Trustworthy Machine Learning From an Optimization Lens

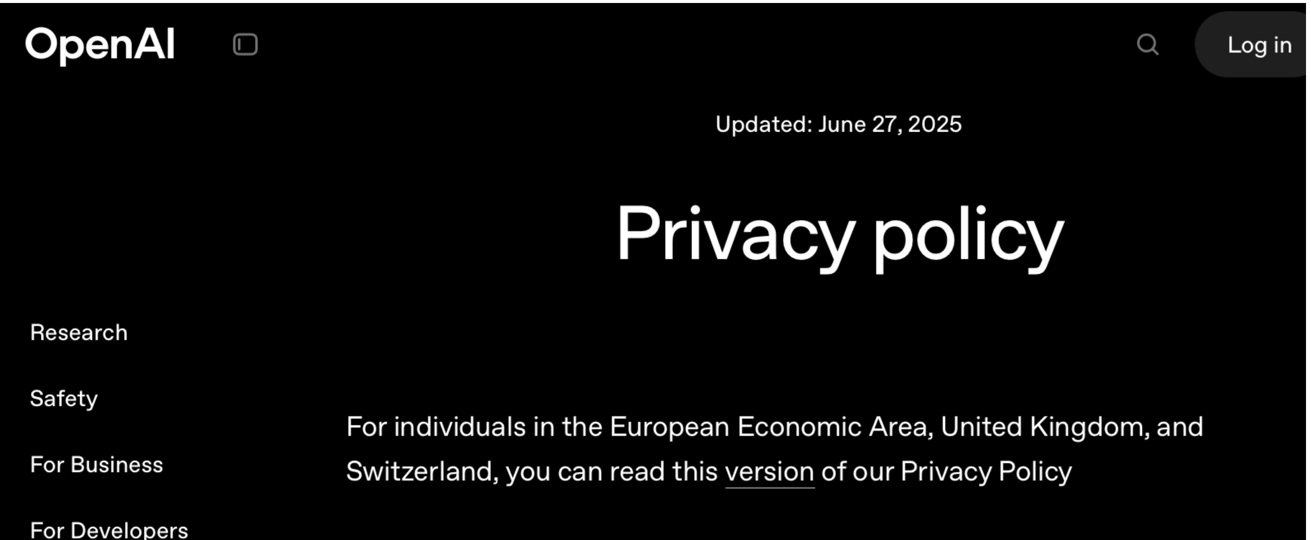
Meisam Razaviyayn

Lecture 9: Properties of DP and Output Perturbation

razaviya@usc.edu

Privacy in the Headlines (2024-2025)

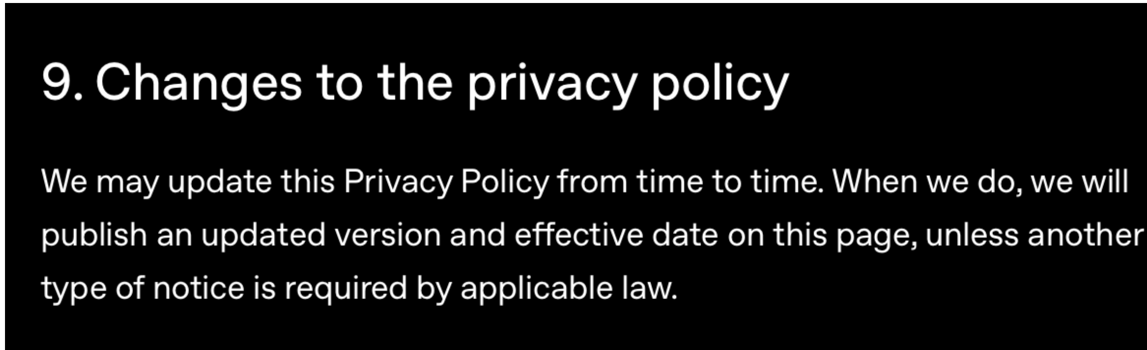
- **Healthcare Breaches** (2024): Change Healthcare attack exposed vulnerabilities in health data systems
- **Data Broker Scrutiny** (2024–2025) : FTC penalized companies (e.g., Avast, X-Mode) for selling sensitive location and browsing data
  - What may location data reveal?
- **LLMs and proprietary data**: Samsung employees leaked code via ChatGPT
- Privacy policy



