Differential Privacy for Machine Learning

Master IASD, Université PSL

February 2024

PSL UNIVERSITÉ PARIS

Recommended Readings

Rererences used for this lecture:

- Differentially Private Empirical Risk Minimization: Efficient Algorithms and Tight Error Bounds – R. Bassily, A. Smith & A. Thakurta. FOCS 2014.
- Differentially Private Empirical Risk Minimization – R. Chaudhuri, C. Monteleoni & A. Sarwate. JMLR 2011
- Understanding Machine Learning: From Theory to Algorithms – Shalev-Shwartz, Shai, and Shai Ben-David. Cambridge University Press, 2014. Chapters 12, 14.

Table of Contents

Problem Formulation

- Dataset $\mathcal{D} = \{d_1, \ldots, d_n\}$.
- Parameter set $\mathcal{C} \subseteq \mathbb{R}^p$ closed and convex. $\theta \in \mathcal{C}$.
- Loss function $\ell(\theta; d_i)$ - loss incurred by θ on d_i.
- Empirical Risk $\mathcal{L}(\theta; \mathcal{D}) = \sum_{i=1}^{n} \ell(\theta; d_i)$.
- Regularized Empirical Risk $\mathcal{L}(\theta; \mathcal{D}) = \sum_{i=1}^{n} \ell(\theta; d_i) + r(\theta)$, where $r(\cdot)$ is a regularization function that is independent of data.
- (Regularized) Empirical Risk Minimization (ERM):

$$\min_{\theta \in \mathcal{C}} \mathcal{L}(\theta; \mathcal{D}).$$

ERM Examples

Linear regression:

- $d_i = (x_i, y_i)$, $x_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$, $\theta \in \mathbb{R}^p$.
- Squared loss: $\ell(\theta; d_i) = (x_i^T\theta - y_i)^2$.
- Ridge regression: $r(\theta) = \|\theta\|_2^2$, Lasso: $r(\theta) = \|\theta\|_1$.

Logistic regression:

- $d_i = (x_i, y_i)$, $x_i \in \mathbb{R}^p$, $y_i \in \{-1, +1\}$, $\theta \in \mathbb{R}^p$.
- Cross-entropy loss: $\ell(\theta; d_i) = -y_i \log(p_i) - (1 - y_i) \log(1 - p_i)$, $p_i = 1/(1 + e^{-x_i^T\theta})$.

ERM Examples

2-layered neural network

- $d_i = (x_i, y_i)$, $x_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}^k$ is one-hot encoding of label in $\{1, \ldots, k\}$
- $\theta = (W_1, b_1, W_2, b_2)$, $W_1 \in \mathbb{R}^{n \times h}$, $b_1 \in \mathbb{R}^h$, $W_2 \in \mathbb{R}^{h \times k}$, $b_2 \in \mathbb{R}^k$.
- Cross-entropy loss: $\ell(\theta; d_i) = \sum_{i=1}^{n} \sum_{j=1}^{k} -y_{ij} \log(p_{ij})$,
  $p_i = \sigma(W_2^T \sigma(W_1^T x_i + b_1) + b_2) \in \mathbb{R}^k$, activation function σ.

ERM Examples

Maximum Likelihood Estimation (MLE)

$$\theta_{MLE} = \arg\max_{\theta} \mathbb{P}(\mathcal{D}|\theta) = \arg\max_{\theta} \prod_{i=1}^{n} \mathbb{P}(d_i|\theta) = \arg\min_{\theta} \sum_{i=1}^{n} \ell(\theta; d_i)$$

- Negative log-likelihood loss function: $\ell(\theta; d_i) = -\log \mathbb{P}(d_i|\theta)$.

Properties of Loss Functions

Lipschitz Continuity

$\ell : \mathcal{C} \times \mathcal{D} \to \mathbb{R}$ is L-Lipschitz if for all $\theta_1, \theta_2 \in \mathcal{C}$ and $d \in \mathcal{D}$,

$$|\ell(\theta_1; d) - \ell(\theta_2; d)| \leq L\|\theta_1 - \theta_2\|_2.$$

If ℓ is differentiable, then we have the following equivalent condition.

$$\|\nabla \ell(\theta)\|_2 \leq L, \text{ for all } \theta \in \mathcal{C}.$$

Examples:

- Is the squared-loss function lipschitz?
- Is the logistic-loss function lipschitz?

Properties of Loss Functions

Convexity

$\ell : \mathcal{C} \times \mathcal{D} \to \mathbb{R}$ is convex if for all $\theta_1, \theta_2 \in \mathcal{C}$ and $\alpha \in [0, 1]$,

$$\ell(\alpha\theta_1 + (1-\alpha)\theta_2; d) \leq \alpha\ell(\theta_1; d) + (1-\alpha)\ell(\theta_2; d).$$

If ℓ is differentiable, then we have the following equivalent condition.

$$\ell(\theta_2) \geq \ell(\theta_1) + \nabla\ell(\theta_1)^T(\theta_2 - \theta_1).$$

Alternatively, $(\nabla\ell(\theta_2) - \nabla\ell(\theta_1))^T(\theta_2 - \theta_1) \geq 0$ for all $\theta_1, \theta_2 \in \mathcal{C}$.
Examples:

- Is the squared-loss function convex?
- Is the loss incurred by a 2-layer neural network convex?

