have*

1. *$f(c, S; \Gamma)$ is strictly decreasing in S for all $0 < c < 1$ and $\Gamma > 0$.*
2. *Consequently, $\min_{c \in (0,1)} f(c, S; \Gamma)$ is strictly decreasing in S.*
3. *$f(c, S; \Gamma)$ is strictly decreasing in c for all $S > 1$ and $\Gamma > 0$.*

We consider a thresholding ratio $r > 0$ and we will focus on the regime that $S < r$. This r will turn out to measure the minimum gradient norm at convergence: informally speaking, $\|\boldsymbol{g}_t\|$ converges to ξ/r.

By the law of total expectation, (C.11) can be relaxed as follows.

$$\begin{aligned}
&\|\boldsymbol{g}_t\|\mathbb{E}\left(f(c,S;\Gamma)\Big|\Delta\in H_+\right)\\
=&\|\boldsymbol{g}_t\|\mathbb{E}\left(f(c,S;\Gamma)\Big|\Delta\in H_+,S<r\right)\mathbb{P}(r\|\boldsymbol{g}_t\|>\|\Delta\|\Big|\Delta\in H_+)\\
&+\|\boldsymbol{g}_t\|\mathbb{E}\left(f(c,S;\Gamma)\Big|\Delta\in H_+,S>r\right)\mathbb{P}(r\|\boldsymbol{g}_t\|<\|\Delta\|\Big|\Delta\in H_+)\\
\geq&\|\boldsymbol{g}_t\|\mathbb{E}\left(f(c,S;\Gamma)\Big|\Delta\in H_+,S<r\right)\mathbb{P}(r\|\boldsymbol{g}_t\|>\|\Delta\|\Big|\Delta\in H_+)\\
\geq&\|\boldsymbol{g}_t\|\mathbb{E}\left(f(c,r;\Gamma)\Big|\Delta\in H_+,S<r\right)\mathbb{P}(r\|\boldsymbol{g}_t\|>\|\Delta\|\Big|\Delta\in H_+)\\
=&\|\boldsymbol{g}_t\|\mathbb{E}\left(f(c,r;\Gamma)\Big|\Delta\in H_+,S<r\right)\mathbb{P}(r\|\boldsymbol{g}_t\|>\|\Delta\|)\\
\geq&\min_{c\in(0,1]}f(c,r;\Gamma)\cdot\underbrace{\|\boldsymbol{g}_t\|\mathbb{P}(r\|\boldsymbol{g}_t\|>\|\Delta\|)}_{\textcircled{$\star$}}
\end{aligned}\tag{C.12}$$

where in the first inequality, the ignoring of last term is justified by $f(c,S;\Gamma)\geq\min_{c\in(0,1]}f(c,S;\Gamma)\geq\min_{c\in(0,1]}f(c,\infty;\Gamma)=0$, from the monotonicity (second statement) in Theorem 7.

We first lower bound \textcircled{$\star$} by applying the Markov's inequality:

$$\mathbb{P}(r\|\boldsymbol{g}_t\|>\|\Delta_t\|)\geq 1-\frac{\mathbb{E}\|\Delta_t\|}{r\|\boldsymbol{g}_t\|}$$

and hence by Assumption 5.3,

$$\|\boldsymbol{g}_t\|\mathbb{P}(r\|\boldsymbol{g}_t\|>\|\Delta_t\|)\geq\|\boldsymbol{g}_t\|-\mathbb{E}\|\Delta\|/r\geq\|\boldsymbol{g}_t\|-\xi/r.$$

Finally, the conditional expectation of interest in (C.9) gives

$$\mathbb{E}\left(\frac{\|\boldsymbol{g}_t\|^2+\boldsymbol{g}_t^\top\Delta_t}{\|\boldsymbol{g}_t+\Delta_t\|}+\frac{\|\boldsymbol{g}_t\|^2-\boldsymbol{g}_t^\top\Delta_t}{\|\boldsymbol{g}_t-\Delta_t\|}\Big|\Delta_t\in H_+\right)\geq\min_{0<c\leq 1}f(c,r;\frac{\gamma}{\|\boldsymbol{g}_t\|})\cdot(\|\boldsymbol{g}_t\|-\xi/r)$$

□

C.4 Proof of Theorem 8

To derive some properties of $\min_c f(c,r;\Gamma)$, we need to compute separately for AUTO-V (without the stability constant, $\Gamma=0$) and for AUTO-S (with the stability constant, $\Gamma>0$), as shown in Theorem 8. As we will show, as the number of training iterations $T\to\infty$, DP-SGD with AUTO-V clipping can only compress $\|\boldsymbol{g}_t\|$ to ξ/r for $r<1$. However, DP-SGD with AUTO-S clipping can compress $\|\boldsymbol{g}_t\|$ to ξ/r to any $r>1$.

Theorem 8.

1. *For $0<r<1$ and $\Gamma=0$, we have $\min_{c\in(0,1]}f(c,r;0)>0$. Then Equation (C.11) is lower bounded by*
$$\min_{c\in(0,1]}f(c,r;0)\cdot(\|\boldsymbol{g}_t\|-\xi/r)$$
which is increasing in $\|g\|-\xi/r$.

2. *For $r\geq 1$ and $\Gamma=0$, we have $\min_{c\in(0,1]}f(c,r;\Gamma)=f(1,r;0)=0$. In words, (C.9) has a trivial lower bound and Theorem 6 cannot compress $\|\boldsymbol{g}_t\|$ to ξ/r.*

3. *For $r\geq 1$ and $\Gamma>0$, we have $\min_{c\in(0,1]}f(c,r;\Gamma)=f(1,r;\Gamma)=\left(\frac{\Gamma}{r+\Gamma-1}-\frac{\Gamma}{r+\Gamma+1}\right)$. Then Equation (C.11) is lower bounded by*
$$\left(\frac{\gamma}{(r-1)\|\boldsymbol{g}_t\|+\gamma}-\frac{\gamma}{(r+1)\|\boldsymbol{g}_t\|+\gamma}\right)\cdot(\|\boldsymbol{g}_t\|-\xi/r)$$
which is increasing in $\|\boldsymbol{g}_t\|-\xi/r$.

