18 Differentially Private (Stochastic) Gradient Descent

This section is based heavily on the course Privacy in Statistics and Machine Learning by Adam Smith and Jonathan Ullman, primarily Lectures 9 and 12 (Spring 2025) [Smi25, SU21a].

We have already seen how to make gradient descent robust by estimating the mean of the individual gradients. To make it private, we will instead add Gaussian noise to that mean. The updates and the final model are then postprocessing of private gradient estimates.

We use the same convex optimization setup as in Theorem 10.1, with constants $R, L > 0$. The domain $\Theta \subseteq \mathbb{R}^d$ is nonempty, closed, and convex, with $\|\theta\|_2 \le R$ for every $\theta \in \Theta$. For every possible record $x \in \mathcal{X}$, the loss $\ell(\cdot; x)$ is convex and differentiable, with $\|\nabla_\theta \ell(\theta; x)\|_2 \le L$. Given $S = \{X_1, \dots, X_n\}$, our goal is to minimize

$$\ell(\theta; S) = \frac{1}{n} \sum_{i=1}^{n} \ell(\theta; X_i), \qquad \theta_S^* \in \arg\min_{\theta \in \Theta} \ell(\theta; S).$$

18.1 Noisy Gradient Descent

Algorithm 11: Private projected gradient descent

Input: A dataset S, a number of iterations T, a step size $\eta > 0$, and a noise level $\sigma > 0$. Choose $\theta_0 \in \Theta$ independently of the data. For $t = 0, \dots, T-1$, draw independent $W_t \sim \mathcal{N}(0, \sigma^2 I_d)$ and compute

$$g_t = \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta \ell(\theta_t; X_i), \qquad \widetilde{g}_t = g_t + W_t,$$

$$\theta_{t+1} = \Pi_\Theta(\theta_t - \eta \widetilde{g}_t),$$

Output: $\hat{\theta} = \frac{1}{T} \sum_{t=0}^{T-1} \theta_t$.

Privacy. For any fixed θ, replacing one record changes the average gradient by at most $2L/n$ in ℓ_2 norm. We can therefore use the Gaussian mechanism at each iteration. The query point θ_t depends on earlier noisy answers, so we need adaptive composition.

Lemma 18.1. *For $0 < \varepsilon \le 1$ and $\delta \in (0, 1)$, Algorithm 11 is (ε, δ)-differentially private if*

$$\sigma \ge \frac{2L\sqrt{T}}{\varepsilon n} \left(1 + \sqrt{2 \log \frac{1}{\delta}}\right).$$

Proof. We use the sharper composition bound for Gaussian noise. For T adaptive queries of ℓ_2-sensitivity at most Δ, each with independent noise $\mathcal{N}(0, \sigma^2 I_d)$, the transcript is $(\varepsilon_{\text{tot}}, \delta)$-DP with

$$\varepsilon_{\text{tot}} = \rho + 2\sqrt{\rho \log(1/\delta)}, \qquad \rho = \frac{T\Delta^2}{2\sigma^2}.$$

This follows by summing the conditional Gaussian privacy losses: each has mean at most $\Delta^2/2\sigma^2$ and a centered Gaussian part with variance at most Δ^2/σ^2. The conditional exponential moment

bound controls their sum under adaptivity. Take $\Delta = 2L/n$; the stated σ makes $\varepsilon_{\text{tot}} \le \varepsilon$. The output is postprocessing of this transcript. □

Accuracy. The noise is unbiased, so its effect on the convergence analysis is especially simple.

Theorem 18.2. *For every fixed dataset S, Algorithm 11 satisfies*

$$\mathbb{E}[\ell(\hat{\theta}; S)] - \ell(\theta_S^*; S) \le \frac{2R^2}{\eta T} + \frac{\eta}{2}(L^2 + d\sigma^2).$$

Proof. Conditional on the previous iterations, the noisy gradient is unbiased and has second moment at most $L^2 + d\sigma^2$. Thus the potential proof of Theorem 10.1 applies in expectation, with the squared-gradient bound L^2 replaced by $L^2 + d\sigma^2$. □