Properties of Loss Functions

Strong Convexity

$\ell : \mathcal{C} \times \mathcal{D} \to \mathbb{R}$ is λ-strongly convex if for all $\theta_1, \theta_2 \in \mathcal{C}$ and $\alpha \in [0, 1]$,

$$\ell(\alpha\theta_1 + (1-\alpha)\theta_2; d) \leq \alpha\ell(\theta_1; d) + (1-\alpha)\ell(\theta_2; d) - \lambda\alpha(1-\alpha)\|\theta_1 - \theta_2\|_2^2.$$

If ℓ is differentiable, then we have the following equivalent condition.

$$\ell(\theta_2) \geq \ell(\theta_1) + \nabla\ell(\theta_1)^T(\theta_2 - \theta_1) + \frac{\lambda}{2}\|\theta_2 - \theta_1\|^2.$$

Alternatively, $(\nabla\ell(\theta_2) - \nabla\ell(\theta_1))^T(\theta_2 - \theta_1) \geq \lambda\|\theta_2 - \theta_1\|^2$ for all $\theta_1, \theta_2 \in \mathcal{C}$.
Examples:

- Is the squared-loss function strongly convex?
- If ℓ is convex, then $\ell'(\theta; d) = \ell(\theta; d) + \frac{\lambda}{2}\|\theta\|_2^2$ is λ-strongly convex.

Properties of Loss Functions

Margin-based Loss

For $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ where $x \in \mathbb{R}^p, y \in \mathbb{R}$, the loss function $\ell : \mathcal{C} \times \mathcal{D} \to \mathbb{R}$ is called a margin-based loss function if $\ell(\theta; (x_i, y_i)) = \ell_m(y_i x_i^T \theta)$ for some function $\ell_m : \mathbb{R} \to \mathbb{R}$.

Examples:

- Hinge loss for Support Vector Machines (SVM) : $\ell_m(t) = \max\{0, 1 - t\}$.

$$\text{SVM}: \quad \min_\theta \sum_{i=1}^n \max\{0, 1 - y_i x_i^T \theta\} + \lambda \|\theta\|_2^2.$$

- Cross entropy loss with single layered neural network.

Properties of Loss Functions

Margin-based Loss

For $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ where $x \in \mathbb{R}^p, y \in \mathbb{R}$, the loss function $\ell : \mathcal{C} \times \mathcal{D} \to \mathbb{R}$ is called a margin-based loss function if $\ell(\theta; (x_i, y_i)) = \ell_m(y_i x_i^T \theta)$ for some function $\ell_m : \mathbb{R} \to \mathbb{R}$.

Example: Generalized Linear Model (GLM)

$$\mathbb{P}(y|x) \propto \exp\left(\frac{y x^T \theta^* - \Phi(x^T \theta^*)}{c(\sigma)}\right).$$

- $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, $x_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$.
- $\Phi : \mathbb{R} \to \mathbb{R}$ link function, $c(\sigma)$ scale parameter.
- Negative log-likelihood loss: $\ell(\theta; (x, y)) = -y x^T \theta^* + \Phi(x^T \theta^*)$.

Projected Gradient Descent

Projected Gradient Descent

Input: Dataset $\mathcal{D}$, loss function ℓ, convex set $\mathcal{C}$.
Input: learning rate η, iterations T, initialization $\theta_1 \in \mathcal{C}$.

1. For $t = 1, \ldots, T$ do

$$\begin{aligned} g_t &\leftarrow \nabla \mathcal{L}(\theta_t; \mathcal{D}), \\ u_t &\leftarrow \theta_t - \eta g_t, \\ \theta_{t+1} &\leftarrow \Pi_{\mathcal{C}}(u_t). \end{aligned}$$

2. $\theta_{GD} \leftarrow \frac{1}{T} \sum_{t=1}^{T} \theta_t$.

Output: θ_{GD}.

Projected Gradient Descent

PGD Convergence

Let $\mathcal{L}$ be convex and L-Lipschitz. Set $\eta = \frac{\|\mathcal{C}\|}{L\sqrt{T}}$. Then $\mathcal{L}(\theta_{GD};\mathcal{D}) \leq \mathcal{L}(\theta^*;\mathcal{D}) + \frac{2L\|\mathcal{C}\|}{\sqrt{T}}$.