Proof. To prove statement 1, we use the second statement from Theorem 7 and show that $\min_c f(c,r;0)>\min_c f(c,\infty;0)=0$. To prove statement 2 and 3, we use the third statement from Theorem 7 and see that $\min_c f(c,r;\Gamma)=f(1,r;\Gamma)$ with an explicit formula. □

D Convergence rate of standard SGD

Theorem 9. *Under Assumption 5.1, 5.2, 5.3 (without the symmetry assumption), running the standard non-DP SGD for T iterations gives, for $\eta \propto 1/\sqrt{T}$,*

$$\min_t \mathbb{E}\left(\|\boldsymbol{g}_t\|\right) \le \frac{1}{T^{1/4}}\sqrt{2(\mathcal{L}_0-\mathcal{L}_*)L+\frac{\xi^2}{B}}$$

Proof of Theorem 9. Consider the standard SGD

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta\frac{\sum_i \tilde{\boldsymbol{g}}_{t,i}}{B}$$

where $\tilde{\boldsymbol{g}}_{t,i}$ is i.i.d. unbiased estimate of $\boldsymbol{g}_t$, with a bounded variance as described in Assumption 5.3. By Lipschitz smoothness assumption in Assumption 5.2,

$$\mathcal{L}_{t+1}-\mathcal{L}_t \le \boldsymbol{g}_t^\top(\boldsymbol{w}_{t+1}-\boldsymbol{w}_t)+\frac{L}{2}\|\boldsymbol{w}_{t+1}-\boldsymbol{w}_t\|^2 = -\eta\boldsymbol{g}_t^\top\left(\sum_i\frac{1}{B}\tilde{\boldsymbol{g}}_{t,i}\right)+\frac{L\eta^2}{2}\left\|\sum_i\frac{1}{B}\tilde{\boldsymbol{g}}_{t,i}\right\|^2$$

The expected improvement at one iteration is

$$\begin{aligned}\mathbb{E}(\mathcal{L}_{t+1}-\mathcal{L}_t|\boldsymbol{w}_t) &\le -\eta\boldsymbol{g}_t^\top\mathbb{E}\tilde{\boldsymbol{g}}_{t,i}+\frac{L\eta^2}{2}\mathbb{E}\|\sum_i\frac{1}{B}\tilde{\boldsymbol{g}}_{t,i}\|^2\\ &\le -\eta\|\boldsymbol{g}_t\|^2+\frac{L\eta^2}{2}\left(\|\boldsymbol{g}_t\|^2+\frac{\xi^2}{B}\right)\end{aligned}\tag{D.1}$$

Now we extend the expectation over randomness in the trajectory, and perform a telescoping sum over the iterations

$$\mathcal{L}_0-\mathcal{L}_* \ge \mathcal{L}_0-\mathbb{E}\mathcal{L}_T=\sum_t\mathbb{E}(\mathcal{L}_t-\mathcal{L}_{t+1})\ge\left(\eta-\frac{L\eta^2}{2}\right)\mathbb{E}(\sum_t\|\boldsymbol{g}_t\|^2)-\frac{TL\eta^2\xi^2}{2B}$$

Notice that we do not need the symmetry assumption in Assumption 5.3 in the non-DP SGD analysis. We apply the same learning rate as in [5], $\eta=\frac{1}{L\sqrt{T}}$,

$$2(\mathcal{L}_0-\mathcal{L}_*)\ge\left(\frac{2}{L\sqrt{T}}-\frac{1}{LT}\right)\mathbb{E}\left(\sum_t\|\boldsymbol{g}_t\|^2\right)-\frac{T\xi^2}{BLT}\ge\frac{\sqrt{T}}{L}\mathbb{E}\left(\frac{1}{T}\sum_t\|\boldsymbol{g}_t\|^2\right)-\frac{\xi^2}{BL}$$

and finally

$$\min_t\mathbb{E}\left(\|\boldsymbol{g}_t\|^2\right)\le\mathbb{E}\left(\frac{1}{T}\sum_t\|\boldsymbol{g}_t\|^2\right)\le\frac{1}{\sqrt{T}}\left[2(\mathcal{L}_0-\mathcal{L}_*)L+\frac{\xi^2}{B}\right]$$

Using the Jensen's inequality, we can have

$$\min_t \mathbb{E}\left(\|\boldsymbol{g}_t\|\right) \le \frac{1}{T^{1/4}}\sqrt{2(\mathcal{L}_0-\mathcal{L}_*)L+\frac{\xi^2}{B}}$$

□

E Auxiliary proofs

E.1 Proof of Theorem 7

Proof. We first show $\frac{df(c,S;\Gamma)}{dS}<0$ for all $0<c<1,\Gamma>0$ and $S>0$, as visualized in the left plot of Figure 9. We can explicitly write down the derivative, by WolframAlpha

$$\frac{df(c,S;\Gamma)}{dS}=\frac{-(A\Gamma^2+B\Gamma+C)}{\sqrt{S^2-2cS+1}\sqrt{S^2+2cS+1}(\Gamma+\sqrt{S^2-2cS+1})^2(\Gamma+\sqrt{S^2+2cS+1})^2}\tag{E.1}$$

with

$$\begin{aligned}
A(c,S) &= \sqrt{S^2+2cS+1}\left(3c^2S-2c(S^2+1)+S\right)+\sqrt{S^2-2cS+1}\left(3c^2S+2c(S^2+1)+S\right)\\
B(c,S) &= 4S\left[(S^2+1)(1-c^2)+c^2\sqrt{S^2+2cS+1}\sqrt{S^2-2cS+1}\right]\\
C(c,S) &= (1-c^2)S\left[(S^2-2cS+1)^{3/2}+(S^2+2cS+1)^{3/2}\right]
\end{aligned}$$

It is obvious that, since $c<1$,

$$S^2 \pm 2cS + 1 > S^2 \pm 2cS + c^2 = (S \pm c)^2 \geq 0. \tag{E.2}$$

From (E.2), the denominator in (E.1) is positive and it suffices to show $A\Gamma^2 + B\Gamma + C > 0$ for all $0 < c < 1$ and $S > 0$, in order to show $\frac{df}{dS} < 0$.

Also from (E.2), we can easily see $B(c,S) > 0$ and $C(c,S) > 0$. We will show that $A(c,S) > 0$ in Lemma E.1, after very heavy algebraic computation.

Now we can claim that $A\Gamma^2 + B\Gamma + C > 0$ by Fact E.3, and complete the proof of the first statement.

To further see that $\min_c f(c,S;\Gamma)$ is decreasing in S, let us denote $c^*(x;\Gamma) := \arg\min_{c\in[0,1]} f(c,x;\Gamma)$. Then considering $S < S'$, we prove the second statement by observing

$$\min_c f(c,S;\Gamma) = f(c^*(S;\Gamma),S;\Gamma) > f(c^*(S;\Gamma),S';\Gamma) \geq \min_c f(c,S';\Gamma).$$

This statement is also visualized in the right plot of Figure 9.