Updated: June 27, 2025

Privacy policy

For individuals in the European Economic Area, United Kingdom, and Switzerland, you can read this version of our Privacy Policy



9. Changes to the privacy policy

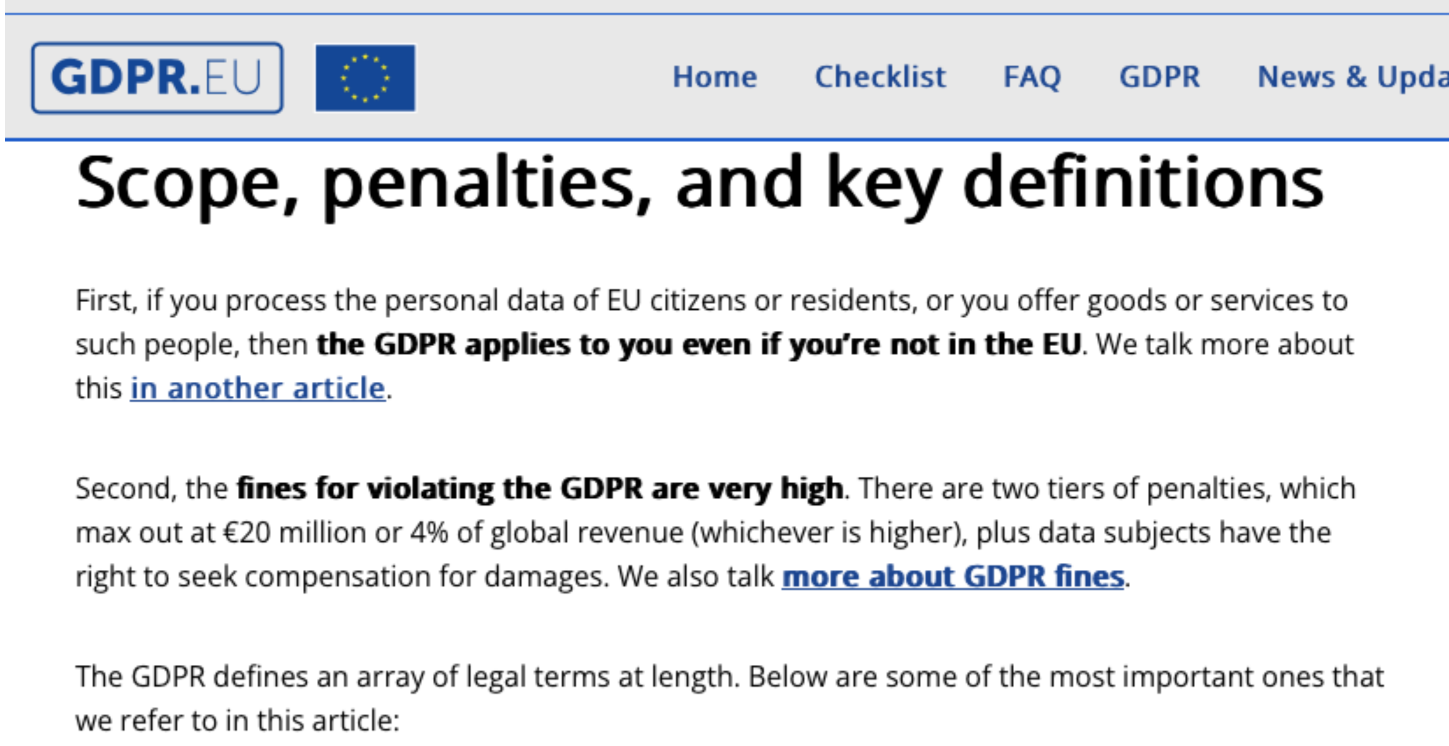
We may update this Privacy Policy from time to time. When we do, we will publish an updated version and effective date on this page, unless another type of notice is required by applicable law.