Proof: Since $\mathcal{L}$ is convex, $\mathcal{L}(\theta^*) \geq \mathcal{L}(\theta_t) + g_t^T(\theta^* - \theta_t)$. Hence,

$$
\begin{aligned}
\mathcal{L}(\theta_t) - \mathcal{L}(\theta^*) \leq \frac{1}{\eta}\eta g_t^T(\theta_t - \theta^*) &\leq \frac{1}{2\eta}\left(\|\eta g_t\|^2 + \|\theta_t - \theta^*\|^2 - \|\theta_t - \theta^* - \eta g_t\|^2\right) \\
&= \frac{\eta\|g_t\|^2}{2} + \frac{1}{2\eta}\left(\|\theta_t - \theta^*\|^2 - \|u_t - \theta^*\|^2\right) \\
&\leq \frac{\eta L^2}{2} + \frac{1}{2\eta}\left(\|\theta_t - \theta^*\|^2 - \|\theta_{t+1} - \theta^*\|^2\right).
\end{aligned}
$$

Projected Gradient Descent

Proof: (continued)

$$
\begin{aligned}
\mathcal{L}(\theta_{GD}) - \mathcal{L}(\theta^*) = \mathcal{L}\left(\frac{1}{T}\sum_{t=1}^{T}\theta_t\right) - \mathcal{L}(\theta^*) &\leq \frac{1}{T}\sum_{t=1}^{T}\mathcal{L}(\theta_t) - \mathcal{L}(\theta^*) \\
&= \frac{1}{T}\sum_{t=1}^{T}(\mathcal{L}(\theta_t) - \mathcal{L}(\theta^*)) \\
&\leq \frac{\eta L^2}{2} + \frac{1}{2\eta T}\sum_{t=1}^{T}\left(\|\theta_t - \theta^*\|^2 - \|\theta_{t+1} - \theta^*\|^2\right) \\
&= \frac{\eta L^2}{2} + \frac{1}{2\eta T}\left(\|\theta_1 - \theta^*\|^2 - \|\theta_{T+1} - \theta^*\|^2\right) \\
&\leq \frac{\eta L^2}{2} + \frac{\|\mathcal{C}\|^2}{\eta T}.
\end{aligned}
$$

$\square$

Table of Contents

DP-ERM

- **Neighboring datasets:** Datasets $\mathcal{D}$ and $\mathcal{D}'$ are said to be neighboring if they differ in exactly one data point d_i.
- **ERM Mechanism:** An ERM mechanism M takes a dataset $\mathcal{D}$ as input and outputs a random $\theta \in \mathcal{C}$.

(ε, δ)–Differentially Private ERM

An ERM mechanism M is (ε, δ)–DP if, for all neighboring $\mathcal{D}$ and $\mathcal{D}'$, and any $\mathcal{S} \subset \mathcal{C}$,

$$\mathbb{P}(M(\mathcal{D}) \in \mathcal{S}) \leq e^{\varepsilon}\, \mathbb{P}(M(\mathcal{D}') \in \mathcal{S}) + \delta$$

where $\mathbb{P}$ refers to the randomization in M.

DP-ERM Performance

- **Excess empirical risk** of an ERM mechanism M is defined as,

$$\mathbb{E}[\mathcal{L}(M(\mathcal{D});\mathcal{D}) - \mathcal{L}(\theta^*;\mathcal{D})],$$

  where $\theta^* = \arg\min_{\theta \in \mathcal{C}} \mathcal{L}(\theta;\mathcal{D})$, and the expectation is over randomness in M.
- We can measure the performance of a DP-ERM mechanism M by upper bounding its worst-case excess empirical risk over all possible datasets $\mathcal{D}$.
- Alternatively, we can use a high probability tail bound of the following form.

$$\mathbb{P}(\mathcal{L}(M(\mathcal{D});\mathcal{D} \leq \mathcal{L}(\theta^*;\mathcal{D}) + t) \geq \ 1 - \delta.$$

DP-ERM with Exponential Mechanism

DP-ERM with Exponential Mechanism (M_{exp})

Input: Dataset $\mathcal{D}$, loss function ℓ, privacy parameter ε, convex set $\mathcal{C}$.
Input: Lipschitz constant L.

1. Sample θ^{priv} such that

$$\mathbb{P}(M_{exp}(\mathcal{D}) = \theta) \propto \exp\left(-\frac{\varepsilon}{2L\|\mathcal{C}\|_2}\mathcal{L}(\theta; \mathcal{D})\right).$$

Output: θ^{priv}

Privacy Guarantee of DP-ERM with Exponential Mechanism

Theorem

If the loss function ℓ is L-Lipschitz, then M_{exp} is ε-DP.