We next show $\frac{df(c,S;\Gamma)}{dc} < 0$ for all $0 < c < 1, \Gamma > 0$ and $S > 1$. We can explicitly write down the derivative, by WolframAlpha

$$\frac{df(c,S;\Gamma)}{dc} = \frac{-S(A'\Gamma^2 + B'\Gamma + C')}{\sqrt{S^2-2cS+1}\sqrt{S^2+2cS+1}(\Gamma+\sqrt{S^2-2cS+1})^2(\Gamma+\sqrt{S^2+2cS+1})^2} \tag{E.3}$$

with

$$\begin{aligned}
A'(c,S) &= \left[(S^2+3cS+2)\sqrt{S^2-2cS+1}-(S^2-3cS+2)\sqrt{S^2+2cS+1}\right]\\
B'(c,S) &= 4Sc\left[\sqrt{S^2+2cS+1}\sqrt{S^2-2cS+1}+(S^2-1)\right]\\
C'(c,S) &= S\left[(c+S)(S^2-2cS+1)^{3/2}+(c-S)(S^2+2cS+1)^{3/2}\right]
\end{aligned}$$

Clearly $B'(c,S) > 0$ and $C'(c,S) > 0$, since $S^2 + 2cS + 1 > S^2 - 2cS + c^2 = (S-c)^2 \geq 0$. And we will show $A'(c,S) > 0$ in Lemma E.2, after some algebra.

We again claim that $A'\Gamma^2 + B'\Gamma + C' > 0$ by Fact E.3, which guarantees that the numerator in (E.3) is negative and that $\frac{df}{dc} < 0$. This is visualized in Figure 10. □

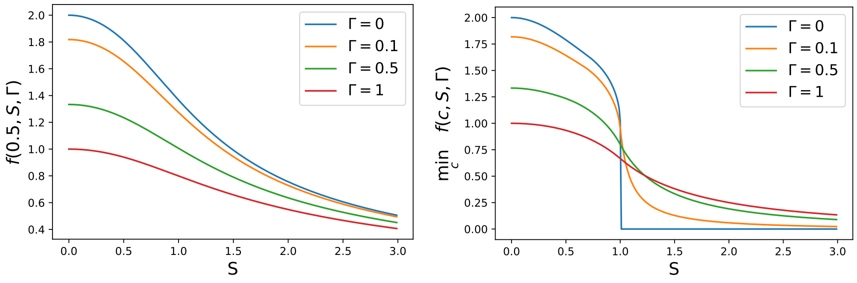

Figure 9: Visualization of $f(0.5,S,\Gamma)$ (left) and $\min_{0\leq c\leq 1} f(c,S,\Gamma)$ over $S>0$.

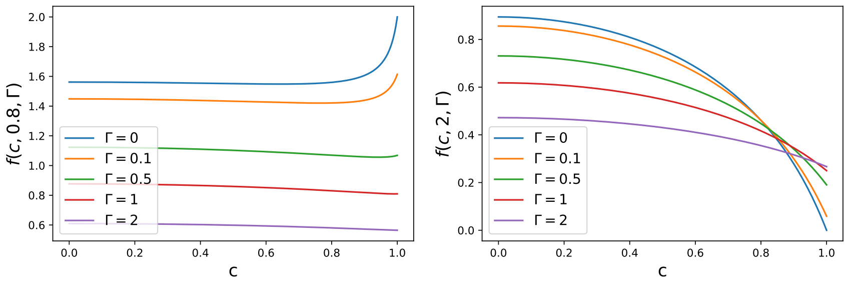

Figure 10: Visualization of $f(c, 0.8, \Gamma)$ (left) and $f(c, 2, \Gamma)$ over $0 \leq c \leq 1$.

E.2 Proof of Lemma E.1

Lemma E.1. *For all* $0 < c < 1$ *and* $S > 0$*,*

$$A := \sqrt{S^2 + 2cS + 1}\left(3c^2S - 2c(S^2+1) + S\right) + \sqrt{S^2 - 2cS + 1}\left(3c^2S + 2c(S^2+1) + S\right) > 0.$$

Proof. We prove by contradiction. Suppose

$$\sqrt{S^2 + 2cS + 1}\left(3c^2S - 2c(S^2+1) + S\right) + \sqrt{S^2 - 2cS + 1}\left(3c^2S + 2c(S^2+1) + S\right) < 0.$$

Then

$$0 < \sqrt{S^2 - 2cS + 1}\left(3c^2S + 2c(S^2+1) + S\right) < -\sqrt{S^2 + 2cS + 1}\left(3c^2S - 2c(S^2+1) + S\right).$$

where the first inequality comes from $S^2 - 2cS + 1 > S^2 - 2cS + c^2 = (S-c)^2 \geq 0$.

Squaring everything gives

$$(S^2 - 2cS + 1)\left(3c^2S + 2c(S^2+1) + S\right)^2 < (S^2 + 2cS + 1)\left(3c^2S - 2c(S^2+1) + S\right)^2.$$

Taking the difference gives

$$4cS(2 + 3S^2 - 9c^4S^2 + 2S^4 + 2c^2(1 - S^2 + S^4)) < 0$$

Given that $c > 0, S > 0$, we have

$$2 + 3S^2 - 9c^4S^2 + 2S^4 + 2c^2(1 - S^2 + S^4) < 0$$

Denoting $X := S^2$ and viewing the above as a quadratic polynomial of X, we have

$$\underbrace{(2c^2+2)X^2 + (3 - 2c^2 - 9c^4)X + (2c^2+2)}_{\textcircled{1}} < 0$$

Using the closed-form minimizer of quadratic polynomial $\textcircled{1}$, after some heavy algebra, one can check the minimum of $\textcircled{1}$ is

$$\frac{(1+3c^2)^2(1-c^2)(7+9c^2)}{8(1+c^2)}$$

which is clearly positive. Contradiction! □

E.3 Proof of Lemma E.2

Lemma E.2. *For all* $0 < c < 1$ *and* $S > 1$*,*

$$(S^2 + 3cS + 2)\sqrt{S^2 - 2cS + 1} - (S^2 - 3cS + 2)\sqrt{S^2 + 2cS + 1} > 0.$$

Proof. Notice that $(S^2 + 3cS + 2) > S^2 + 2 > 0$ and $\sqrt{S^2 \pm 2cS + 1} > 0$. Therefore if $S^2 - 3cS + 2 \leq 0$, we are done.

Otherwise, we prove by contradiction and suppose

$$0 < (S^2 + 3cS + 2)\sqrt{S^2 - 2cS + 1} < (S^2 - 3cS + 2)\sqrt{S^2 + 2cS + 1}.$$

under the condition that $S^2 - 3cS + 2 > 0$.

