Trustworthy Machine Learning From an Optimization Lens

Meisam Razaviyayn

Lecture 10: DP Optimization - Regularization and Objective Perturbation

razaviya@usc.edu

Recap: Optimization and Privacy

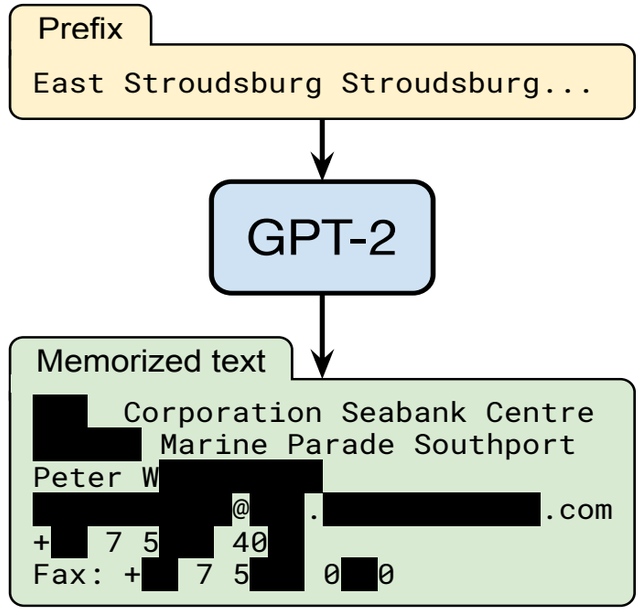
- We solve training/optimization problems that depend on data

$$\min_{\mathbf{w}} \frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{w}, \mathbf{z}_i)$$

- The optimizer w^* may contain a lot of information
  - This high-dimensional w^* can encode a lot of info about data

Model
+
Individual's name

[Fredrikson et al 2015]



[Carlini et al 2021]

- Attacks: Membership Inference, Model Inversion and Reconstruction, Training Data Extraction

Recap: (Pure) differential privacy

Let $\epsilon > 0$ and X be the set of possible datasets. A randomized algorithm $M(\cdot): X \to O$ is said to be $\epsilon -$ differentially private if

$$\Pr(M(D) \in \Omega) \leq e^{\epsilon} \Pr(M(D') \in \Omega)$$ for **all** $\Omega \subseteq O$ and **all** neighboring datasets $D, D' \in X$

- Proposed by Dwork, McSherry, Nissim, and. Smith [2017 Godel Prize]
- Property of the algorithm and not a particular output
- $M(\cdot)$ can even be public; only the randomness of the algorithm should be private
- Smaller ϵ means more privacy
- It hold even if the adversary has arbitrary auxiliary information
- Hypothesis testing viewpoint

Recap: (Approximate) differential privacy

- **Motivation**: Pure DP (δ=0) is often too restrictive

- **Example:** Pure DP does not allow simply using additive Gaussian noise (why?)
  - Gaussian noise has infinite tail
  - To satisfy Pure DP, the probability ratio must be bounded **everywhere**

$$\frac{\Pr(f(D)+z=x)}{\Pr(f(D')+z=x)}=?$$

- $(\epsilon,\delta)-DP$ (Approximate DP) is a relaxation of $\epsilon-DP$ (Pure DP)

Definition: Let $\epsilon,\delta>0$ and X be the set of possible datasets. A randomized algorithm $M(\cdot): X \to O$ is said to be $(\epsilon,\delta)-$ differentially private if

$\Pr(M(D)\in\Omega) \le e^{\epsilon}\Pr(M(D')\in\Omega)+\delta$ for all $\Omega \subseteq O$ and all neighboring datasets $D, D' \in X$

Recap: (Approximate) differential privacy

Definition: Let $\epsilon, \delta > 0$ and X be the set of possible datasets. A randomized algorithm $M(\cdot): X \to O$ is said to be $(\epsilon, \delta) -$ differentially private if

$\Pr(M(D) \in \Omega) \leq e^{\epsilon} \Pr(M(D') \in \Omega) + \delta$ for all $\Omega \subseteq O$ and all neighboring datasets $D, D' \in X$

Connection to pure DP

- **Definition**: Consider two fixed datasets D, D', and a randomized mechanism M. The privacy loss random variable (PLRV) draws an outcome o from $M(D)$ and outputs $\ln\left(\frac{P(M(D)=o)}{P(M(D')=o)}\right)$. In other words, the random variable takes the value of $\ln\left(\frac{P(M(D)=o)}{P(M(D')=o)}\right)$ with probability $P(M(D) = o)$.

