CS292A Introduction to Differential Privacy **Fall 2021**

Lecture 12: NoisyGD and NoisySGD (November 12)

Lecturer: Yu-Xiang Wang *Scribes: Xuandong Zhao*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

12.1 Noisy Gradient Descent Mechanism

12.1.1 Algorithm

$$\theta_{t+1} = \theta_t + \eta_t \left[\sum_{i=1}^{n} \nabla \ell_i(\theta_t) + \mathcal{N}(0, \sigma^2 \mathbf{I}_d) \right], \text{for } t = 1, 2, \ldots, T \tag{12.1}$$

As shown in Equation 12.1, the NoisyGD mechanism is straightforward, which simply adds gaussian noise to the gradient. Note that $\sum_{i=1}^{n} \nabla \ell_i(\theta_t)$ is $\nabla f(\theta_t)$, and $\mathcal{N}(0, \sigma^2 \mathbf{I}_d)$ is the noise.

If we set $g_t = \sum_{i=1}^{n} \nabla \ell_i(\theta_t) + \mathcal{N}(0, \sigma^2 \mathbf{I}_d)$, the expected value of g_t is $\mathbb{E}[g_t|\theta_t] = \nabla f(\theta_t)$ and variance is $\mathbb{E}[\|g_t - \mathbb{E}[g_t]\|_{\theta_t}] = d\sigma^2$.

12.1.2 Privacy analysis

Global sensitivity of NoisyGD is L, because ℓ_i is L-lipschitz. Each iteration of NoisyGD is ρ-zCDP with $\rho = \frac{L^2}{2\sigma^2}$. Since NoisyGD is a composition of T Gaussian mechanisms, the whole algorithm of NoisyGD is $T\rho$-zCDP with $\rho_{\text{total}} = \frac{TL^2}{2\sigma^2}$. And we can get that $\frac{\sigma^2}{T} = \frac{L^2}{2\rho}, T = \frac{2\rho\sigma^2}{L^2}$.

12.2 Convergence of NoisyGD

12.2.1 Nonconvex / smooth problems

Lemma 12.1. *(Descent Lemma): For the NoisyGD update: $x_{t+1} = x_t - \eta_t \hat{g}_t$ in smooth/nonconvex case, the convergence guarantee is:*

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[\|\nabla f(x_t)\|^2 \right] \leq \frac{2(f(x_1) - f^*)}{T\eta} + \eta n \beta d \sigma^2$$

Proof. Since $f(x)$ is smooth and use update rule,

$$\begin{aligned} f(x_{t+1}) &\leq f(x_t) + \langle x_{t+1} - x_t, \nabla f(x_t) \rangle + \frac{\beta \|x_{t+1} - x_t\|^2}{2} \\ &= f(x_t) - \eta_t \langle \hat{g}_t, \nabla f(x_t) \rangle + \frac{\beta}{2} \eta_t^2 \|\hat{g}_t\|^2 \end{aligned}$$

We assume $\mathbb{E}\left[\hat{g}_t | x_t\right] = \nabla f(x_t)$ and $\mathbb{E}\left[\|\hat{g}_t - \mathbb{E}[\hat{g}_t]\| | x_t\right] \leq d\sigma^2$. If we set constant learning rate $\eta_t = \eta < \frac{1}{\beta}$ and take conditional expectation on both side,

$$\begin{aligned} \mathbb{E}\left[f(x_{t+1}) | x_t\right] &\leq f(x_t) - \eta_t \|\nabla f(x_t)\|^2 + \frac{\beta}{2} \eta_t^2 \left(\|\nabla f(x_t)\|^2 + d\sigma^2 \right) \\ &= f(x_t) - \eta \|\nabla f(x_t)\|^2 + \frac{\eta}{2} \|\nabla f(x_t)\|^2 + \frac{\eta^2 \beta \sigma^2 d}{2} \\ &= f(x_t) - \frac{\eta}{2} \|\nabla f(x_t)\|^2 + \frac{\eta^2 \beta \sigma^2 d}{2} \end{aligned}$$

Take full expectation on both side,

$$\mathbb{E}\left[f(x_{t+1})\right] \leq \mathbb{E}[f(x_t)] - \frac{\eta}{2} \mathbb{E}\left[\|\nabla f(x_t)\|^2\right] + \frac{\eta^2 \beta \sigma^2 d}{2}$$