Regulations on Privacy

- European Union’s General Data Protection Regulation (GDPR)
- California Consumer Privacy Act of 2018
- AI Bill of Rights

- Key Principles:
  - Privacy by Design
  - Data minimization

- Right to be forgotten



Scope, penalties, and key definitions

First, if you process the personal data of EU citizens or residents, or you offer goods or services to such people, then **the GDPR applies to you even if you’re not in the EU**. We talk more about this in another article.

Second, the **fines for violating the GDPR are very high**. There are two tiers of penalties, which max out at €20 million or 4% of global revenue (whichever is higher), plus data subjects have the right to seek compensation for damages. We also talk **more about GDPR fines**.

The GDPR defines an array of legal terms at length. Below are some of the most important ones that we refer to in this article:

THE WHITE HOUSE

DATA PRIVACY

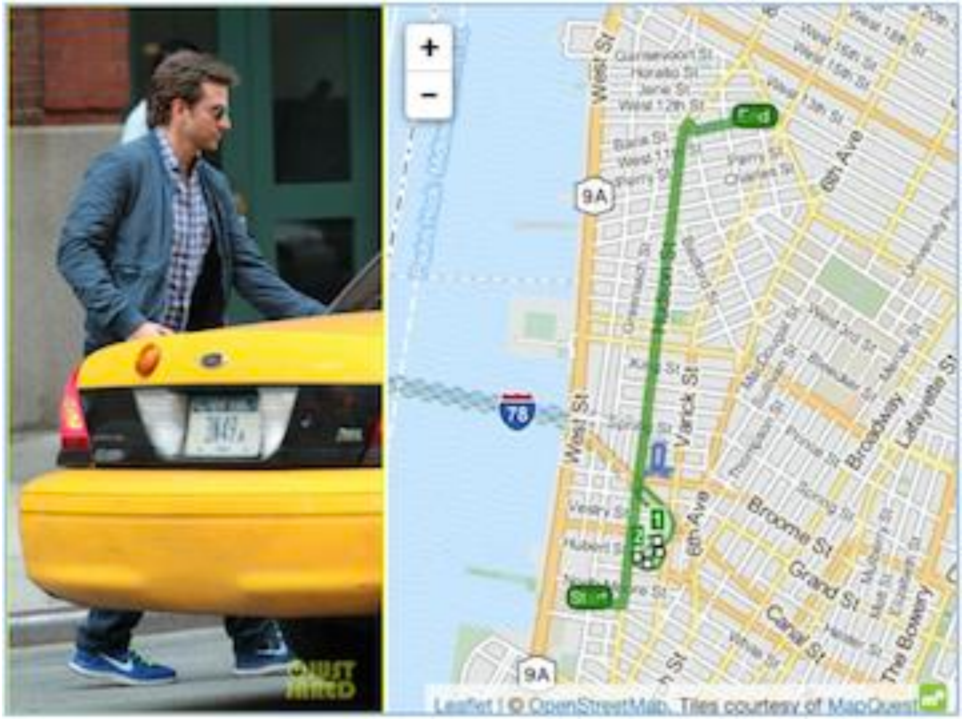
YOU SHOULD BE PROTECTED FROM ABUSIVE DATA PRACTICES VIA BUILT-IN PROTECTIONS AND YOU SHOULD HAVE AGENCY OVER HOW DATA ABOUT YOU IS USED

The Illusion of Anonymization (PII Removal)

- **Naive Approach**: Simply removing obvious Personally Identifiable Information (PII) like names or SSNs

- **Quasi-Identifiers (QIs)**: Attributes that don’t uniquely identify on their own but can re-identify individuals when combined (e.g., Zipcode, Age, Gender).