Squaring everything gives

$$(S^2 + 3cS + 2)^2(S^2 - 2cS + 1) < (S^2 - 3cS + 2)^2(S^2 + 2cS + 1).$$

Taking the difference gives

$$cS(8 + 20S^2 - 36c^2S^2 + 8S^4) < 0$$

Given that $c > 0, S > 0$, we have

$$2 + 5S^2 - 9c^2S^2 + 2S^4 < 0$$

Denoting $X := S^2$ and viewing the above as a quadratic polynomial of X, we have, for $X > 1$,

$$\underbrace{2X^2 + (5 - 9c^2)X + 2}_{\textcircled{2}} < 0$$

The closed-form minimizer of quadratic polynomial $\textcircled{2}$ is $\frac{(9c^2-5)}{4}$. Given that $0 < c < 1$, we must have $-\frac{5}{4} < \frac{9c^2-5}{4} < 1$. Hence the minimizer is not within the feasible domain $(1, \infty)$ of X. Thus the minimum of $\textcircled{2}$ is achieved with $X = 1$ at $9(1 - c^2)$. This is positive. Contradiction! □

E.4 Proof of Fact E.3

Fact E.3. For a quadratic polynomial $Ax^2 + Bx + C$ with $A, B, C > 0$, the minimum value on the domain $x \geq 0$ is C, at $x = 0$. Therefore $Ax^2 + Bx + C > 0$.

Proof. Since $A > 0$, the quadratic polynomial is convex and increasing on the domain $x > -\frac{B}{2A}$. Since $B > 0$ as well, we know $-\frac{B}{2A} < 0$ and hence the quadratic polynomial is strictly increasing on $x > 0$. Therefore the minimum value is achieved when $x = 0$, and we obtain $Ax^2 + Bx + C \geq C > 0$ for all $x \geq 0$. □

F Examples of lazy regions

F.1 Balanced binary classification

We describe the data generation in Section 3.3. The label is uniformly ±1, that is $\mathbb{P}(y_i = +1) = \mathbb{P}(y_i = -1) = 0.5$. We have 10000 positive and negative samples $x_i \sim \mathcal{N}(y_i, 1)$. We consider a logistic regression model $\mathbb{P}(Y = y|x) = \mathbb{I}(y = 1) \cdot \text{Sigmoid}(x + \theta) + \mathbb{I}(y = -1) \cdot (1 - \text{Sigmoid}(x + \theta)) = \frac{1}{1+e^{-y(\theta+x)}}$, where $\theta \in \mathbb{R}$ is the intercept. The gradient with respect to this only trainable parameter is $\frac{\partial \mathcal{L}_i}{\partial \theta} = -y\left(1 - \frac{1}{1+e^{-y(\theta+x)}}\right)$. We set the clipping threshold $R = 0.01$ and the stability constant $\gamma = 0.01$.

F.2 Mean estimation on Gaussian mixture data

We also observe the lazy region issue in the mean estimation problem $\min_\theta \frac{1}{2}\|\theta - x_i\|^2$. Here $\mathbb{P}(x_i \sim \mathcal{N}(4, 1)) = \mathbb{P}(x_i \sim \mathcal{N}(4, 1)) = 0.5$. We have 10000 samples from each Gaussian distribution. The regular minimum is clearly $\sum_i x_i \to 0$, where the regular gradient and AUTO-S clipped gradient vanish. Yet both AUTO-V and Abadi's clipping lose motivation to update the mean estimator on the interval $(-1, 1)$. We set the clipping threshold $R = 0.01$ and the stability constant $\gamma = 0.1$.

G Experiments settings

G.1 Image classification settings

We give the experiments settings for computer vision tasks in Table 1.

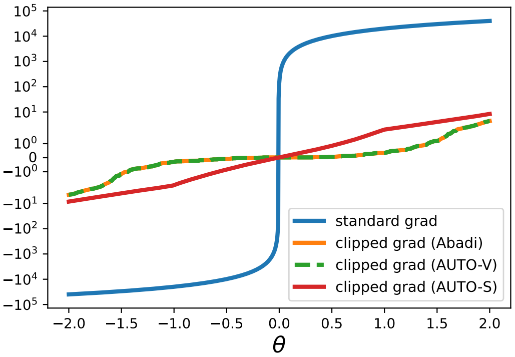

Figure 11: Scalar gradient $\frac{\partial \mathcal{L}}{\partial \theta}$ at each θ.

- **MNIST**: We use the network architecture from [53, 65, 61], with 40 epochs, 512 batch size, 0.5 learning rate (or 0.005 non-DP learning rate), 0.1 clipping threshold, DP-SGD with 0.9 momentum, and without pretraining. This setting is the same as [65].
- **FashionMNIST**: We use the same network architecture as MNIST, with 40 epochs, 2048 batch size, 4 learning rate (or 0.04 non-DP learning rate), DP-SGD with 0.9 momentum, and without pretraining. This setting is the same as [65].
- **CIFAR10 pretrained**: We use the SimCLR model from [13][10], with 50 epochs, 1024 batch size, 4 learning rate (or 0.04 non-DP learning rate), 0.1 clipping threshold, and DP-SGD with 0.9 momentum. The SimCLR model is pretrained on unlabelled ImageNet dataset. After pretraining, we obtain a feature of dimension 4096 on which a linear classifier is trained privately. This setting is the same as [65].
- **ImageNette**: We use the ResNet9 (2.5 million parameters) with Mish activation function [50]. We set 50 epochs, 1000 batch size, 0.0005 learning rate (or 0.000005 non-DP learning rate), 1.5 clipping threshold, and use DP-NAdam, without pretraining. This setting is the same as [35] except we did not apply the learning rate decaying scheduler.
- **CelebA (Smiling and Male and Multi-label)** We use the same ResNet9 as above, with 10 epochs, 500 batch size, 0.001 DP learning rate (or 0.00001 non-DP learning rate), 0.1 clipping threshold, and use DP-Adam, without pretraining. We use the labels 'Smiling' and 'Male' for two binary classification tasks, with cross-entropy loss. For the multi-label task uses a scalar loss by summing up the 40 binary cross-entropy losses from each label.

We refer the code for MNIST, FashionMNIST, CIFAR10, CIFAR10 pretrained to `https://github.com/ftramer/Handcrafted-DP` by [65]. ResNet9 can be found in `https://github.com/cbenitez81/Resnet9`.

Throughout all experiments, we do not apply tricks such as random data augmentation (single or multiple times [16]), weight standardization [58], or parameter averaging [57].

G.2 Sentence classification settings

We experiment on five datasets in Table 2 and Table 3.

- **MNLI(m)** MNLI-matched, the matched validation and test splits from Multi-Genre Natural Language Inference Corpus.
- **MNLI(mm)** MNLI-mismatched, the matched validation and test splits from Multi-Genre Natural Language Inference Corpus.
- **QQP** The Quora Question Pairs2 dataset.
- **QNLI** The Stanford Question Answering dataset.
- **SST2** The Stanford Sentiment Treebank dataset.

