Lecture 10 Convex Differentially Private Empirical Risk Minimization

Yu-Xiang Wang

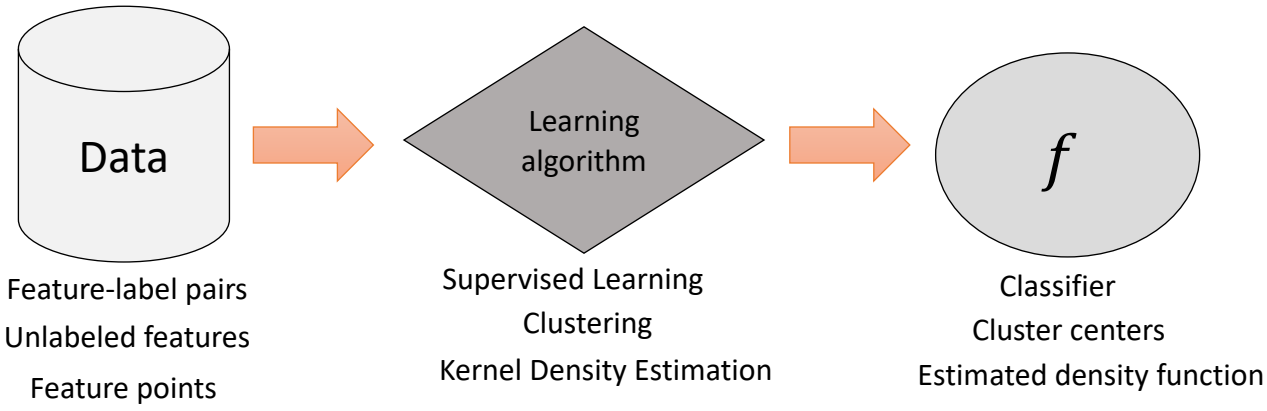
UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

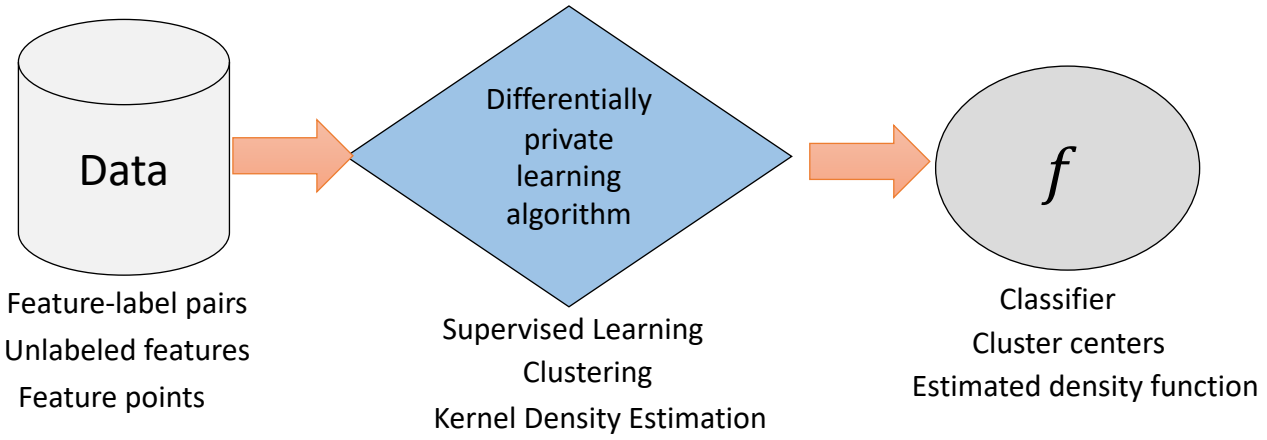
Recap: Last lecture

- Introduction to Differentially Private Machine Learning
  - Problem setup and notations
  - Examples
  - A learning theoretic study of the problem

- Posterior sampling mechanism
  - When the loss functions are bounded
  - A new analysis of the when they are not (not covered)

Recap: Differentially Private Machine Learning





Recap: Statistical Learning Jargons / Notations

- Data space / data points

$$(x_1, y_1), ..., (x_n, y_n) \in \mathcal{X} \times \mathcal{Y} = \mathcal{Z}$$

- Hypothesis class $\mathcal{H}$
- Loss function, risk and empirical risk

$$\hat{R}_n(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(h, (x_i, y_i)) \qquad R(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(h, (x, y))]$$