- **Real-World Example 1**: NYC Taxi Data
  - Originally anonymized by removing driver names and medallion numbers
  - Researchers combined timestamps, pickup/dropoff locations (QIs) to re-identify drivers or trips
  - Reddit discussion: https://www.reddit.com/r/bigquery/comments/28ialf/comment/cicr3n2/
  - Chris Whong Analysis: https://chriswhong.com/open-data/foil_nyc_taxi/



Anonymization ≠ Privacy

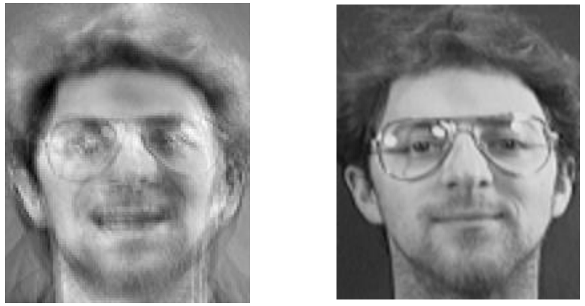
But we are not releasing any data in optimization

- We solve training/optimization problems that depend on data

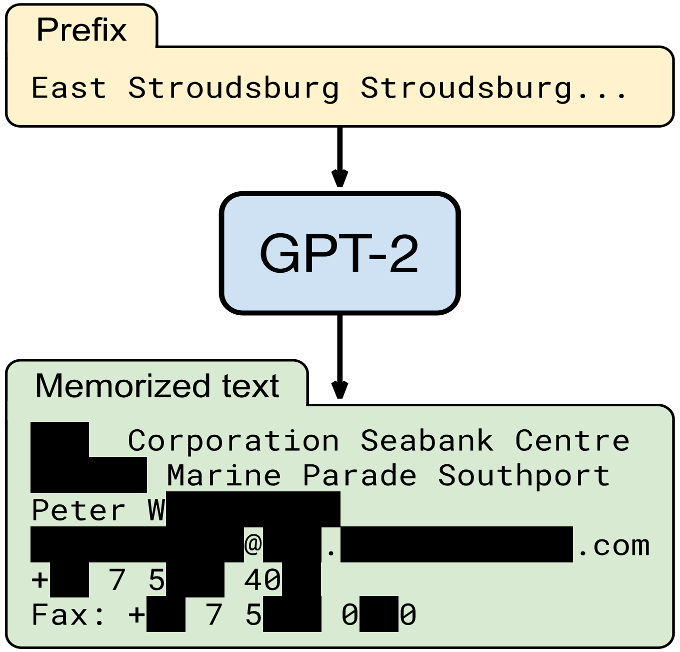
$$\min_{\mathbf{w}} \ \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{w},\mathbf{z}_i)$$

- 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]



[redacted] Corporation Seabank Centre
[redacted] Marine Parade Southport
Peter W[redacted]
[redacted]@[redacted].[redacted].com
+[redacted] 7 5[redacted] 40[redacted]
Fax: +[redacted] 7 5[redacted] 0[redacted]0

[Carlini et al 2021]

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

(Pure) differential privacy

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

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

- 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

(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

$\Pr(M(D) \in \Omega) \leq \mathrm{e}^{\epsilon} \Pr(M(D') \in \Omega) + \delta$ 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

Connection to pure DP

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

- Proof: Let us denote $M(D)$ random variable with Y, and $M(D')$ with Z. Define $G = \{t \mid \ln\left(\frac{P(Y=t)}{P(Z=t)}\right) \leq \epsilon\}$.