Then we add up $t = 1, \ldots, T$

$$\begin{aligned} \mathbb{E}\left[f(x_2)\right] &\leq \mathbb{E}[f(x_1)] - \frac{\eta}{2} \mathbb{E}\left[\|\nabla f(x_1)\|^2\right] + \frac{\eta^2 \beta}{2} \sigma^2 d \\ \mathbb{E}\left[f(x_3)\right] &\leq \mathbb{E}[f(x_2)] - \frac{\eta}{2} \mathbb{E}\left[\|\nabla f(x_2)\|^2\right] + \frac{\eta^2 \beta}{2} \sigma^2 d \\ &\cdots \\ \mathbb{E}\left[f(x_T)\right] &\leq \mathbb{E}[f(x_{T-1})] - \frac{\eta}{2} \mathbb{E}\left[\|\nabla f(x_{T-1})\|^2\right] + \frac{\eta^2 \beta}{2} \sigma^2 d \end{aligned}$$

We finally get

$$\begin{aligned} \mathbb{E}\left[f(x_T)\right] - \mathbb{E}[f(x_1)] &\leq -\frac{\eta}{2} \mathbb{E}\left[\sum_t \|\nabla f(x_t)\|^2\right] + \frac{T\eta^2 \beta}{2} \sigma^2 d \\ \mathbb{E}\left[\frac{1}{T} \sum_t \|\nabla f(x_t)\|^2\right] &\leq \frac{2\left(f(x_1) - f(x^\star)\right)}{T\eta} + \beta \eta n d \sigma^2 \end{aligned}$$

$\square$

Utility bound

We can choose the learning rate $\eta = \min\left\{\frac{1}{n\beta}, \frac{\sqrt{2(f(x_1)-f^*)}}{\sqrt{n\beta d\sigma^2 T}}\right\}$

$$\begin{aligned}\mathbb{E}\left[\min_{t\in[T]} \|\nabla f(x_t)\|^2\right] &\le \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla f(x_t)\|^2\right] \\ &\le \frac{2(f(x_1)-f^*)}{T\eta} + \eta n\beta d\sigma^2 \\ &\le \frac{2(f(x_1)-f^*)}{T}\max\left\{n\beta, \frac{\sqrt{n\beta d\sigma^2 T}}{\sqrt{2(f(x_1)-f^*)}}\right\} + \sqrt{\frac{2n\beta d\sigma^2(f(x_1)-f^*)}{T}} \\ &\le \frac{2n\beta(f(x_1)-f^*)}{T} + 2\sqrt{\frac{2n\beta d\sigma^2(f(x_1)-f^*)}{T}}\end{aligned}$$

Recall that for ρ-zCDP, $\frac{\sigma^2}{T} = \frac{L^2}{2\rho}$, if we substitute it in the second term.

$$\sqrt{\frac{n\beta d(f(x_1)-f^*)L^2}{2\rho}} \asymp \sqrt{\frac{n\beta d(f(x_1)-f^*)L^2}{\epsilon^2/\log\frac{1}{\delta}}}$$

If we substitute it in the first term, the first term becomes $\frac{2n\beta(f(x_1)-f^*)L^2}{2\sigma^2\rho}$. We can make it arbitrarily small by choosing large noise and more number of iterations to get $\sigma^2 \to \infty$. So we can only consider the second term for utility guarantee.

12.2.2 Convex /smooth problems

Following similar analysis as Lemma 12.1 and applying convex property we can get

$$\mathbb{E}\left[f\left(\frac{1}{T}\sum_{t=1}^{T}x_t\right) - f^\star\right] \le \mathbb{E}\left[\frac{1}{T}\sum_t (f(x_t)-f^*)\right] \le \frac{\|x_1-x^*\|^2}{T\eta} + \eta d\sigma^2$$

Utility bound

We can choose the learning rate $\eta = \min\left\{\frac{1}{n\beta}, \frac{\|x_1-x^*\|}{\sqrt{d\sigma^2 T}}\right\}$, where the first apply to GD and the second apply to SGD. Following the same analysis in nonconvex/smooth problems,

$$\frac{\|x_1-x^*\|^2}{T\eta} + \eta d\sigma^2 \le \frac{n\beta\|x_1-x^*\|^2}{T} + \frac{2\|x_1-x^*\|\sqrt{d\sigma^2}}{\sqrt{T}}$$

Substitute $\frac{\sigma^2}{T} = \frac{L^2}{2\rho}$ for ρ-zCDP in the second term, the final utility bound is

$$2\|x_1-x^*\|\sqrt{\frac{dL^2}{\rho}} \asymp \|x_1-x^*\|\frac{\sqrt{dL^2\log\frac{1}{\delta}}}{\epsilon}$$