Proof Sketch:

$$
\begin{aligned}
u(\mathcal{D}, \theta) &= \mathcal{L}(\theta^*; \mathcal{D}) - \mathcal{L}(\theta; \mathcal{D}), \\
\Delta_u &= \max_{\theta \in \mathcal{C}} \max_{\substack{\mathcal{D}, \mathcal{D}': \\ d(D, D')=1}} |u(\mathcal{D}, \theta) - u(\mathcal{D}', \theta)| \\
&= \max_{\theta \in \mathcal{C}} \max_{\substack{\mathcal{D}, \mathcal{D}': \\ d(D, D')=1}} |(\ell(\theta; d_i) - \ell(\theta)^*; d_i) - (\ell(\theta; d_i') - \ell(\theta)^*; d_i')| \\
&= \max_{\theta \in \mathcal{C}} 2L\|\theta - \theta^*\|_2 \\
&= 2L\|\mathcal{C}\|_2.
\end{aligned}
$$

Utility Guarantee of DP-ERM with Exponential Mechanism

Theorem

If the loss function ℓ is L-Lipschitz then $\mathbb{E}[M_{exp}(\mathcal{D})] \leq \mathcal{L}(\theta^*; \mathcal{D}) + \mathcal{O}\left(\frac{L\|\mathcal{C}\|_2}{\varepsilon}\right)$.

Proof Sketch: Define $S_t = \{\theta : \mathcal{L}(\theta; \mathcal{D}) \leq \mathcal{L}(\theta^*; \mathcal{D}) + t\} \subset \mathcal{C}$.

$$\frac{\mathbb{P}(S_{2t}^c)}{\mathbb{P}(S_t)} = \frac{\int_{\theta \in S_{2t}^c} \exp\left(-\frac{\varepsilon}{2L\|\mathcal{C}\|_2}\mathcal{L}(\theta; \mathcal{D})\right) d\theta}{\int_{\theta \in S_t} \exp\left(-\frac{\varepsilon}{2L\|\mathcal{C}\|_2}\mathcal{L}(\theta; \mathcal{D})\right) d\theta} \leq \frac{\int_{\theta \in S_{2t}^c} \exp\left(-\frac{\varepsilon}{2L\|\mathcal{C}\|_2}(\mathcal{L}(\theta^*; \mathcal{D}) + 2t)\right) d\theta}{\int_{\theta \in S_t} \exp\left(-\frac{\varepsilon}{2L\|\mathcal{C}\|_2}(\mathcal{L}(\theta^*; \mathcal{D}) + t)\right) d\theta}$$

$$= \exp\left(-\frac{t\varepsilon}{2L\|\mathcal{C}\|_2}\right) \frac{\mathrm{Vol}(S_{2t}^c)}{\mathrm{Vol}(S_t)}.$$

$$\therefore \mathbb{P}(S_{2t}^c) \leq \exp\left(-\frac{t\varepsilon}{2L\|\mathcal{C}\|_2}\right) \frac{\mathrm{Vol}(\mathcal{C})}{\mathrm{Vol}(S_t)}.$$

Utility Guarantee of DP-ERM with Exponential Mechanism

Proof Sketch: (continued) Choose $t_0 > 0$ big enough so that $\mathrm{Vol}(S_{t_0}) > c' \,\mathrm{Vol}(\mathcal{C})$ for some $c' > 0$. How big should t_0 be? User Lipschitz continuity.

$$
\begin{aligned}
\mathbb{E}[\mathcal{L}(\theta^{priv};\mathcal{D}) - \mathcal{L}(\theta^*;\mathcal{D})] &= \int_0^\infty \mathbb{P}(\mathcal{L}(\theta^{priv};\mathcal{D}) - \mathcal{L}(\theta^*;\mathcal{D}) \geq t)dt \\
&= \int_0^{t_0} + \int_{t_0}^\infty \mathbb{P}(\mathcal{L}(\theta^{priv};\mathcal{D}) - \mathcal{L}(\theta^*;\mathcal{D}) \geq t)dt \\
&\leq t_0 + \int_{t_0}^\infty \exp\left(-\frac{t\varepsilon}{2L\|\mathcal{C}\|_2}\right)\frac{\mathrm{Vol}(\mathcal{C})}{\mathrm{Vol}(S_t)}dt \\
&< t_0 + \frac{1}{c'}\int_{t_0}^\infty \exp\left(-\frac{t\varepsilon}{2L\|\mathcal{C}\|_2}\right)dt \\
&= t_0 + \frac{2L\|\mathcal{C}\|_2}{\varepsilon c'}\exp\left(-\frac{t_0\varepsilon}{2L\|\mathcal{C}\|_2}\right).
\end{aligned}
$$

Choose the best t_0 by differentiating.

How to sample for θ^{priv} ?

Simple example: Linear regression

- $d_i = (x_i, y_i)$, $x_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$, $\theta \in \mathbb{R}^p$.
- Define $Y = [y_1 \ldots y_n]^T \in \mathbb{R}^n$ and $X = [x_1^T \ldots x_n^T]^T \in \mathbb{R}^{n \times p}$.
- $\mathcal{L}(\theta; \mathcal{D}) = \|Y - X\theta\|^2$, $\theta^* = (X^T X)^{-1} X^T Y$.