$$P(Y \in \Omega) = \int_{\Omega} P(Y=t)dt = \int_{\Omega\cap G} P(Y=t)dt + \int_{\Omega \cap G^c} P(Y=t)dt$$

$$\leq \int_{\Omega\cap G} e^{\epsilon} P(Z=t)dt + P(Y \in \Omega \cap G^c) \leq e^{\epsilon} P(Z \in \Omega) + P(Y \in G^c) \leq e^{\epsilon} P(Z \in \Omega) + \delta$$

- Converse? Consider the following algorithm

$$P(M(D)=0) = \frac{1}{2} \qquad P(M(D)=1) = \frac{1}{2}$$

$$P(M(D')=0) = \frac{1}{3} \qquad P(M(D')=1) = \frac{2}{3}$$

- The algorithm is $\left(\epsilon = 0, \delta = \frac{1}{6}\right)$ $-$DP. Is the privacy loss bounded by 0 w.p 5/6?

Gaussian Mechanism

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

$$\Delta_2 = \sup_{D,D'} \parallel f(D) - f(D') \parallel_2$$

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 ($||x_i||_2 \leq c$), the sensitivity of the mean (assuming neighboring datasets defined by replacement) is $\Delta_2 \ = 2c/n$

Properties of approximate DP

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

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

Group Privacy: Let $M: X \to O$ be $(\epsilon, \delta) - \mathrm{DP}$ 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?

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?

Intuition for Improvement Over Linear Scaling

- Linear scaling $O(T\epsilon)$ is highly pessimistic for large T

- **Why this can be an issue?**
  - To achieve a final ϵ, we need $\epsilon_i = \epsilon/T$.
  - Noise magnitude (e.g., Gaussian mechanism) scales as $O(1/\epsilon_i) = O(T)$
  - The total accumulated variance over T steps scales as T$O(T^2) = O(T^3)$. This rapidly destroys utility.

- **The Intuition for Improvement**:
  - Basic composition assumes PLRV $L_i \leq \epsilon_i$ always.
  - However, L_i's are random variables. We expect some cancellation (hopefully concentration).
  - The expected privacy loss $E[L_i]$ is often much smaller than ϵ_i (typically $O\left(\epsilon_i^2\right)$ for small ϵ_i).
  - We expect the fluctuations of privacy loss to scale with $O\left(\sqrt{T}\right)$ rather than the worst-case $O(T)$.

Composition Theorems

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

Advanced Composition: Let $M = (M_1, .., M_T)$ be a sequence of $(\epsilon, \delta) - \mathrm{DP}$ algorithms (may be chosen adaptively). Then, for any $\delta' > 0$, $M(\cdot)$ is $(\epsilon', T\delta + \delta') - \mathrm{DP}$ 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 $(\epsilon, \delta) - \mathrm{DP}$ 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 $\epsilon \leq \frac{\epsilon'}{\sqrt{8T\ln(1/\delta')}}$.

Advanced Composition: Proof Idea

- Assume $\delta = 0$
- Total privacy loss $L_{tot} = \sum_{i=1}^{T} L_i$, we need to get the bound $P(L_{tot} > \epsilon') \le \delta'$
- We want to do concentration; however, L_i's are dependent (since M_i's are selected adaptively)
- Idea: do martingale decomposition
  - Define $F_i = (s_1, s_2, \ldots, s_i)$ and $\mu_i = E[L_i \mid F_{i-1}]$
  - Therefore, $E[L_{tot}] = \underbrace{\sum_{i=1}^{T} \mu_i}_{\text{Drift}} + \underbrace{\sum_{i=1}^{T} (L_i - \mu_i)}_{\text{Random Fluctuations}}$
- We will show E[Drift] $\le \frac{T\epsilon(\exp(\epsilon) - 1)}{\exp(\epsilon) + 1} \le T\epsilon(\exp(\epsilon) - 1)$
- We will show E[Random Fluctuations] $\le$ …
- Apply Chernoff bound