[10] See implementation in `https://github.com/google-research/simclr`.

The datasets are processed and loaded from Huggingface [38], as described in https://huggingface.co/datasets/glue. We follow the same setup as [74] and [40]. We refer the interested readers to Appendix G,H,I,K,N of [40] for more details.

We emphasize that our automatic clipping uses exactly the same hyperparameters as the Abadi's clipping in [40], which is released in their Private-Transformers library [11].

Dataset	MNLI(m/mm)	QQP	QNLI	SST2
Epoch	18	18	6	3
Batch size	6000	6000	2000	1000
clipping threshold R	0.1	0.1	0.1	0.1
DP learning rate	5e-4	5e-4	5e-4	5e-4
non-DP learning rate	5e-5	5e-5	5e-5	5e-5
learning rate decay	Yes	Yes	Yes	Yes
AdamW weight decay	0	0	0	0
Max sequence length	256	256	256	256

Table 5: Hyperparameters of automatic clipping and Abadi's clipping, for sentence classification in Table 2 and Table 3, using either RoBERTa base or large.

Notice that we use DP learning rate 5e-4 across tasks for the R-dependent automatic DP-Adam, which is equivalent to R-independent automatic DP-Adam with the same learning rate. We demonstrate that the results are not sensitive to learning rates around the optimal choice. That is, the automatic clipping does not eliminate R at the cost of more difficult tuning of learning rate.

learning rate	1e-4	3e-4	5e-4	8e-4	1e-3
RoBERTa-base	93.92	94.38	94.49	94.72	93.35
RoBERTa-large	95.76	96.21	96.21	96.33	95.99

Table 6: SST2 accuracy with respect to learning rate.

G.3 Table-to-text generation settings

We experiment multiple GPT2 models on E2E dataset from Huggingface [38] in Table 4. We follow the same setup as [40], and our automatic clipping uses exactly the same hyperparameters as the Abadi's clipping in [40], which is released in their Private-Transformer library [12].

Model	GPT2	GPT2 medium	GPT2 large
Epoch	10	10	10
Batch size	1024	1024	1024
clipping threshold R	0.1	0.1	0.1
DP learning rate	2e-3	2e-3	2e-3
non-DP learning rate	2e-4	1e-4	1e-4
learning rate decay	No	No	No
AdamW weight decay	0.01	0.01	0.01
Max sequence length	100	100	100

Table 7: Hyperparameters of automatic clipping and Abadi's clipping, for the E2E generation task in Table 4.

[11] See https://github.com/lxuechen/private-transformers/blob/main/examples/classification/run_wrapper.py

[12] See https://github.com/lxuechen/private-transformers/blob/main/examples/table2text/run.sh

H Figure zoo

H.1 Frequency of clipping

We show that in all sentence classification tasks, Abadi's clipping happens on a large proportion of per-sample gradients. This supports the similarity between Abadi's clipping and AUTO-V in (3.1).

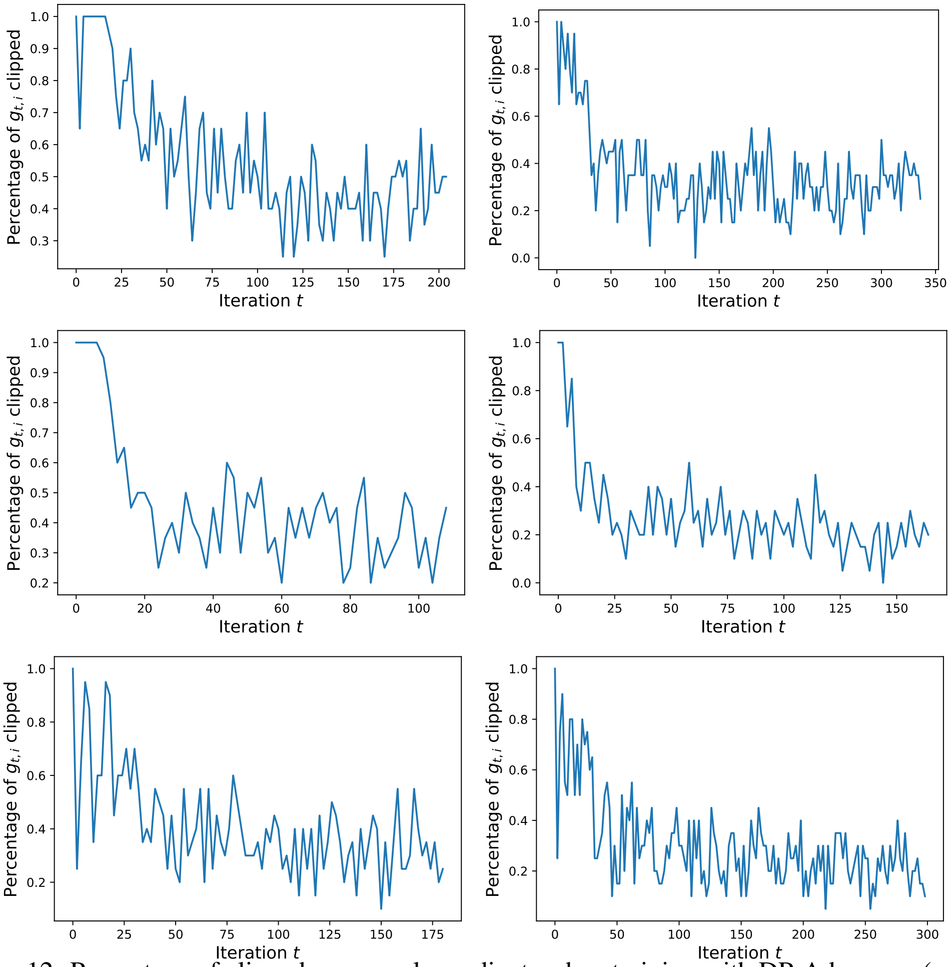

Figure 12: Percentage of clipped per-sample gradients when training with DP-Adam$_{\text{Abadi}}$ ($\epsilon = 3$), as in Section 6.2. Left panel is RoBERTa-base and right panel is RoBERTa-large. Top row: MNLI. Middle row: QNLI. Bottom row: QQP.

We note that for GPT2, GPT2 medium and GPT2 large, empirically in all iterations 100% of the per-sample gradients are clipped by the Abadi's clipping, making the performance of Abadi's clipping equivalent to AUTO-V clipping, as shown in Table 4.

H.2 Stability constant helps AUTO clipping reduce gradient norm

To corroborate our claim in Theorem 6, that the stability γ reduces the gradient norm, we plot the actual gradient norm by iteration.

