Lecture 13

Differentially Private ERM

A Quick Primer on Non-Private Machine Learning

Formulation

feature vector

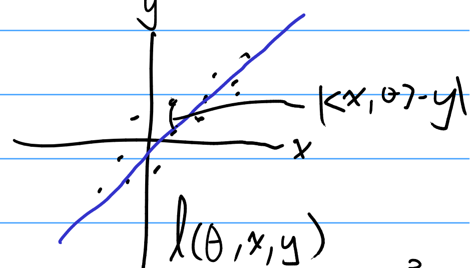
Dataset D of (x,y) pairs

label

loss function: $\ell(\theta, x, y)$

parameter vector

Linear regression



$\ell(\theta, x, y) = (\langle x,\theta\rangle - y)^2$

$$\mathcal{L}(\theta, D) = \sum_{i=1}^{n} \ell(\theta, x_i, y_i)$$

Empirical Risk Minimization (ERM)

Goal: minimize $\mathcal{L}(\theta, D)$

Given: Dataset D of (x,y), "Parameter space" C

Ideal $\theta^* = \underset{\theta \in C}{\arg\min}\, \mathcal{L}(\theta, D)$

Algo will output $\hat{\theta}$.

$f_{\hat{\theta}}(x) = \langle \hat{\theta}, x\rangle$

Expected excess empirical risk

$$E\left[\mathcal{L}(\hat{\theta}, D) - \mathcal{L}(\theta^*, D)\right]$$

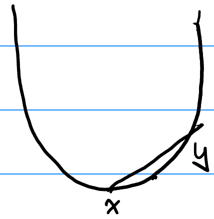
over algo's randomness

Restrictions:

- Diameter of C is bounded
- $\ell(\cdot, x_i, y_i)$ convex, L-Lipschitz $\forall (x_i, y_i) \in \mathcal{X}$

Terminology

- Gradient: $\ell(\theta): \mathbb{R}^d \to \mathbb{R}$
  Gradient at $\tilde{\theta}$ is $\nabla \ell(\tilde{\theta}) \in \mathbb{R}^d$ where ith coord is
  $\frac{\partial}{\partial \theta_i}(\ell(\tilde{\theta}))$
- Diameter of C, $\|C\|_2$
  Max dist between 2 pts in C



- $\ell: C \to \mathbb{R}$ is convex if $\forall x, y \in C$, $\forall t \in [0,1]$
  $f(tx+(1-t)y) \leq tf(x)+(1-t)f(y)$
- $\ell: C \to \mathbb{R}$ is L-Lipschitz$_2$ if $\forall x, y \in C$
  $|\ell(x)-\ell(y)| \leq L\|x-y\|_2$

$\Rightarrow \|\nabla \ell\|_2 \leq L$

Loss Functions

$x, \theta \in \mathbb{R}^d$

1. Linear regression: $y \in \mathbb{R}$, $\ell(\theta, x, y) = (\langle x, \theta \rangle - y)^2$
2. Logistic Regression: $y \in \{\pm 1\}$, $\ell(\theta, x, y) = \log(1 + e^{-y\langle x, \theta \rangle})$
3. (Geometric) Median: $\ell(\theta, x, y) = \|\theta - x\|_2$. 1-Lipschitz.
   Mean: $\ell(\theta, x, y) = \|\theta - x\|_2^2$

   $\downarrow$ 1D

   $|\theta - x|$

4. Support Vector Machine (SVM): $y \in \{\pm 1\}$, $\ell(\theta, x, y) = \max(0, 1 - y\langle x, \theta \rangle)$

Optimization

$\theta^* = \arg\min \mathcal{L}(\theta, D)$

Gradient descent

Assume non-priv. optimizer exists

Generalization

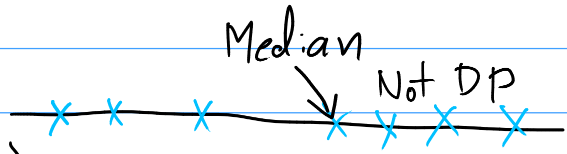
Uniform convergence + ERM $\Rightarrow$ generalization