Bounding the drift

- **Lemma**: for an $(\epsilon, 0) - DP$ mechanism with PLRV L, we have

$$E[L] \leq \frac{\epsilon(\exp(\epsilon)\ - 1)}{\exp(\epsilon) +\ 1} \leq \epsilon(\exp(\epsilon)\ - 1) \approx O(\epsilon^2)$$

- **Proof Idea**: find max $E[L]$ s.t. constraints imposed by $(\epsilon, 0)$ −DP
- Recall $L(o) = \log\left(\frac{P(o)}{Q(o)}\right)$ with probability $P(o)$
- Define $X(o) = P(o)/Q(o)$ → $E[L] = E_Q[X \log X]$
- Constraints: $X(o) \in [e^{-\epsilon}, e^{\epsilon}]$ and $E_Q[X] = 1$
- Problem: max $E_Q[X \log X]$ s.t. $X \in [e^{-\epsilon}, e^{\epsilon}]$ and $E_Q[X] = 1$
- We can show that it is better to put the mass on two points only → $X^* \in \{e^{-\epsilon}, e^{\epsilon}\}$
- Solve the two-dimensional problem: $P(X = e^{\epsilon}) = 1\ -\ P(X = e^{-\epsilon}) = \frac{1}{e^{\epsilon}+1}$
- Plugging in the solution: $E[L] \leq \frac{\epsilon(e^{\epsilon}-1)}{e^{\epsilon}+1}$

Bounding the fluctuations

- Define $S_T = \sum_{i=1}^{T}(X_i \coloneqq L_i - \mu_i)$
- We know (conditionally) $E[X_i \, | F_{i-1}] = 0$ and $X_i \in [\alpha, \alpha + 2\epsilon]$
- Apply Hoeffding's (conditionally) → $E\left[e^{\lambda X_i} \mid F_{i-1}\right] \le e^{\frac{\lambda^2 \epsilon^2}{2}}$
- Recursive calculation: $E\left[e^{\lambda \sum_i X_i}\right] = E\left[e^{\lambda S_{T-1}} \; E\left[e^{\lambda X_T} \mid F_{T-1}\right]\right] \le E\left[e^{\lambda S_{T-1}}\right] e^{\frac{\lambda^2 \epsilon^2}{2}} \le e^{T \frac{\lambda^2 \epsilon^2}{2}}$
- Apply Chernoff → $P(S_T > \gamma) \le \inf_\lambda e^{-\lambda\gamma + T\frac{\lambda^2 \epsilon^2}{2}} = e^{-\frac{\gamma^2}{2T\,\epsilon^2}} \coloneqq \delta'$
- Rearranging → fluctuations = $\gamma = \epsilon\sqrt{2T \log(1/\delta')}$
- Therefore, $\epsilon' =$ drift + fluctuations $\le \epsilon\sqrt{2T \ln\left(\frac{1}{\delta'}\right)} + \frac{T\epsilon(e^{\epsilon}-1)}{e^{\epsilon}+1}$

Hoeffding's lemma: If $E[Y] = 0$ and Y bounded in an interval of size w, then $E\left[e^{\lambda Y}\right] \le e^{\frac{\lambda^2 w^2}{8}}$

Chernoff bound:
$P(Y \ge \gamma) \le \inf_\lambda e^{-\lambda\gamma} \, E\left[e^{\lambda Y}\right]$

Q.E.D.

Remarks on composition

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.

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 $\epsilon' = \epsilon\sqrt{2T\ln\left(\frac{1}{\delta'}\right)} + \frac{T\epsilon(e^{\epsilon}-1)}{e^{\epsilon}+1}$.

- ➢ See [Kairouz et al., 2015] for slightly tighter results that also hold when we have $(\varepsilon_i, \delta_i)$-DP
- ➢ BUT these composition results are typically not tight in practice (Why?)
- ➢ Some variants of (ε, δ)-DP, such as Rényi DP [Mironov, 2017] and zero-concentrated DP (zCDP) [Bun and Steinke, 2016], can enable tighter bounds for specific mechanisms (such as Gaussian mechanism)

Applying DP to practical ML models

Applying DP to practical ML models

- Privatizing the model
  - Privatizing training data
  - **DP training/optimization**
  - Synthetic data
  - …
- Privatizing the output

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

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?