H.3 Choice of stability constant is robust

We claim in Theorem 6 that, as long as $\gamma > 0$ in our automatic clipping, the asymptotic convergence rate of gradient norm is the same as that by standard non-private SGD. We plot the ablation study of learning rate and the stability constant γ to show that it is easy to set γ: in Table 2 and Table 3,

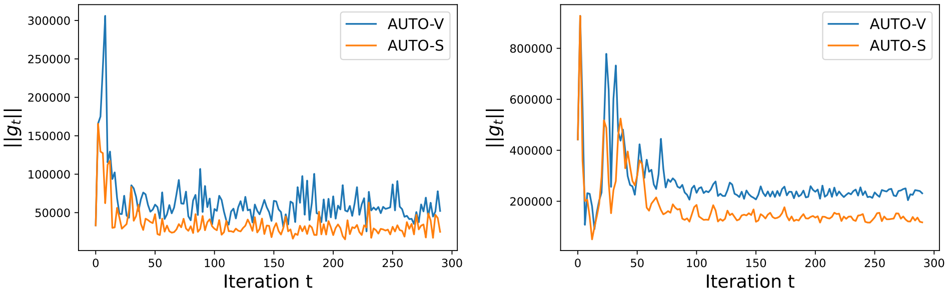

Figure 13: Gradient norm by different automatic clipping methods, on SST2 (left) and MNLI (right), trained with RoBERTa-base.

we adopt learning rate 0.0005, under which a wide range of $0.0001 < \gamma < 1$ gives similar accuracy. Note that the largest good γ is 1000 times bigger than the smallest good γ.

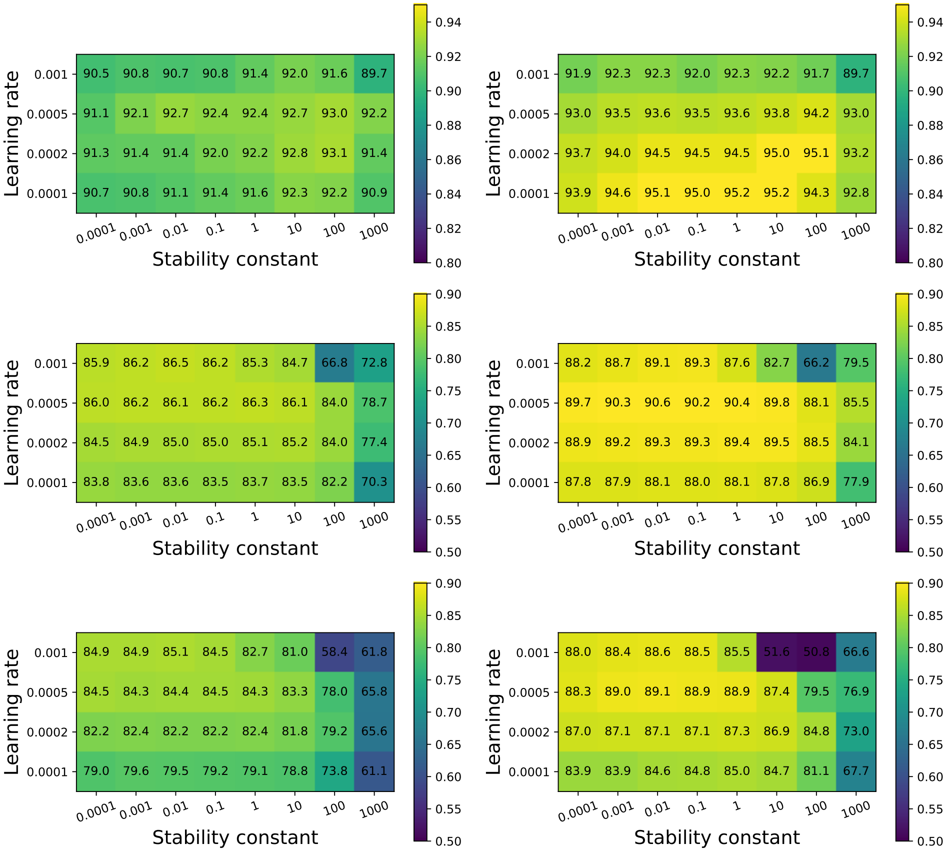

Figure 14: Test accuracy by different stability constant γ and learning rate η in automatic clipping ($\epsilon = 3$). Upper row: SST2 for full 3 epochs. Middle row: QNLI for full 6 epochs. Lower row: QNLI for one epoch. Trained with RoBERTa-base (left) and RoBERTa-large (right).

H.4 Automatic clipping avoids ablation study

We plot the ablation study of learning rate and clipping threshold in Abadi's clipping below. This demonstrates that, AUTO-S clipping only requires 1D grid search to tune the learning rate, avoiding the expensive 2D grid search that is unfortunately necessary for the Abadi's clipping. Hence our automatic clipping can save the tuning effort substantially.

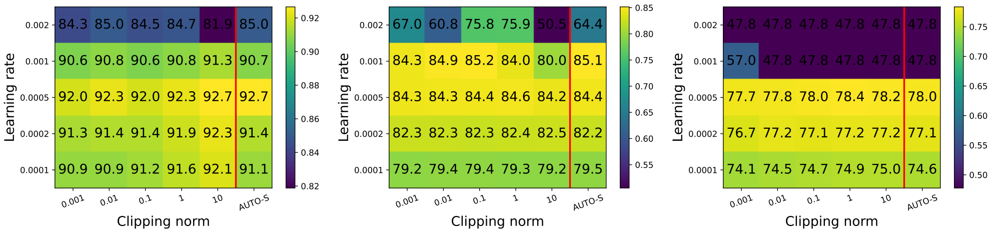

Figure 15: Test accuracy by different clipping threshold R in DP-Adam$_{\text{Abadi}}$ and learning rate η, on SST2 (left, 3 epochs) / QNLI (middle, 1 epoch) / MNLI (right, 1 epoch), $\epsilon = 3$, trained with RoBERTa-base.

I Full table of GPT2 generation task on E2E dataset

This is the extended version of Table 4 on E2E dataset. The performance measures are BLEU [54], ROGUE-L [41], NIST [60], METEOR [4], and CIDEr [67] scores. Here ϵ is accounted by RDP [49], where $\epsilon = 3$ corresponds to 2.68 if accounted by Gaussian DP [18, 7] or to 2.75 if accounted by numerical composition [29], and $\epsilon = 8$ corresponds to 6.77 if accounted by Gaussian DP or to 7.27 if accounted by numerical composition.