Remark

$$\mathbb{P}(M_{exp}(\mathcal{D}) = \theta) \propto \exp\left(-\frac{\varepsilon}{2L\|\mathcal{C}\|_2}\|Y - X\theta\|^2\right).$$

The above probability distribution is identical to the following

$$M_{exp}(\mathcal{D}) \sim MVN\left(\theta^*, \frac{\varepsilon}{2L\|\mathcal{C}\|_2} X^T X\right).$$

How to sample for θ^{priv} ?

- For general ERM, it is much harder to sample θ^{priv} from the exact exponential distribution for ε-DP.
- Approximations exist in the form of Markov Chain Monte Carlo (MCMC) methods.

DP-ERM with Output Perturbation

DP-ERM with Output Perturbation (M_{out})

Input: Dataset $\mathcal{D}$, loss function ℓ, privacy parameter ε, convex set $\mathcal{C}$.
Input: Lipschitz constant L, strong convexity parameter λ.

1. Compute $\theta^* = \arg\min \ell(\theta; d)$.
2. Sample $\theta_n \in \mathcal{C} \subseteq \mathbb{R}^p$ such that $\mathbb{P}(\theta_n) \propto \exp\left(-\frac{\lambda}{4L\varepsilon}\|\theta_n\|_2\right)$.
3. $\theta^{priv} \leftarrow \theta^* + \theta_n$.

Output: θ^{priv}.

Privacy Guarantee of DP-ERM with Output Perturbation

Theorem

If the loss function ℓ is λ-strongly convex and L-Lipschitz, then M_{out} is ε-DP.

Proof Sketch: Define $\theta^*(\mathcal{D}) = \arg\min_{\theta \in \mathcal{C}} \mathcal{L}(\theta; D)$. Take any two neighboring datasets $\mathcal{D}$ and $\mathcal{D}'$. By Lipschitzness, we have the following inequality.

$$
\begin{aligned}
\mathcal{L}(\theta^*(\mathcal{D}'); \mathcal{D}) - \mathcal{L}(\theta^*(\mathcal{D}); \mathcal{D}) = & \underbrace{\mathcal{L}(\theta^*(\mathcal{D}'); \mathcal{D}') - \mathcal{L}(\theta^*(\mathcal{D}); \mathcal{D}')}_{\leq 0} + \ell(\theta^*(\mathcal{D}'); d_i) - \mathcal{L}(\theta^*(\mathcal{D}'); d_i') \\
& - (\ell(\theta^*(\mathcal{D}); d_i) - \mathcal{L}(\theta^*(\mathcal{D}); d_i')) \\
\leq & \; 2L \|\theta^*(\mathcal{D}') - \theta^*(\mathcal{D})\|_2.
\end{aligned}
$$

By strong convexity, $\mathcal{L}(\theta^*(\mathcal{D}'); \mathcal{D}) - \mathcal{L}(\theta^*(\mathcal{D}); \mathcal{D}) \geq \frac{\lambda}{2} \|\theta^*(\mathcal{D}') - \theta^*(\mathcal{D})\|_2^2$. Combining with the above, $\|\theta^*(\mathcal{D}') - \theta^*(\mathcal{D})\| \leq \frac{4L}{\lambda}$. The rest is similar to proof of ε-DP for Laplace Mechanism.

Utility Guarantee of DP-ERM with Output Perturbation

Theorem

If the loss function ℓ is λ-strongly convex and L-Lipschitz, then $\mathbb{E}[M_{out}(\mathcal{D})] \leq \mathcal{L}(\theta^*; \mathcal{D}) + \mathcal{O}\left(\frac{\lambda p}{L\varepsilon}\right)$.

Proof Sketch:

$$\begin{aligned}
\mathbb{E}[M_{out}(\mathcal{D})] - \mathcal{L}(\theta^*; \mathcal{D}) &\leq \mathbb{E}[\mathcal{L}(\theta^{priv}; \mathcal{D}) - \mathcal{L}(\theta^*; \mathcal{D})] \\
&\leq L\mathbb{E}[\|\theta^{priv} - \theta^*\|_2] \\
&= L\mathbb{E}[\|\theta_n\|_2] \\
&= \mathcal{O}\left(\frac{\lambda p}{L\varepsilon}\right).
\end{aligned}$$

DP-ERM with Objective Perturbation

DP-ERM with Objective Perturbation (M_{obj})

Input: Dataset $\mathcal{D}$, loss function ℓ, privacy parameter ε, convex set $\mathcal{C}$.
Input: Strong convexity parameter λ, Bounds $c_x, c_y, c_{\ell'}, c_{\ell''}$.

1. Compute $\varepsilon' \leftarrow \varepsilon - 2\log\left(1 + \frac{c_x^2 c_y^2 c_{\ell''}}{\lambda}\right)$.
2. If $\varepsilon' > 0$, then $\gamma \leftarrow 0$. Else, $\gamma \leftarrow \frac{c_x^2 c_y^2 c_{\ell''}}{e^{\varepsilon/4} - 1}$ and $\varepsilon' \leftarrow \varepsilon/2$.
3. Sample $w \in \mathcal{C} \subseteq \mathbb{R}^p$ such that $\mathbb{P}(w) \propto \exp\left(-\frac{\varepsilon' \|w\|_2}{2 c_x c_y c_{\ell'}}\right)$.
4. $\theta^{priv} \leftarrow \arg\min_{\theta \in \mathcal{C}} \mathcal{L}(\theta; \mathcal{D}) + w^T\theta + \gamma\|\theta\|_2^2$.