The geometric step in that proof is illustrated in Figure 15. If $p = \Pi_\Theta(y)$, the projection optimality condition gives $\langle y - p\,,\ \theta_S^* - p\rangle \le 0$. Expanding the squared distance therefore gives

$$\|y - \theta_S^*\|_2^2 = \|y - p\|_2^2 + \|p - \theta_S^*\|_2^2 - 2\langle y - p\,,\ \theta_S^* - p\rangle \ge \|p - \theta_S^*\|_2^2.$$

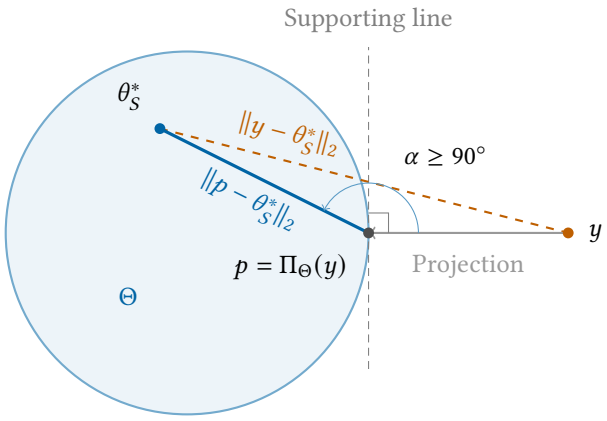


Figure 15: Projection cannot increase distance to $\theta_S^* \in \Theta$. At $p = \Pi_\Theta(y)$, the angle between the outward vector $y - p$ and $\theta_S^* - p$ is at least 90°, so the orange distance is at least the blue distance.

With $\eta = 2R/\sqrt{T(L^2 + d\sigma^2)}$ and σ equal to the threshold in Lemma 18.1, the bound becomes

$$\mathbb{E}[\ell(\hat{\theta}; S)] - \ell(\theta_S^*; S) = O\left(\frac{RL}{\sqrt{T}} + \frac{RL\sqrt{d\log(2/\delta)}}{\varepsilon n}\right).$$

The first term decreases with more iterations; the second is the privacy cost. These are empirical excess-risk bounds. A population-risk guarantee also requires a generalization argument.

18.2 Stochastic Gradients and Clipping

Draw a fresh uniform batch $B_t \subseteq [n]$ of size m, without replacement, and use $g_t = \frac{1}{m}\sum_{i\in B_t} \nabla_\theta \ell(\theta_t; X_i)$. This unbiased estimate has second moment at most L^2, so the same accuracy bound applies using m gradient evaluations per iteration.

The fixed-batch sensitivity is $2L/m$. Hidden sampling gives *privacy amplification*: a mechanism that is $(\varepsilon_0, \delta_0)$-DP on m records gives per-step parameters

$$\varepsilon_{\text{step}} = \log\Big(1 + \frac{m}{n}(e^{\varepsilon_0} - 1)\Big), \qquad \delta_{\text{step}} = \frac{m}{n}\delta_0.$$

For small ε_0, $\varepsilon_{\text{step}} \approx m\varepsilon_0/n$. Compose these bounds across iterations, or use a tighter sampled Gaussian analysis.

For unbounded gradients, *clip each individual's gradient* before averaging:

$$v_{t,i} = \frac{\nabla_\theta \ell(\theta_t; X_i)}{\max\{1, \|\nabla_\theta \ell(\theta_t; X_i)\|_2 / L\}}, \qquad \widetilde{g_t} = \frac{1}{m}\sum_{i\in B_t} v_{t,i} + W_t.$$

Now $\|v_{t,i}\|_2 \le L$, giving sensitivity at most $2L/m$ and the DP-SGD update. Privacy holds without convexity; the accuracy theorem requires convexity and unbiased gradients, and active clipping can introduce bias.

References

[Smi25] Adam Smith. *Lecture 10: Advanced Composition*. Privacy in Statistics and Machine Learning, Spring 2025. Based on materials developed with Jonathan Ullman. Linked as the notes for Lecture 9 in the course schedule. 2025. (Link).

[SU21a] Adam Smith and Jonathan Ullman. *Lecture 12: Private Gradient Descent*. Privacy in Statistics and Machine Learning, Spring 2021. Linked as the notes for Lecture 12 in the Spring 2025 course schedule. 2021. (Link).