Metric	DP guarantee	GPT2 large	GPT2 medium	GPT2							
		full AUTO-S	full AUTO-S	full AUTO-S	full AUTO-V	full [40]	LoRA [32]	RGP [74]	prefix [39]	top2 [40]	retrain [40]
BLEU	$\epsilon = 3$	**64.180**	**63.850**	**61.340**	**61.519**	**61.519**	58.153	58.482	47.772	25.920	15.457
	$\epsilon = 8$	**64.640**	**64.220**	**63.600**	63.189	63.189	**63.389**	58.455	49.263	26.885	24.247
	non-DP	66.840	68.500	69.463	69.463	69.463	69.682	68.328	68.845	65.752	65.731
ROGUE-L	$\epsilon = 3$	**67.857**	**67.071**	**65.872**	65.670	65.670	**65.773**	65.560	58.964	44.536	35.240
	$\epsilon = 8$	**68.968**	**67.533**	**67.073**	66.429	66.429	**67.525**	65.030	60.730	46.421	39.951
	non-DP	70.384	71.458	71.359	71.359	71.359	71.709	68.844	70.805	68.704	68.751
NIST	$\epsilon = 3$	**7.937**	**7.106**	**7.071**	**6.697**	**6.697**	5.463	5.775	5.249	1.510	0.376
	$\epsilon = 8$	**8.301**	**8.172**	**7.714**	7.444	7.444	**7.449**	6.276	5.525	1.547	1.01
	non-DP	8.730	8.628	8.780	8.780	8.780	8.822	8.722	8.722	8.418	8.286
METEOR	$\epsilon = 3$	**0.403**	**0.387**	**0.387**	**0.384**	**0.384**	0.370	0.331	0.363	0.197	0.113
	$\epsilon = 8$	**0.420**	**0.418**	**0.404**	0.400	0.400	**0.407**	0.349	0.364	0.207	0.145
	non-DP	0.460	0.449	0.461	0.461	0.461	0.463	0.456	0.445	0.443	0.429
CIDEr	$\epsilon = 3$	**2.008**	**1.754**	**1.801**	**1.761**	**1.761**	1.581	1.300	1.507	0.452	0.116
	$\epsilon = 8$	**2.163**	**2.081**	**1.938**	1.919	1.919	**1.948**	1.496	1.569	0.499	0.281
	non-DP	2.356	2.137	2.422	2.422	2.422	2.491	2.418	2.345	2.180	2.004

Table 8: Test performance on E2E dataset with GPT2. The best two GPT2 models for each row are marked in bold.

We observe that GPT2 (163 million parameters), GPT2-medium (406 million), and GPT2-large (838 million), Table 4 trained with our automatic clipping consistently perform better in comparison to other methods. In some cases, LoRA trained with Abadi's clipping also demonstrates strong performance and it would be interesting to see how LoRA trained with the automatic clipping will behave.

J Further experiments on CelebA dataset

In this section, we present a complete summary of accuracy results, with DP constraint or not, for the CelebA dataset. We do not apply any data-preprocessing. In the first experiment, we apply a single ResNet on the 40 labels as the multi-task/multi-label learning. In the second experiment, we apply one ResNet on one label. As expected, our automatic DP optimizers have comparable test accuracy to the Abadi's DP optimizers, but we do not need to tune the clipping threshold for each individual task/label. We also notice that, learning different labels separately gives better accuracy than learning all labels together, though at the cost of heavier computational burden.

J.1 Multi-label classification

We apply ResNet9 as in Appendix G.1 on the multi-label classification task. I.e. the output layer has 40 neurons, each corresponding to one sigmoid cross-entropy loss, that are summed to a single loss and all labels are learnt jointly.

Index	Attributes	Abadi's	AUTO-S	Abadi's	AUTO-S	non-DP
		$\epsilon = 3$	$\epsilon = 3$	$\epsilon = 8$	$\epsilon = 8$	$\epsilon = \infty$
0	5 o Clock Shadow	90.64	90.99↑	90.81	91.28↑	93.33
1	Arched Eyebrows	75.15	76.31↑	76.84	77.11↑	81.52
2	Attractive	75.85	76.10↑	77.50	77.74↑	81.15
3	Bags Under Eyes	80.75	81.12↑	82.15	82.13↓	84.81
4	Bald	97.84	97.87↑	98.04	97.98↓	98.58
5	Bangs	92.71	92.68↓	93.46	93.55↑	95.50
6	Big Lips	67.51	67.78↑	68.34	68.44↑	71.33
7	Big Nose	78.01	80.23↑	76.69	80.59↑	83.54
8	Black Hair	81.92	80.95↓	83.33	83.28↓	88.55
9	Blond Hair	92.25	92.38↑	93.52	93.09↓	95.49
10	Blurry	94.91	94.82↓	95.08	94.90↓	95.78
11	Brown Hair	80.13	82.50↑	83.74	83.89↑	87.79
12	Bushy Eyebrows	88.06	88.23↑	89.72	88.80↓	92.19
13	Chubby	94.72	94.54↓	94.54	94.50↓	95.56
14	Double Chin	95.19	95.49↑	95.50	95.51↑	96.09
15	Eyeglasses	97.06	97.64↑	98.32	98.06↓	99.39
16	Goatee	95.68	95.45↓	95.84	95.87↑	97.06
17	Gray Hair	96.77	96.79↑	97.02	97.03↑	98.06
18	Heavy Makeup	84.96	85.70↑	87.58	87.29↓	90.76
19	High Cheekbones	81.46	81.42↓	82.62	82.72↑	86.62
20	Male	92.05	92.17↑	93.32	93.17↓	97.46
21	Mouth Slightly Open	86.20	86.32↑	87.84	88.48↑	93.07
22	Mustache	96.05	95.96↓	96.08	95.99↓	96.74
23	Narrow Eyes	84.90	84.78↓	85.14	85.18↑	86.98
24	No Beard	91.55	91.67↑	92.29	92.45↑	95.18
25	Oval Face	71.26	71.42↑	71.98	71.25↓	74.62
26	Pale Skin	96.09	96.04↓	96.15	96.17↑	96.93
27	Pointy Nose	70.34	72.11↑	72.23	73.01↑	75.68
28	Receding Hairline	91.53	91.37↓	91.75	91.74↓	92.87
29	Rosy Cheeks	93.26	93.02↓	93.56	93.35↓	94.86
30	Sideburns	96.16	96.09↓	96.27	96.46↑	97.44
31	Smiling	86.39	87.08↑	88.87	88.63↓	92.25
32	Straight Hair	76.20	77.95↑	78.78	78.52↓	80.66
33	Wavy Hair	70.30	71.79↑	73.58	73.19↓	79.15
34	Wearing Earrings	80.53	81.52↑	82.29	82.20↓	87.56
35	Wearing Hat	96.99	96.83↓	97.46	97.31↓	98.68
36	Wearing Lipstick	88.95	88.04↓	89.87	90.72↑	93.49
37	Wearing Necklace	84.59	85.83↑	85.93	85.42↓	86.61
38	Wearing Necktie	93.91	93.91–	94.43	94.08↓	96.30
39	Young	81.35	81.21↓	82.18	82.52↑	87.18

Table 9: Accuracy on CelebA dataset with settings in Appendix G.1 from one run. The green arrow indicates AUTO-S is better than Abadi's clipping under the same ϵ; the red arrow indicates otherwise; the black bar indicates the same accuracy.