Output perturbation: sensitivity lemma

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

- Where $x_2 = x_2', x_3 = x_3', \dots, x_n = x_n{}'$
- 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?

Proof of the sensitivity lemma

- Let $G(w) = \frac{1}{n}\sum_i \ell(w, x_i) + R(w)$ and $g(w) = \frac{1}{n}(\ell(w, x_1') - \ell(w, x_1))$

- Then, $w^* = \arg\min_w G(w)$ and $w^{*\prime} = \arg\min_w G(w) + g(w)$

- By strong convexity we have,

  - $G(w^*) \geq G(w^{*\prime}) + \langle \nabla G(w^{*\prime}), w^* - w^{*\prime} \rangle + \frac{\mu}{2} \parallel w - w^* \parallel^2$
  - $G(w^{*\prime}) \geq G(w^*) + \frac{\mu}{2} \parallel w - w^* \parallel^2$

- Therefore, $\langle \nabla G(w^{*\prime}), w^{*\prime} - w^* \rangle \geq \mu \parallel w - w^* \parallel^2$

- Using Cauchy-Schwarz inequality, and noticing that $\nabla G(w^{*\prime}) = \nabla G(w^{*\prime}) + \nabla g(w^{*\prime}) - \nabla g(w^{*\prime}) = -\nabla g(w^{*\prime})$, will complete the proof.

Output perturbation: how much noise to add?

- Recall:

Theorem: Let $f: X \rightarrow R^k$ have the $\ell_2 -$ sensitivity Δ_2 and $M(D) = f(D) + (Z_1, .., Z_k)$ Where $Z_1, \dots, Z_k$ are iid Gaussian with variance $\sigma = \frac{\Delta_2 \sqrt{2 \ln \frac{1.25}{\delta}}}{\epsilon}$. Then $M(\cdot)$ is $(\epsilon, \delta) -$ DP.

- For the minimizer, we have $\Delta_2 = \frac{2L}{\mu n}$.

- For achieving $(\epsilon, \delta) -$ DP, we need to add Gaussian noise with $\sigma = \frac{2L\sqrt{2 \ln \frac{1.25}{\delta}}}{\epsilon\, \mu\, n}$

$$w_{priv} = w^* + z \quad with \;\; z \sim N(0, \sigma^2 I)$$

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,

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

- 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?

Proof of the Utility Bound

$$
\begin{aligned}
E\left[L\left(w_{priv}\right) - L(w^*)\right] &= E\left[\left\langle \nabla L(\widehat{w}), w_{priv} - w^* \right\rangle\right] \\
&\leq E\left[\left|\left|\nabla L(\widehat{w})\right|\right| \cdot ||w_{priv} - w^*||\right] \\
&\leq L\, E\left[||w_{priv} - w^*||\right] \\
&\leq L\, E\left[||Z||_2\right] \\
&\leq L \sqrt{E\left[||Z||_2^2\right]} \\
&= L\sqrt{d\sigma^2}
\end{aligned}
$$

- ➢ Plugging the value $\sigma = \frac{2L\sqrt{2\ln\frac{1.25}{\delta}}}{\epsilon\, \mu\, n}$ will complete the proof.

What if we had no regularizer

- 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

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

- 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: $\tilde{O}\left(LD \min\left\{1, \left(\frac{\sqrt{d}}{\epsilon n}\right)^{\frac{1}{2}}\right\}\right)$

Limitations of output perturbation

- Only applicable to (strongly) convex setting or problems with similar behaviors (why?)

- We may want to maintain privacy during training

- Why not add noise to the steps of an algorithm?