- **Theorem:** If privacy loss is bounded by ϵ with probability $\geq 1 - \delta$, then the algorithm is $(\epsilon, \delta) -$ DP

Converse? Proof

Recap: Gaussian Mechanism

Definition: Let $f: X \to R^k$. The $\ell_2 -$ sensitivity of f is defined as

$$\Delta_2 = \sup_{D,D'} \| f(D) - f(D') \|_2$$

where the supremum is taken over all neighboring datasets D and D'

Recall: Zero-mean Gaussian distribution with variance σ^2: $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right)$

Theorem: Let $f: X \to R^k$ have the $\ell_2 -$ sensitivity Δ_2 and

$$M(D) = f(D) + (Z_1, .., Z_k)$$

Where $Z_1, \ldots, Z_k$ are iid Gaussian with variance $\sigma = \frac{\Delta_2 \sqrt{2 \ln \frac{1.25}{\delta}}}{\epsilon}$. Then $M(\cdot)$ is $(\epsilon, \delta) -$ DP.

Example: *Mean estimation*: If data is bounded ($\left|\left|x_i\right|\right|_2 \leq c$), the sensitivity of the mean (assuming neighboring datasets defined by replacement) is $\Delta_2 = 2c/n$

Recap: Properties of approximate DP

Post-Processing: Let $M: X \to O$ be $(\epsilon, \delta) -$ DP and let $G: O \to T$ be an arbitrary (potentially randomized) mapping. Then, $G(M(\cdot))$ is also $(\epsilon, \delta) -$ DP.

Use cases: Integer optimization/decisions, projections involved, etc.

Group Privacy: Let $M: X \to O$ be $(\epsilon, \delta) -$ DP and let D and D' be two datasets that differ in k entries. Then, for any $\Omega \in O$, we have $\Pr(M(D) \in \Omega) \leq \mathrm{e}^{k\epsilon} \Pr(M(D') \in \Omega) + \delta \frac{e^{k\epsilon}-1}{e^{\epsilon}-1}$.

Proof?

Recap: Composition Theorems

Basic Adaptive Composition: Let $M = (M_1, .., M_T)$ be a sequence of algorithms where M_i is $(\epsilon_i, \delta_i) -$ DP.

The algorithms may be chosen adaptively. Then, $M(\cdot)$ is $(\sum_i \epsilon_i, \sum_i \delta_i) -$ DP.

What does adaptive algorithm mean?

Set initial state s_0
for $t = 1, \dots, T$
    $M_t \leftarrow Pick\_Alg(s_0, \dots, s_{t-1})$
    $s_t \leftarrow M_t(D)$
return $(s_1, \dots, s_T)$

- $Pick\_Alg(\cdot)$ may be randomized
- Proof sketch
  - Condition on the randomness of the $Pick\_Alg(\cdot)$
  - Unroll carefully
- Linear scaling in $T\epsilon$. Can we improve it?

Recap: Composition Theorems

Basic Composition: Let $M = (M_1, .., M_T)$ be a sequence of algorithms where M_i is $(\epsilon_i, \delta_i) -$ DP. The algorithms may be chosen adaptively. Then, $M(\cdot)$ is $(\sum_i \epsilon_i, \sum_i \delta_i) -$ DP.

Advanced Composition: Let $M = (M_1, .., M_T)$ be a sequence of $(\epsilon, \delta) -$ DP algorithms (may be chosen adaptively). Then, for any $\delta' > 0$, $M(\cdot)$ is $(\epsilon', T\delta + \delta') -$ DP where $\epsilon' = \epsilon\sqrt{2T \ln\left(\frac{1}{\delta'}\right)} + T\epsilon(e^{\epsilon} - 1)$.

The two composition theorems do not contradict each other; they hold simultaneously

Corollary: Let $M = (M_1, .., M_T)$ be a sequence of $(\epsilon, \delta) -$ DP algorithms (may be chosen adaptively). To guarantee target privacy level $(\epsilon', T\delta + \delta')$ with $0 < \epsilon' < 1$ and $\delta' > 0$, it suffices to choose $\epsilon \leq \frac{\epsilon'}{\sqrt{8T \ln(1/\delta')}}$.