Note that if we use large T, the first term can be arbitrarily small

$$\frac{n\beta \left\|x_1 - x^*\right\|^2}{T} \leq \left\|x_1 - x^*\right\| \sqrt{\frac{dL^2}{\rho}}$$
$$T \geq \frac{n\beta \left\|x_1 - x^*\right\| \sqrt{\rho}}{L\sqrt{d}} = \mathcal{O}(n\epsilon)$$

12.2.3 Convex / Lipschitz problems

Following similar analysis as Lemma 12.1 and applying convex and Lipschitz property (Refer to notes in CS292F Convex Optimization Lecture 8) we can get

$$\mathbb{E}\left[\frac{1}{T}\sum_t \left(f\left(x_t\right) - f^*\right)\right] \leq \frac{\left\|x_1 - x^*\right\|^2}{T\eta} + \eta\left(\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\left\|\partial f\left(x_t\right)\right\|^2\right] + d\sigma^2\right)$$

Utility bound

By choosing learning rate optimally,

$$\mathbb{E}\left[\frac{1}{T}\sum_t \left(f\left(x_t\right) - f^*\right)\right] \leq \frac{\left\|x_1 - x^*\right\| \sqrt{d\sigma^2 + n^2L^2}}{\sqrt{T}}$$
$$\leq \frac{\left\|x_1 - x^*\right\| nL}{\sqrt{T}} + \left\|x_1 - x^*\right\| \sqrt{\frac{d\sigma^2}{T}},$$

where the first inequality follows f is nL-Lipschitz so that $\frac{1}{T}\sum_{t=1}^{T}\left\|\partial f\left(x_t\right)\right\|^2 \leq n^2L^2$ and the second inequality follows $\sqrt{x^2+y^2} \leq x + y$ for $x, y \geq 0$.

Substitute $\frac{\sigma^2}{T} = \frac{L^2}{2\rho}$ for ρ-zCDP in the second term, the final utility bound is

$$\left\|x_1 - x^*\right\| \sqrt{\frac{dL^2}{\rho}} = \left\|x_1 - x^*\right\| \sqrt{\frac{d\log\frac{1}{\delta}L^2}{\epsilon^2}}$$

Note that we can also use large T to make the first term be arbitrarily small

$$\frac{\left\|x_1 - x^*\right\| nL}{\sqrt{T}} \leq \left\|x_1 - x^*\right\| \sqrt{\frac{dL^2}{\rho}}$$
$$T \geq \frac{n^2L^2\rho}{dL^2} = \mathcal{O}(n^2\epsilon^2)$$

12.2.4 Strongly convex / Lipschitz problems

If f is λ-strongly convex and L-Lipschitz, convergence is even faster[1]. For learning rate $\eta_t = \frac{1}{\lambda t}$,

$$\mathbb{E}\left[f\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right)\right] - f\left(x^*\right) \leq \frac{n^2L^2 + d\sigma^2}{2\lambda T}(1 + \log T)$$

For learning rate $\eta_t = \frac{1}{\lambda(t+1)}$,

$$\begin{aligned}\mathbb{E}\left[f\left(\frac{2}{T(T+1)}\sum_{t=1}^{T} t x_t\right)\right] - f(x^*) &\leq \frac{4\left(n^2L^2 + d\sigma^2\right)}{\lambda(T+1)} \\ &= c\left(\frac{n^2L^2}{\lambda T} + \frac{d\sigma^2}{\lambda T}\right)\end{aligned}$$

Utility bound

Following the same utility analysis, we substitute $\frac{\sigma^2}{T} = \frac{L^2}{2\rho}$ for ρ-zCDP in the second term.

$$\frac{d\sigma^2}{\lambda T} = \frac{dL^2}{\lambda\rho} \asymp \frac{dL^2 \log\frac{1}{\delta}}{\lambda\epsilon^2}$$

Note that we can also use large T to make the first term be arbitrarily small

$$\frac{n^2L^2}{\lambda T} \leq \frac{dL^2}{\lambda\rho}$$

$$T \geq \frac{n^2\rho}{\lambda} \asymp \mathcal{O}(\frac{n^2\epsilon^2}{\lambda})$$

12.2.5 Summary

The advantage of NoisyGD:

- It is more generally applicable
- Results in stronger guarantees
- Do not require exact optimal solution