- Goal of learning

- Realizable vs Agnostic

Recap: Learning with differential privacy in theory

- Finite hypothesis class

	ERM	Exponential mechanism
Excess risk (Agnostic)	$\sqrt{\frac{\log(\lvert\mathcal{H}\rvert/\beta)}{n}}$	$\sqrt{\frac{\log(\lvert\mathcal{H}\rvert/\beta)}{n}} + \frac{\log(\lvert\mathcal{H}\rvert/\beta)}{n\epsilon}$
Excess risk (Realizable)	$\frac{\log(\lvert\mathcal{H}\rvert/\beta)}{n}$	$\frac{\log(\lvert\mathcal{H}\rvert/\beta)}{n} + \frac{\log(\lvert\mathcal{H}\rvert/\beta)}{n\epsilon}$

- Continuous hypothesis class (bounded VC-dim)
  - No DP algorithm gives consistent learning
  - A Packing lower bound
- However, there are weak assumptions we can add
  - Lipschitz loss functions
  - data-distributions with bounded density

Recap: Packing lower bound in the problem of learning threshold function

- Assume K-data distributions with disjoint support
- Assume a DP algorithm A works on (K-1) of them
  - Expected error < 0.05 => need to stay within support w.p. 0.9
- Implies that it does not work on the remaining one using the definition of DP
  - Probability of outside support > 0.9 => Expected error > 0.45

Recap: Alternative analysis of the exponential mechanism.

- Bayesian learning:
  - Posterior sampling is private for free, if the log-likelihood function is bounded (or clipped).
  - Utility: (1+ 4B/ε) multiplicative factor in efficiency.

- More advanced analysis by Minami et al:
  - For strongly log-concave prior

This lecture

- Differentially Private Empirical Risk Minimization
  - Output perturbation
  - Objective perturbation

Optional readings

- “Differentially Private Empirical Risk Minimization” by Chaudhuri, Monteleone, Sarwate.
  - https://www.jmlr.org/papers/volume12/chaudhuri11a/chaudhuri11a.pdf

- “Private Convex Empirical Risk Minimization and High-dimensional regression” by Kifer, Smith and Thakurta
  - http://proceedings.mlr.press/v23/kifer12/kifer12.pdf

The problem of convex empirical risk minimization

- Data
- Hypothesis class
- Loss function

A specific class of problems of interest: generalized Linear models

- Loss function is of a particular form

- Examples:
  - Linear regression
  - Logistic regression
  - Support vector machine

(Detour) Convex Optimization 101

- Convex functions
- Strongly convex functions
- Optimality condition for differentiable convex functions

(Detour) Convex Optimization 101

- Lipschitz constant of a function
- Smoothness constant of a function

Can we apply Laplace / Gaussian mechanism?

- What is the global sensitivity for the minimizer?

Let’s regularize and make it strongly convex

- Regularized objective function

- Now what is the global sensitivity again?

Output perturbation mechanism

1. Solve for the optimal solution

2. Add noise to the optimal solution

Utility analysis of the output perturbation mechanism

- Let’s appeal to the smoothness of the loss function

Objective perturbation

Algorithm 1 Release $\hat{\theta}^P$ via `Obj-Pert` (Kifer et al., 2012)

Input: Dataset D, noise parameter σ, regularization parameter λ, loss function $L(\theta; D) = \sum_i \ell(\theta; z_i)$, convex and twice-differentiable regularizer $r(\theta)$, convex set Θ.
Output: $\hat{\theta}^P$, the minimizer of the perturbed objective.
Draw noise vector $b \sim \mathcal{N}(0, \sigma^2 I)$.
Compute $\hat{\theta}^P$ according to (1).

$$\hat{\theta}^P = \underset{\theta \in \Theta}{\operatorname{argmin}} L(\theta; D) + r(\theta) + \frac{\lambda}{2}||\theta||_2^2 + b^T\theta, \tag{1}$$

- Remark:
  - Equivalent to adding a random synthetic data point with linear loss
  - One could also get pure DP if we choose b to be:

KKT condition of Objective Perturbation implies a relationship between b and the optimal solution.

$$\hat{\theta}^P = \underset{\theta \in \Theta}{\operatorname{argmin}} L(\theta; D) + r(\theta) + \frac{\lambda}{2}||\theta||_2^2 + b^T\theta,$$