Recap: Differentially private optimization

- Assume we want to train a model by solve

$$\min_{\mathbf{w}} f(\mathbf{w}, D)$$

$$\text{s.t.} \quad \mathbf{w} \in \mathcal{W}$$

- Our algorithm returns w^* that may reveal sensitive data.
- How can we solve this optimization problem in a DP fashion?
  - Output perturbation
  - Exponential mechanism
  - Objective perturbation
  - Privatizing the algorithm: DP-SGD

Recap: DP optimization via output perturbation

- Consider the ERM setting

$$\min_{w} \left(L(\mathbf{w}, \mathbf{X}) \triangleq \frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{w}, \mathbf{x}_i) + R(\mathbf{w}) \right)$$

- Can we get the solution (output) and add noise to it to make it DP?

- Output perturbation: $w_{priv} = w^* + z$

- How much noise should we add?

Recap: Output perturbation: sensitivity lemma

$$\mathbf{w}^* = \arg\min_w \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{w}, \mathbf{x}_i) + R(\mathbf{w}) \qquad \mathbf{w}^{*'} = \arg\min_w \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{w}, \mathbf{x}_i') + R(\mathbf{w})$$

- Where $x_2 = x_2', x_3 = x_3', \ldots, x_n = x_n{}'$
- Bounding sensitivity: we need to bound $||w^* - w^{*\prime}||$

Theorem: Let $\ell(\cdot, x)$ be $L-$ Lipschitz and convex; and $R(w)$ be $\mu-$ strongly convex. Then, $\| w^* - w^{*\prime} \|_2 \leq \frac{2L}{\mu n}$

- Do we need to add strongly convex regularizer?
- Do we need Lipschitzness of the loss?
- This bound is tight, why?

Recap: Output perturbation: utility

- How much do we lose after adding noise?

$$L(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{w}, \mathbf{x}_i)$$

Theorem: Let $\ell(\cdot, x)$ be $L-$ Lipschitz and convex; and $R(w)$ be $\mu-$ strongly convex regularizer. Then,

$$E\left[L(w_{priv}) - L(w^*)\right] \leq \frac{3L^2\sqrt{d\log(1.25/\delta)}}{\epsilon\mu\, n}$$

- Tradeoffs in choosing μ
- We can obtain tighter bounds assuming the loss function is $\beta-$ smooth is as well.
- You can make these bounds with high probability (we need to change it a bit though)
- Can we achieve Pure-DP by output perturbation?

How to regularize?

- Goal: $\min_w \left(L(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(w, x_i) \right)$
- Sensitivity = infinity, but we can still add a regularizer and solve the problem

$$\mathbf{w}_R^* = \arg\min_{w \in \mathcal{W}} \frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{w}, \mathbf{x}_i) + \frac{\mu}{2} \|\mathbf{w}\|^2 \qquad \mathbf{w}_{priv} = \mathbf{w}_R^* + \mathbf{z}$$

- What is the minimum excess risk $E\left[L(w_{priv}) \right] - \min_w L(w)$?

Theorem: Let $\ell(\cdot, x)$ be $L-$ Lipschitz and convex and $diam(W) = D$. Then, there exists $\mu > 0$ s.t.

$$E\left[L(w_{priv})\right] - \min_w L(w) = \tilde{O}\left(LD \left(\frac{\sqrt{d}}{\epsilon n} \right)^{1/2} \right)$$

- We ignored $\log\left(\frac{1}{\delta}\right)$ term
- Trivial algorithm provides the bound $L(w_{priv}) - \min_w L(w) \leq LD$
- Excess risk for output perturbation in the convex setting: $\tilde{O}\left(LD \min\left\{ 1, \left(\frac{\sqrt{d}}{\epsilon n} \right)^{\frac{1}{2}} \right\} \right)$

Proof

- Notice that

$$L(w_R^*) + \frac{\mu}{2}||\, w_R^*||^2 \leq L(w^*) + \frac{\mu}{2}\left||w^*|\right|^2 \;\; \rightarrow \;\; L(w_R^*) - L(w^*) \leq \frac{\mu}{2}D^2$$

- Moreover, we have

$$L\left(w_{priv}\right) - L(w_R^*) \leq \frac{3L^2\sqrt{d\,\log(1.25/\delta)}}{\epsilon\mu n}$$

- Adding the two and finding the optimal requlatizer:

$$L\left(w_{priv}\right) - L(w^*) \lesssim LD\left(\frac{\sqrt{d}}{\epsilon n}\right)^{1/2}$$

Limitations of output perturbation

- Only applicable to strongly convex setting or problems with similar behaviors (why?)
- Requires Lipschitz loss functions
- We may want to maintain privacy during training
- Another idea: instead of adding noise to the solution, can we add noise to the objective?