Output: θ^{priv}.

Privacy Guarantee of DP-ERM with Objective Perturbation

Theorem

Suppose the loss function $\mathcal{L}(\theta; \mathcal{D}) = \sum_{i=1}^{n} \ell(y_i x_i^T \theta) + r(\theta)$.
Let ℓ *be twice differentiable and convex and* r *be* λ*-strongly convex.*
Let $\|x_i\| \leq c_x, |y_i| \leq c_y, |\ell'(\cdot)| \leq c_{\ell'}, |\ell''(\cdot)| \leq c_{\ell''}$.
Then M_{obj} *is* ε*-DP.*

Proof Sketch: Since $\theta^{priv} := \arg\min_{\theta \in \mathcal{C}} \mathcal{L}(\theta; \mathcal{D}) + w^T\theta + \gamma\|\theta\|_2^2$, we get the following first-order optimiality condition.

$$w(\theta^{priv}; \mathcal{D}) = -\sum_{i=1}^{n} y_i \ell'_m(y_i x_i^T \theta^{priv}) x_i - \nabla r(\theta^{priv}) - \gamma\theta^{priv}.$$

Let J_w denote the Jacobian of the function that maps θ to w according to the above equation. $J_w(\theta; \mathcal{D}) = -\sum_{i=1}^{n} y_i^2 \ell''_m(y_i x_i^T \theta^{priv}) x_i x_i^T - \nabla^2 r(\theta^{priv}) - \gamma$.

$$\frac{\mathbb{P}(\theta^{priv}|\mathcal{D})}{\mathbb{P}(\theta^{priv}|\mathcal{D}')} = \frac{\mathbb{P}(w|\mathcal{D})}{\mathbb{P}(w'|\mathcal{D}')} \frac{|\det(J_w(\theta^{priv}; \mathcal{D}))|}{|\det(J_w(\theta^{priv}; \mathcal{D}'))|}.$$

Privacy Guarantee of DP-ERM with Objective Perturbation

Proof Sketch: (continued)
Let

$$\begin{aligned} A &= J_w(\theta^{priv}; \mathcal{D}'), \\ E &= J_w(\theta^{priv}; \mathcal{D}) - J_w(\theta^{priv}; \mathcal{D}') \\ &= y_i'^2 \ell_m''(y_i' x_i'^T \theta^{priv}) x_i' x_i'^T - y_i^2 \ell_m''(y_i x_i^T \theta^{priv}) x_i x_i^T. \end{aligned}$$

Observe that E has rank at most 2. Moreover,

$$\begin{aligned} |\lambda_1(E)| + |\lambda_2(E)| &\leq |y_i'^2 \ell_m''(y_i' x_i'^T \theta^{priv})| \|x_i'\|^2 + |y_i^2 \ell_m''(y_i x_i^T \theta^{priv})| \|x_i\|^2 \\ &\leq 2 c_x^2 c_y^2 c_{\ell''}. \end{aligned}$$

Hence, $|\lambda_1(E)| \cdot |\lambda_2(E)| \leq c'^2$. Since r is λ-strongly convex, A is $(\lambda + \gamma)$-strongly convex. Hence, $|\lambda_j(A^{-1}E)| \leq |\lambda_j(E)|/(\lambda + \gamma)$ for $j = 1, 2$.

Privacy Guarantee of DP-ERM with Objective Perturbation

Proof Sketch: (continued)

$$
\begin{aligned}
\left|\frac{\det(J_w(\theta^{priv};\mathcal{D}))}{\det(J_w(\theta^{priv};\mathcal{D}'))}\right| &= \frac{|\det(A+E)|}{|\det(A)|} \\
&= |\det(I + A^{-1}E)| \\
&= |(1+\lambda_1(A^{-1}E))(1+\lambda_2(A^{-1}E))| \\
&= |1+\lambda_1(A^{-1}E)+\lambda_2(A^{-1}E)+\lambda_1(A^{-1}E)\lambda_2(A^{-1}E)| \\
&\le 1 + \frac{2c_x^2c_y^2c_{\ell''}}{\lambda+\gamma} + \frac{c_x^4c_y^4c_{\ell''}^2}{(\lambda+\gamma)^2} = \left(1+\frac{c_x^2c_y^2c_{\ell''}}{\lambda+\gamma}\right)^2 .
\end{aligned}
$$

Case I: $\varepsilon' := \varepsilon - 2\log\left(1+\frac{c_x^2c_y^2c_{\ell''}}{\lambda}\right) > 0$. It follows that $\left(1+\frac{c_x^2c_y^2c_{\ell''}}{\lambda+\gamma}\right)^2 \le e^{\varepsilon-\varepsilon'}$.