Privacy Considerations

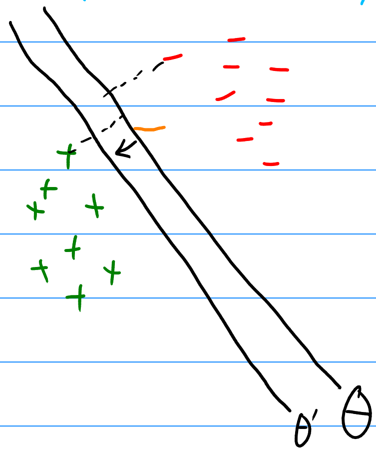
$\hat{\theta}$ which minimizes $\mathcal{L}(\cdot, D)$

$\hat{\theta}$ is DP wrt D

Median:



SVM:



- Output perturbation
- Objective "
- Gradient "

- Input perturbation? $\leftrightarrow$ LDP

Output Perturbation

Chaudhuri, Monteleoni, Sarwate '11 (CMS'11)

Regularization

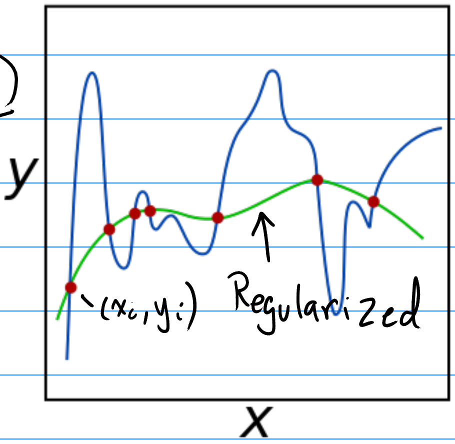
Polynomial Regression. $(x_i, y_i) \in \mathbb{R}^2$. $\theta \in \mathbb{R}^{k+1}$

$$\ell(\theta, x, y) = \left(\sum_{j=0}^{k} \theta_j x^j - y\right)^2 \qquad \hat{y} = \sum \theta_j x^j$$

$k = n$.

$$\mathcal{L}(\theta, D) = \sum \ell(\theta, x_i, y_i) + \lambda \underline{N(\theta)}$$

$$N(\theta) = \sum \theta_j^2$$



- Poor gen.
- Sensitive.

Lemma 1 (Corollary 8 in [CMS11]). *If N is differentiable and 1-strongly convex, and ℓ is convex and 1-Lipschitz, then ~~for all θ~~ global, the ℓ_2-sensitivity of $\left(\arg\min_{\theta \in C} \mathcal{L}(\theta, D)\right)$ is at most $\frac{2n}{\lambda}$.*

$$\tilde{\theta} = \arg\min_{\theta \in C} \mathcal{L}(\theta, D)$$

$$\hat{\theta} = \tilde{\theta} + b \leftarrow \text{noise}$$

$$b \sim N\left(0, O\left(\frac{n^2 \log 1/\delta}{\lambda^2 \varepsilon^2}\right) \cdot I_{d \times d}\right)$$

(ε, δ)-DP

Objective Perturbation

Before: $f(\theta, D) = \sum \ell(\theta, x_i, y_i) + \frac{\lambda}{2} \|\theta\|_2^2$

Random perturb f

$f^{Priv}(\theta, D) = \sum \ell(\theta, x_i, y_i) + \frac{\lambda}{2} \|\theta\|_2^2 + \langle b, \theta \rangle$

b is Gamma or Gaussian

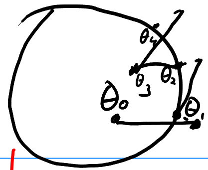
1. Perturb f to get f^{Priv}
2. Solve f^{Priv} non-privately