Objective perturbation

- Introduced by Chaudhuri, Monteleoni, and Sarwate (2011).
- Idea: Instead of adding noise to the solution, we perturb the goal of the training itself
- Standard ERM Objective: $J(w, \mathrm{D}) = \frac{1}{n}\sum_{i=1}^{n} \ell(w, x_i) + R(w)$
- Perturbed ERM Objective: $\widetilde{J}(w, \mathrm{D}) = \frac{1}{n}\sum_{i=1}^{n} \ell(w, x_i) + R(w) + \frac{1}{n} b^T w$ where $b \sim N(0, \sigma^2 I)$
- Added noise effectively "tilts" the loss landscape
- Then solve the tilted objective and release $w_{priv} = \arg\min_{w} J(w)$
- How much noise should we add?

Privacy analysis

- Optimality condition: $\nabla \mathrm{J}\left(w_{priv}, \mathcal{D}\right) + \frac{1}{n} b = 0$
- Assuming strong convexity (uniqueness of minimizer), we have

$$P\left(w_{priv} | \mathcal{D}\right) = P\left(n\, \nabla \mathrm{J}\left(w_{priv}, \mathcal{D}\right) + b = 0\right)$$

- It is like making $n \nabla \mathrm{J}\left(w_{priv}, \mathcal{D}\right)$ private
- Computing sensitivity of $n \nabla \mathrm{J}\left(w_{priv}, \mathcal{D}\right)$:

$$||n \nabla \mathrm{J}(w, \mathcal{D}) - n \nabla \mathrm{J}(w, \mathcal{D}')|| \leq \left|\left|\nabla \ell(w, x) - \nabla \ell(w, x')\right|\right| \leq 2L$$

- Noise variance: $\sigma^2 = \frac{2\Delta_2^2}{\epsilon^2} \log(1.25/\delta) = \frac{8L^2 \log(1.25/\delta)}{\epsilon^2}$

Utility analysis

- Let $w_R^* = \arg\min_w \left(J(w) := \frac{1}{n}\sum_{i=1}^n \ell(w, x_i) + R(w) \right)$ $\qquad w_{priv}^* = \arg\min_w \left(J(w) + \frac{1}{n} b \right)$
- By strong convexity, we have

$$J(w_R^*) \geq J(w_{priv}) + \langle \nabla J(w_{priv}), w_R^* - w_{priv} \rangle + \frac{\mu}{2} \left|\left| w_R^* - w_{priv} \right|\right|^2$$

$$= J(w_{priv}) - \frac{1}{n} \langle b, w_R^* - w_{priv} \rangle + \frac{\mu}{2} \left|\left| w_R^* - w_{priv} \right|\right|^2$$

$$\geq J(w_{priv}) - \frac{1}{n} ||b|| \cdot \left|\left| w_R^* - w_{priv} \right|\right| + \frac{\mu}{2} \left|\left| w_R^* - w_{priv} \right|\right|^2 \geq J(w_{priv}) - \frac{1}{2n^2\mu} ||b||^2$$

- Therefore, $\mathrm{E}\left[J(w_{priv})\right] - J(w_R^*) \leq \frac{d\sigma^2}{2n^2\mu} \approx \frac{d\, L^2 \log(1.25/\delta)}{\mu n^2 \epsilon^2}$
- Going from regularized to non-regularized …. → $\mathrm{E}\left[L\left(w_{priv}\right)\right] - \min_w L(w) \lesssim \frac{DL\sqrt{d\log(1/\delta)}}{n\epsilon}$
- Combining with trivial bound: Excess risk for objective perturbation in the convex setting: $\tilde{O}\left(LD \min\left\{1, \frac{\sqrt{d}}{\epsilon n}\right\} \right)$
- Comparison with the output perturbation?

Limitations of objective perturbation

- Requires strong convexity
- Is the strong convexity necessary?
  - Example: median estimation $J(w, \mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n}|w - x_i|$ → With probability 1/2 catastrophic privacy failure
  - $\mathcal{D} = \{0,100\}$ and $\mathcal{D} = \{0,101\}$
- Requires Lipschitz loss functions
- We may want to maintain privacy during training
- Another idea: instead of adding noise to the solution or objective, inject noise to the steps of the algorithm