Case II: $\varepsilon' := \varepsilon - 2\log\left(1+\frac{c_x^2c_y^2c_{\ell''}}{\lambda}\right) > 0$. Here, $\left(1+\frac{c_x^2c_y^2c_{\ell''}}{\lambda+\gamma}\right)^2 = e^{\varepsilon/2} = e^{\varepsilon-\varepsilon'}$.

Privacy Guarantee of DP-ERM with Objective Perturbation

Proof Sketch: (continued)

$$\|w - w'\| = \|y_i'\ell_m'(y_i'x_i'^T\theta^{priv})x_i' - y_i\ell_m'(y_ix_i^T\theta^{priv})x_i\| \le 2c_xc_yc_{\ell'}.$$

$$\frac{\mathbb{P}(w|\mathcal{D})}{\mathbb{P}(w'|\mathcal{D}')} = \frac{e^{-\frac{\varepsilon'\|w\|_2}{2c_xc_yc_{\ell'}}}}{e^{-\frac{\varepsilon'\|w'\|_2}{2c_xc_yc_{\ell'}}}} = e^{\frac{\varepsilon'(\|w'\|_2 - \|w\|_2)}{2c_xc_yc_{\ell'}}} \le e^{\varepsilon'}.$$

$$\begin{aligned}
\therefore \frac{\mathbb{P}(\theta^{priv}|\mathcal{D})}{\mathbb{P}(\theta^{priv}|\mathcal{D}')} &= \frac{\mathbb{P}(w|\mathcal{D})}{\mathbb{P}(w'|\mathcal{D}')}\frac{|\det(J_w(\theta^{priv};\mathcal{D}))|}{|\det(J_w(\theta^{priv};\mathcal{D}'))|} \\
&\le e^{\varepsilon-\varepsilon'}e^{\varepsilon'} \\
&= e^{\varepsilon}.
\end{aligned}$$

$\square$

Utility Guarantee of DP-ERM with Objective Perturbation

Theorem

Under the assumptions of the previous theorem, assume the setting with $\gamma = 0$. Then $\mathbb{E}[M_{obj}(\mathcal{D})] \leq \mathcal{L}(\theta^*; \mathcal{D}) + \mathcal{O}\left(\frac{c_x c_y c_{\ell'} p}{\varepsilon \lambda}\right)$.

Proof Sketch:
Observe that $\nabla \mathcal{L}(\theta^*) = 0$ and $\nabla \mathcal{L}(\theta^{priv}) + w = 0$. From strong convexity,

$$\|\theta^* - \theta^{priv}\|^2 \leq \frac{1}{\lambda}(\nabla \mathcal{L}(\theta^*) - \nabla \mathcal{L}(\theta^{priv}))^T (\theta^* - \theta^{priv}) \leq \frac{1}{\lambda} w^T (\theta^* - \theta^{priv})$$

Therefore, $\|\theta^* - \theta^{priv}\| \leq \frac{1}{\lambda}\|w\|$. From the definition of θ^{priv}, we have,

$$\mathcal{L}(\theta^{priv}; \mathcal{D}) + w^T \theta^{priv} \leq \mathcal{L}(\theta^*; \mathcal{D}) + w^T \theta^* \implies \mathcal{L}(\theta^{priv}; \mathcal{D}) - \mathcal{L}(\theta^*; \mathcal{D}) \leq w^T (\theta^* - \theta^{priv}).$$

Utility Guarantee of DP-ERM with Objective Perturbation

Proof Sketch: (continued)

$$\mathcal{L}(\theta^{priv};\mathcal{D}) - \mathcal{L}(\theta^*;\mathcal{D}) \leq \|w\|\|\theta^* - \theta^{priv}\| \leq \frac{1}{\lambda}\|w\|^2.$$

Since $\mathbb{P}(w) \propto \exp\left(-\frac{\varepsilon'\|w\|_2}{2c_x c_y c_{\ell'}}\right)$, $\mathbb{E}[\|w\|^2] = \mathcal{O}(2c_x c_y c_{\ell'}\sqrt{p}/\varepsilon')$

$$\mathbb{E}[M_{obj}(\mathcal{D})] \leq \mathcal{L}(\theta^*;\mathcal{D}) + \mathcal{O}\left(\frac{c_x c_y c_{\ell'} p}{\varepsilon\lambda}\right)$$

$\square$

DP-ERM with Gradient Perturbation

DP-GD

Input: Dataset $\mathcal{D}$, loss function ℓ, privacy parameters (ε, δ), convex set $\mathcal{C}$.
Input: learning rate η, iterations T, initialization $\theta_1 \in \mathcal{C}$.

