12.1 Noisy Gradient Descent Mechanism

12.1.1 Algorithm

\[ \theta_{t+1} = \theta_t + \eta \left( \sum_{i=1}^n \nabla \ell_i(\theta_t) + \mathcal{N}(0, \sigma^2 I_d) \right) \text{, for } t = 1, 2, \ldots, T \]

(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 \( \mathcal{N}(0, \sigma^2 I_d) \) is the noise.

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

12.1.2 Privacy analysis

Global sensitivity of NoisyGD is \( L \), because \( \ell_i \) is \( L \)-lipschitz. Each iteration of NoisyGD is \( \rho \)-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 \( \sigma^2 = \frac{T^2 L^2}{4\rho} = \frac{2\sigma^2}{L^2} \).

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:

\[ \frac{1}{T} \sum_{t=1}^T \mathbb{E} \left[ \| \nabla f(x_t) \|^2 \right] \leq \frac{2(f(x_1) - f^*)}{T\eta} + \eta \beta d\sigma^2 \]
Proof. Since $f(x)$ is smooth and use update rule,

$$f(x_{t+1}) \leq f(x_t) + \langle x_{t+1} - x_t, \nabla f(x_t) \rangle + \frac{\beta}{2} \|x_{t+1} - x_t\|^2$$

$$= f(x_t) - \eta_t \langle \hat{g}_t, \nabla f(x_t) \rangle + \frac{\beta}{2} \eta_t^2 \|\hat{g}_t\|^2$$

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

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

$$= f(x_t) - \eta \|\nabla f(x_t)\|^2 + \frac{\eta}{2} \|\nabla f(x_t)\|^2 + \frac{\eta^2 \beta \sigma^2 d}{2}$$

$$= f(x_t) - \frac{\eta}{2} \|\nabla f(x_t)\|^2 + \frac{\eta^2 \beta \sigma^2 d}{2}$$

Take full expectation on both side,

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

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

$$\mathbb{E} [f(x_2)] \leq \mathbb{E} [f(x_1)] - \frac{\eta}{2} \mathbb{E} [\|\nabla f(x_1)\|^2] + \frac{\eta^2 \beta \sigma^2 d}{2}$$

$$\mathbb{E} [f(x_3)] \leq \mathbb{E} [f(x_2)] - \frac{\eta}{2} \mathbb{E} [\|\nabla f(x_2)\|^2] + \frac{\eta^2 \beta \sigma^2 d}{2}$$

$$\ldots$$

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

We finally get

$$\mathbb{E} [f(x_T)] - \mathbb{E} [f(x_1)] \leq - \frac{\eta}{2} \sum_t \mathbb{E} [\|\nabla f(x_t)\|^2] + \frac{T \eta^2 \beta \sigma^2 d}{2}$$

$$\mathbb{E} \left[ \frac{1}{T} \sum_t \|\nabla f(x_t)\|^2 \right] \leq \frac{2(f(x_1) - f(x^*))}{T\eta} + \beta \eta \delta \sigma^2$$
Utility bound

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

\[
E \left[ \min_{t \in [T]} \| \nabla f(x_t) \|^2 \right] \leq \frac{1}{T} \sum_{t=1}^{T} E \left[ \| \nabla f(x_t) \|^2 \right] \\
\leq \frac{2(\hat{f}(x_1) - f^*)}{T\eta} + \eta n \beta \sigma^2 \\
\leq \frac{2(\hat{f}(x_1) - f^*)}{T} \max \left\{ n\beta, \frac{\sqrt{n\beta \sigma^2 T}}{\sqrt{2(\hat{f}(x_1) - f^*)}} \right\} + \frac{2n\beta \sigma^2 (\hat{f}(x_1) - f^*)}{T} \\
\leq \frac{2n\beta (\hat{f}(x_1) - f^*)}{T} + 2 \sqrt{\frac{2n\beta \sigma^2 (\hat{f}(x_1) - f^*)}{T}}.
\]

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

\[
\sqrt{\frac{n\beta d(\hat{f}(x_1) - f^*) L^2}{2\rho}} \approx \sqrt{\frac{n\beta d(\hat{f}(x_1) - f^*) L^2}{\epsilon^2/\log \frac{1}{\delta}}}
\]

If we substitute it in the first term, the first term becomes \( \frac{2n\beta (\hat{f}(x_1) - f^*) L^2}{2\sigma^2 \rho} \). 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

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

Utility bound

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

\[
\frac{\|x_1 - x^*\|^2}{T\eta} + \eta \sigma^2 \leq \frac{n\beta \|x_1 - x^*\|^2}{T} + 2 \frac{\|x_1 - x^*\| \sqrt{\sigma^2}}{\sqrt{T}}
\]

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

\[
2 \frac{\|x_1 - x^*\| \sqrt{\frac{dL^2}{\rho}}}{\epsilon} \leq \frac{\|x_1 - x^*\| \sqrt{dL^2 \log \frac{1}{\delta}}}{\epsilon}
\]
Note that if we use large $T$, the first term can be arbitrarily small
\[
\frac{n\beta \|x_1 - x^*\|^2}{T} \leq \|x_1 - x^*\| \sqrt{\frac{dL^2}{\rho}}
\]
\[
T \geq \frac{n\beta \|x_1 - x^*\| \sqrt{\rho}}{L \sqrt{d}} = O(n\epsilon)
\]