J.2 Multiple binary classification

For the second experiment, we apply ResNet9 on each label as a binary classification task. I.e. the output layer has 1 neuron and we run 40 different models for all labels separately.

Index	Attributes	Abadi's	AUTO-S	Abadi's	AUTO-S	non-DP
		Single	Single	Multi	Multi	Multi
		$\epsilon = 8$	$\epsilon = 8$	$\epsilon = 8$	$\epsilon = 8$	$\epsilon = \infty$
0	5 o Clock Shadow	92.15	92.29↑	90.81	91.28↑	93.33
1	Arched Eyebrows	81.18	80.19↓	76.84	77.11↑	81.52
2	Attractive	79.31	79.79↑	77.50	77.74↑	81.15
3	Bags Under Eyes	83.52	83.48↓	82.15	82.13↓	84.81
4	Bald	97.89	97.88↓	98.04	97.98↓	98.58
5	Bangs	94.52	94.83↑	93.46	93.55↑	95.50
6	Big Lips	67.32	67.53↑	68.34	68.44↑	71.33
7	Big Nose	82.31	82.36↑	76.69	80.59↑	83.54
8	Black Hair	87.08	86.93↓	83.33	83.28↓	88.55
9	Blond Hair	94.29	94.73↑	93.52	93.09↓	95.49
10	Blurry	94.95	95.20↑	95.08	94.90↓	95.78
11	Brown Hair	87.41	87.19↓	83.74	83.89↑	87.79
12	Bushy Eyebrows	91.23	91.43↑	89.72	88.80↓	92.19
13	Chubby	94.70	94.70–	94.54	94.50↓	95.56
14	Double Chin	95.43	95.43–	95.50	95.51↑	96.09
15	Eyeglasses	98.88	99.14↑	98.32	98.06↓	99.39
16	Goatee	96.12	96.07↓	95.84	95.87↑	97.06
17	Gray Hair	97.48	97.34↓	97.02	97.03↑	98.06
18	Heavy Makeup	88.85	88.72↓	87.58	87.29↓	90.76
19	High Cheekbones	85.66	85.45↓	82.62	82.72↑	86.62
20	Male	95.42	95.70↑	95.53	93.17↓	97.46
21	Mouth Slightly Open	92.67	92.74↑	87.84	88.48↑	93.07
22	Mustache	96.13	96.13–	96.08	95.99↓	96.74
23	Narrow Eyes	85.13	85.13–	85.14	85.18↑	86.98
24	No Beard	94.26	94.58↑	92.29	92.45↑	95.18
25	Oval Face	70.77	73.05↑	71.98	71.25↓	74.62
26	Pale Skin	96.38	96.34↓	96.15	96.17↑	96.93
27	Pointy Nose	71.48	73.37↑	72.23	73.01↑	75.68
28	Receding Hairline	91.51	91.51–	91.75	91.74↓	92.87
29	Rosy Cheeks	93.26	93.35↑	93.56	93.35↓	94.86
30	Sideburns	96.46	96.34↓	96.27	96.46↑	97.44
31	Smiling	90.82	90.87↑	88.87	88.63↓	92.25
32	Straight Hair	79.01	79.01–	78.78	78.52↓	80.66
33	Wavy Hair	77.55	78.83↑	73.58	73.19↓	79.15
34	Wearing Earrings	87.33	87.50↑	82.29	82.20↓	87.56
35	Wearing Hat	98.04	98.11↑	97.46	97.31↓	98.68
36	Wearing Lipstick	92.05	90.46↓	89.87	90.72↑	93.49
37	Wearing Necklace	86.21	86.21–	85.93	85.42↓	86.61
38	Wearing Necktie	95.85	95.94↑	94.43	94.08↓	96.30
39	Young	85.19	84.12↓	82.18	82.52↑	87.18

Table 10: Accuracy on CelebA dataset with settings in Appendix G.1 from one run. 'Single' means each attribute is learned separately as a binary classification task. 'Multi' means all attributes are learned jointly as a multi-label classification task. The green arrow indicates AUTO-S is better than Abadi's clipping under the same ϵ and the same task; the red arrow indicates otherwise; the black bar indicates the same accuracy.

K Code implementation of automatic clipping

Changing Abadi's clipping to automatic clipping is easy in available codebases. One can set the clipping $R = 1$ or any other constant, as explained in Theorem 1 and Theorem 2.

K.1 Opacus

For Opacus [73] version 1.1.2 (latest), we can implement the all-layer automatic clipping by changing Line 399-401 in https://github.com/pytorch/opacus/blob/main/opacus/optimizers/optimizer.py to

```
per_sample_clip_factor = self.max_grad_norm /(per_sample_norms + 0.01)
```

The per-layer automatic clipping requires changing Line 61-63 in https://github.com/pytorch/opacus/blob/main/opacus/optimizers/perlayeroptimizer.py to

```
per_sample_clip_factor =max_grad_norm / (per_sample_norms + 0.01)
```

For older version (<1.0, e.g. 0.15) of Opacus, we can implement the all-layer automatic clipping by changing Line 223-225 in https://github.com/pytorch/opacus/blob/v0.15.0/opacus/utils/clipping.py to

```
per_sample_clip_factor = self.flat_value / (norms[0] + 0.01)
```

or implement the per-layer automatic clipping by changing Line 301-302 in https://github.com/pytorch/opacus/blob/main/opacus/optimizers/perlayeroptimizer.py to

```
per_sample_clip_factor = threshold / (norm + 0.01)
clipping_factor.append(per_sample_clip_factor)
```

K.2 ObJAX

For ObJAX version 1.6.0 (latest), we can implement the automatic clipping in https://github.com/google/objax/blob/master/objax/privacy/dpsgd/gradient.py by changing Line 92 to

```
idivisor = self.l2_norm_clip / (total_grad_norm+0.01)
```

and changing Line 145 to

```
idivisor = self.l2_norm_clip/(grad_norms+0.01)
```

K.3 Private-transformers

To reproduce our experiments for sentence classification and table-to-text generation, we modify the 'private-transformers' codebase of [40]. The modification is in https://github.com/lxuechen/private-transformers/blob/main/private_transformers/privacy_utils/privacy_engine.py, by changing Line 349 to

```
return self.max_grad_norm / (norm_sample + 0.01)
```

and Line 510-512 to

```
coef_sample = self.max_grad_norm * scale / (norm_sample + 0.01)
```