1. Set $\sigma = \frac{L\sqrt{2T\log(1.25T/\delta)}}{\varepsilon}$.
2. For $t = 1, \ldots, T$ do

$$\begin{aligned} \widetilde{g}_t &\leftarrow \nabla \mathcal{L}(\theta_t; \mathcal{D}) + \mathcal{N}\left(0, \sigma^2 \mathcal{I}_p\right), \\ u_t &\leftarrow \theta_t - \eta \widetilde{g}_t, \\ \theta_{t+1} &\leftarrow \Pi_{\mathcal{C}}(u_t). \end{aligned}$$

3. $\theta_{GD} \leftarrow \frac{1}{T}\sum_{t=1}^{T} \theta_t$.

Output: θ_{GD}.

Privacy Guarantee of DP-ERM with Gradient Perturbation

Theorem

Let $\mathcal{L}$ be convex and L-Lipschitz. Then M_{GD} is (ε, δ)-DP.

Proof Sketch: Let Observe that ℓ_2 sensitivity of g_t is L. From Gaussian mechanism, each iteration is $(\varepsilon/\sqrt{T}, \delta/T)$-DP. Applying advanced composition gives the desired result. $\square$

Utility Guarantee of DP-ERM with Gradient Perturbation

PGD Convergence

Let $\mathcal{L}$ be convex and L-Lipschitz. Set $\eta = \frac{\|\mathcal{C}\|}{\sqrt{T(L^2+p\sigma^2)}}$. Then
$\mathbb{E}[\mathcal{L}(\theta_{GD};\mathcal{D})] \leq \mathcal{L}(\theta^*;\mathcal{D}) + \frac{2L\|\mathcal{C}\|\sqrt{L^2+p\sigma^2}}{\sqrt{T}}$.

Proof Sketch:

$$
\begin{aligned}
\mathbb{E}[\eta\widetilde{g}_t^T(\theta_t - \theta^*)] &= \mathbb{E}_{\widetilde{g}_{1:T}}[\eta\widetilde{g}_t^T(\theta_t - \theta^*)] \\
&= \mathbb{E}_{\widetilde{g}_{1:t}}[\eta\widetilde{g}_t^T(\theta_t - \theta^*)] \\
&= \mathbb{E}_{\widetilde{g}_{1:t-1}}[\mathbb{E}_{\widetilde{g}_{1:t}}[\eta\widetilde{g}_t^T(\theta_t - \theta^*)|\widetilde{g}_{1:t-1}]] \\
&= \mathbb{E}_{\widetilde{g}_{1:t-1}}[\mathbb{E}_{\widetilde{g}_{1:t}}[\eta\widetilde{g}_t|\widetilde{g}_{1:t-1}]^T(\theta_t - \theta^*)] \\
&= \mathbb{E}_{\widetilde{g}_{1:t-1}}[\eta g_t^T(\theta_t - \theta^*)] \\
&= \mathbb{E}[\eta g_t^T(\theta_t - \theta^*)].
\end{aligned}
$$

Utility Guarantee of DP-ERM with Gradient Perturbation

Proof Sketch: (continued)
Similar to the proof of PGD convergence, we get

$$
\begin{aligned}
\mathbb{E}[\mathcal{L}(\theta_t)] - \mathcal{L}(\theta^*) &\leq \frac{1}{\eta}\mathbb{E}[\eta g_t^T(\theta_t - \theta^*)] \\
&= \frac{1}{\eta}\mathbb{E}[\eta \widetilde{g}_t^T(\theta_t - \theta^*)] \\
&\leq \frac{\eta \mathbb{E}[\|g_t\|^2]}{2} + \frac{1}{2\eta}\left(\mathbb{E}[\|\theta_t - \theta^*\|^2] - \mathbb{E}[\|\theta_{t+1} - \theta^*\|^2]\right) \\
&\leq \frac{\eta(L^2 + p\sigma^2)}{2} + \frac{1}{2\eta}\left(\|\theta_t - \theta^*\|^2 - \|\theta_{t+1} - \theta^*\|^2\right). \square
\end{aligned}
$$

Utility Guarantee of DP-ERM with Gradient Perturbation

Proof Sketch: (continued)

$$
\begin{aligned}
\mathbb{E}[\mathcal{L}(\theta_{GD})] - \mathcal{L}(\theta^*) &\leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\mathcal{L}(\theta_t) - \mathcal{L}(\theta^*)\right] \\
&\leq \frac{\eta(L^2 + p\sigma^2)}{2} + \frac{1}{2\eta T}\sum_{t=1}^{T}\left(\|\theta_t - \theta^*\|^2 - \|\theta_{t+1} - \theta^*\|^2\right) \\
&= \frac{\eta(L^2 + p\sigma^2)}{2} + \frac{1}{2\eta T}\left(\|\theta_0 - \theta^*\|^2 - \|\theta_T - \theta^*\|^2\right) \\
&\leq \frac{\eta(L^2 + p\sigma^2)}{2} + \frac{\|\mathcal{C}\|^2}{\eta T}.
\end{aligned}
$$

$\square$