Change of variable trick: distribution via the Jacobian

Theorem: Let X, Y be random variables in $\mathbb{R}^d$ and $Y = g(X)$ satisfying that function g is bijective (one-to-one map) and differentiable. Then the probability density function of Y is:

$$f_Y(y) = f_X(g^{-1}(y)) \left| \det \left[\frac{\partial g^{-1}(y)}{\partial y} \right] \right|$$

- Apply this theorem to our problem:

The privacy loss random variable of the objective perturbation

$$
\begin{aligned}
\log \frac{\Pr\big(\mathcal{A}(D) = \hat{\theta}^P\big)}{\Pr\big(\mathcal{A}(D_{\pm z}) = \hat{\theta}^P\big)} &= \log \frac{\left|\det\Big(\nabla b(\hat{\theta}^P; D)\Big)\right|}{\left|\det\Big(\nabla b(\hat{\theta}^P; D_{\pm z})\Big)\right|} \frac{\nu(b(\hat{\theta}^P; D); \sigma)}{\nu(b(\hat{\theta}^P; D_{\pm z}); \sigma)} \\
&= \log \frac{\left|\det\Big(\nabla b(\hat{\theta}^P; D)\Big)\right|}{\left|\det\Big(\nabla b(\hat{\theta}^P; D_{\pm z})\Big)\right|} + \log \frac{e^{-\frac{1}{2\sigma^2}||b(\hat{\theta}^P; D)||_2^2}}{e^{-\frac{1}{2\sigma^2}||b(\hat{\theta}^P; D_{\pm z})||_2^2}} \\
&= \underbrace{\log \frac{\left|\det\Big(\nabla b(\hat{\theta}^P; D)\Big)\right|}{\left|\det\Big(\nabla b(\hat{\theta}^P; D_{\pm z})\Big)\right|}}_{(*)} + \underbrace{\frac{1}{2\sigma^2}\Big(||b(\hat{\theta}^P; D_{\pm z})||_2^2 - ||b(\hat{\theta}^P; D)||_2^2\Big)}_{(**)}.
\end{aligned}
$$

Bounding (*): the difference of the log determinants

Lemma 25 (Determinant of Rank-1 perturbation)**.** *For invertible matrix A and vector c, d of compatible dimension*

$$\det(A + cd^T) = \det(A)(1 + d^T A^{-1} c).$$

Recall:

$$b(\hat{\theta}^P; D) = -\left(\nabla \hat{\mathcal{L}}(\hat{\theta}^P; D) + \nabla r(\hat{\theta}^P) + \lambda \hat{\theta}^P\right).$$

$$\nabla b(\hat{\theta}^P; D) = -\left(\nabla^2 \hat{\mathcal{L}}(\hat{\theta}^P; D) + \nabla^2 r(\hat{\theta}^P) + \lambda I_d\right).$$

- Iterative application of Lemma 25

$$\begin{aligned}\left|\det\left(\nabla b(\hat{\theta}^P; D_{\pm z})\right)\right| &= \left|\det\left(\nabla b(\hat{\theta}^P; D) \mp \nabla^2 \ell(\hat{\theta}^P; z)\right)\right| \\ &= \left|\det\left(\nabla b(\hat{\theta}^P; D) \mp \sum_{k=1}^{d} \lambda_k u_k u_k^T\right)\right|\end{aligned}$$

- For generalized linear models

Recall:

$$b(\hat{\theta}^P; D) = -\left(\nabla\hat{\mathcal{L}}(\hat{\theta}^P; D) + \nabla r(\hat{\theta}^P) + \lambda\hat{\theta}^P\right).$$

$$\nabla b(\hat{\theta}^P; D) = -\left(\nabla^2\hat{\mathcal{L}}(\hat{\theta}^P; D) + \nabla^2 r(\hat{\theta}^P) + \lambda I_d\right).$$

Bounding (**)

$$(**) = \frac{1}{2\sigma^2}\left(||b(\hat{\theta}^P; D_{\pm z})||_2^2 - ||b(\hat{\theta}^P; D)||_2^2\right)$$
$$= \frac{1}{2\sigma^2}\Big[\mp\nabla\ell(\hat{\theta}^P; z)\Big]\Big[2b(\hat{\theta}^P; D) \mp \nabla\ell(\hat{\theta}^P; z)\Big]$$

- For generalized linear models:

Putting everything together

- For general loss functions

Next lecture

- Utility analysis of ObjPert

- Noisy Gradient Descent