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

Utility bound

By choosing learning rate optimally,
\[
\mathbb{E} \left[ \frac{1}{T} \sum_{t=1}^{T} (f(x_t) - f^*) \right] \leq \left\| x_1 - x^* \right\| \frac{nL}{\sqrt{T}} + \|x_1 - x^*\| \sqrt{\frac{d\sigma^2}{T}},
\]

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

Substitute $\frac{d^2}{2T} = \frac{L^2}{2\rho}$ for $\rho$-zCDP in the second term, the final utility bound is
\[
\left\| x_1 - x^* \right\| \sqrt{\frac{d\sigma^2}{T}} = \left\| x_1 - x^* \right\| \sqrt{\frac{d \log \frac{1}{\delta}}{\epsilon^2}}
\]

Note that we can also use large $T$ to make the first term be arbitrarily small
\[
\frac{nL}{\sqrt{T}} \leq \left\| x_1 - x^* \right\| \sqrt{\frac{dL^2}{\rho}}
\]
\[
T \geq \frac{nL^2 \rho}{dL^2} = O(n^2 \epsilon^2)
\]

12.2.4 Strongly convex / Lipschitz problems

If $f$ is $\lambda$-strongly convex and $L$-Lipschitz, convergence is even faster[1]. For learning rate $\eta_t = \frac{1}{\lambda t}$,
\[
\mathbb{E} \left[ f \left( \frac{1}{T} \sum_{t=1}^{T} x_t \right) \right] - f(x^*) \leq \frac{n^2 L^2 + d\sigma^2}{2AT} (1 + \log T)
\]
For learning rate $\eta_t = \frac{1}{\lambda(T+1)}$,

$$
E \left[ f \left( \frac{2}{T(T+1)} \sum_{t=1}^{T} tx_t \right) \right] - f(x^*) \leq \frac{4}{\lambda(T+1)} \left( n^2 L^2 + d\sigma^2 \right)
$$

$$
= c \left( \frac{n^2 L^2}{\lambda T} + \frac{d\sigma^2}{\lambda T} \right)
$$

Utility bound

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

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

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

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

$$
T \geq \frac{n^2 \rho}{\lambda} \approx O\left( \frac{n^2 \epsilon^2}{\lambda} \right)
$$

12.2.5 Summary

The advantage of NoisyGD:

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

<table>
<thead>
<tr>
<th>Function</th>
<th>Utility Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lipschitz+convex</td>
<td>$\frac{\sqrt{dL} |\theta^*| \sqrt{\log \left( \frac{1}{\delta} \right)}}{\sqrt{\frac{dL^2 \rho}{\lambda \rho}}}$</td>
</tr>
<tr>
<td>Lipschitz+Strongly convex</td>
<td>$\frac{dL^2 \log(1/\delta)}{\sqrt{n^2 \rho \log(1/\delta)}}$</td>
</tr>
<tr>
<td>Lipschitz+Smooth+Nonconvex</td>
<td>$\frac{\sqrt{dL^2 (f(\theta^<em>) - f^</em>) \log(1/\delta)}}{\sqrt{n^2 \rho \log(1/\delta)}}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Function</th>
<th>Computational Complexity</th>
<th># of call</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lipschitz+convex</td>
<td>$T \geq \frac{n^2 \rho}{|x_1 - x^*|}$ = $O(n^2 \epsilon^2)$</td>
<td>$O(n^3 \epsilon^2)$</td>
</tr>
<tr>
<td>Smooth+convex</td>
<td>$T \geq \frac{n^2 \rho}{\sqrt{dL}}$ = $O(n \epsilon)$</td>
<td>$O(n^3 \epsilon)$</td>
</tr>
<tr>
<td>Lipschitz+Strongly convex</td>
<td>$T \geq \frac{n^2 \rho}{\sqrt{dL}}$ = $O(n^2 \epsilon^2)$</td>
<td>$O(n^3 \epsilon^2)$</td>
</tr>
</tbody>
</table>
12.3 Noisy Stochastic Gradient Descent Mechanism

12.3.1 Privacy Amplification by Sampling

Lemma 12.2. (Subsampling Lemma): If $M$ obeys $(\epsilon, \delta)$-DP, then $M \circ \text{Subsample}$ obeys $(\epsilon', \delta')$-DP with

$$\delta' = \gamma \delta, \epsilon' = \log(1 + \gamma(\epsilon' - 1)) = O(\gamma \epsilon)$$

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 $\gamma \cdot \mathbb{E}[$batch size$] = \gamma \cdot n$. Poisson Sampling works well for add/remove.

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

12.3.2 Algorithm

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

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

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

12.3.3 Utility analysis

The estimate of the gradient is

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

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

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

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

$$\sqrt{\frac{\|x_1 - x^*\|^2}{T} \left( \frac{\sigma^2}{\gamma^2} + \frac{nL^2}{\gamma} \right)} \leq \|x_1 - x^*\| \left( \sqrt{\frac{\sigma^2}{T\gamma^2}} + \sqrt{\frac{nL^2}{\gamma T}} \right)$$

$$= \|x_1 - x^*\| \sqrt{\frac{dL^2}{\rho}}$$
References