Function	Utility Bound
Lipschitz+convex	$\frac{\sqrt{d}L\|\theta^*\|\sqrt{\log\left(\frac{1}{\delta}\right)}}{n\epsilon}$
Lipschitz+Strongly convex	$\frac{dL^2\log(1/\delta)}{n\lambda\epsilon^2}$
Lipschitz+Smooth+Nonconvex	$\frac{\sqrt{n\beta dL^2(f(\theta_1)-f^*)\log(1/\delta)}}{n\epsilon}$

Function	Computational Complexity	# of call
Lipschitz+convex	$T \geq \frac{n^2\rho}{\|x_1-x^\star\|d} = \mathcal{O}(n^2\epsilon^2)$	$\mathcal{O}(n^3\epsilon^2)$
Smooth+convex	$T \geq \frac{2n\beta\sqrt{\rho}\|x_1-x^\star\|}{\sqrt{d}L} = \mathcal{O}(n\epsilon)$	$\mathcal{O}(n^2\epsilon)$
Lipschitz+Strongly convex	$T \geq \frac{n^2\rho}{d} = \mathcal{O}(n^2\epsilon^2)$	$\mathcal{O}(n^3\epsilon^2)$

12.3 Noisy Stochastic Gradient Descent Mechanism

12.3.1 Privacy Amplification by Sampling

Lemma 12.2. *(Subsampling Lemma): If $\mathcal{M}$ obeys (ϵ, δ)-DP, then $\mathcal{M} \circ$Subsample obeys (ϵ', δ')-DP with*

$$\delta' = \gamma\delta, \epsilon' = \log(1 + \gamma(e^{\epsilon} - 1)) = \mathcal{O}(\gamma\epsilon)$$

There are two types of sampling schemes for privacy amplification, one is Poisson Sampling and another is Sampling without Replacement.

Poisson Sampling: include datapoint i in the minibatch by sampling from a Bernoulli Distribution with probability γ ($\mathbb{E}[\text{batch size}] = \gamma \cdot n$). Poisson Sampling works well for add/remove.

Random subset: choose a subset with size equal to m from $\{1, \dots, n\}$, so that $\gamma_i = \frac{m}{n}$. Random subset works well for replace-one.

12.3.2 Algorithm

$$\hat{g}_t = \frac{1}{\gamma}\left(\sum_{i \in \text{Batch}} \nabla \ell_i(\theta_t) + \mathcal{N}(0, \sigma^2 \mathbf{I}_d)\right) \tag{12.2}$$

$$\theta_{t+1} = \theta_t + \eta_t \hat{g}_t, \text{for } t = 1, 2, \dots, T \tag{12.3}$$

The privacy analysis is just simply adds up RDP. NoisySGD satisfy ρ-tCDP with $\rho = \frac{\gamma^2 L^2 T}{2\sigma^2}$. In the "nice" regimes of the conversion $\rho \asymp \epsilon^2 \log \frac{1}{\delta}$.

12.3.3 Utility analysis

The estimate of the gradient is

$$\frac{1}{\gamma}\left(\sum_{i \in \text{Batch}} \nabla \ell_i(\theta_t) + \mathcal{N}(0, \sigma^2 \mathbf{I}_d)\right)$$

It has same bounds as before, but noise gets larger: $d\sigma^2 \to \frac{d\sigma^2}{\gamma^2}$. Then we have:

$$\mathbb{E}\left[\|\hat{g} - \mathbb{E}[\hat{g}]\|^2\right] = \frac{d\sigma^2}{\gamma^2} + \frac{nL^2}{\gamma}$$

For the convex/smooth case $\frac{2\eta\beta\|x_1 - x^\star\|^2}{T} + \sqrt{\frac{d\|x_1 - x^\star\|^2\sigma^2}{T}}$, if we substitute it in the second term

$$\begin{aligned}
\sqrt{\frac{\|x_1 - x^\star\|^2}{T}\left(\frac{d\sigma^2}{\gamma^2} + \frac{nL^2}{\gamma}\right)} &\leq \|x_1 - x^\star\|\left(\sqrt{\frac{d\sigma^2}{T\gamma^2}} + \sqrt{\frac{nL^2}{\gamma T}}\right) \\
&= \|x_1 - x^\star\|\sqrt{\frac{dL^2}{\rho}}
\end{aligned}$$

References

[1] Simon Lacoste-Julien, Mark Schmidt, Francis Bach ”A simpler approach to obtaining an O(1/t) convergence rate for the projected stochastic subgradient method” *arXiv preprint arXiv:1212.2002 (2012).*