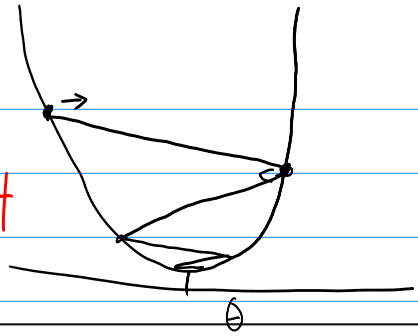
Drawbacks:

- Exact optimization req'd
  - ↳ Numerical prec   Mironov '12
  - ↳ Iterative methods (!! ☹)

$\Rightarrow$ Approx optimum.

- Restrictive assns.

$C =$ 



Gradient Perturbation

Non-Private Gradient Descent

Algorithm 1: Projected Gradient Descent

Set $\theta_0 \in \mathcal{C}$ arbitrarily
for $t = 1$ *to* T **do**
  Compute $\theta_t = \Pi_{\mathcal{C}}\left(\theta_{t-1} - \eta(t)\nabla\mathcal{L}(\theta_{t-1}, D)\right)$
end
return $\hat{\theta} = \theta_T$

learning rate

projection onto C

Theorem 2 (Theorem 2 from [SZ13]). *Let F be a convex function and let $\theta^* = \arg\min_{\theta\in\mathcal{C}} F(\theta)$. Let θ_0 be an arbitrary point in $\mathcal{C}$, and $\theta_{t+1} = \Pi_{\mathcal{C}}\left(\theta_t - \eta(t)G_t(\theta_t)\right)$, where $\mathbf{E}[G_t(\theta_t)] = \nabla F(\theta_t)$ and $\mathbf{E}[\|G_t(\theta_t)\|_2^2] \le G^2$, and the learning rate function $\eta(t) = \frac{\|\mathcal{C}\|_2}{G\sqrt{t}}$. Then for any $T > 0$, we have the following:*

$$\mathbf{E}[F(\theta_T) - F(\theta^*)] = O\left(\frac{\|\mathcal{C}\|_2 G \log T}{\sqrt{T}}\right).$$

$\nabla \mathcal{L}(\theta_{t-1}, D) = G_t(\theta_t)$

$G = \underline{n \cdot L}$ $\qquad \tilde{O}\left(\frac{\|C\|_2 nL}{\sqrt{T}}\right)$ $\qquad T \geq \Omega\left(\frac{n^2L^2\|C\|_2^2}{\alpha^2}\right)$

$\Rightarrow \leq \alpha.$

$T \cdot n$ Grad comps

$\approx n^3$.

Algorithm 2: Stochastic Projected Gradient Descent

Set $\theta_0 \in \mathcal{C}$ arbitrarily
for $t = 1$ *to* T **do**
  Select $i \in [n]$ uniformly at random
  Compute $\theta_t = \Pi_{\mathcal{C}}\left(\theta_{t-1} - \eta(t)\left(n \cdot \nabla\ell(\theta_{t-1}, x_i, y_i)\right)\right)$
end
return $\hat{\theta} = \theta_T$

$\nabla \mathcal{L}(\theta_{t-1}, D)$

$G_t(\theta_t) = n\nabla\ell(\theta_{t-1}, x_i, y_i)$

$E[G_t(\theta_t)] = \nabla\mathcal{L}(\theta_t, D)$

$E[\|G_t(\theta_t)\|_2^2] \leq n^2L^2 \Rightarrow G = nL$

$Tn \approx n^2$ comps. vs n^3

Private Stochastic Gradient Descent

Algorithm 3: Private Stochastic Projected Gradient Descent

Define $\sigma^2 \leftarrow \frac{32L^2n^2\log(n/\delta)\log(1/\delta)}{\varepsilon^2}$
Set $\theta_0 \in \mathcal{C}$ arbitrarily
for $t = 1$ *to* n^2 **do**
  Select $i \in [n]$ uniformly at random
  Compute $\theta_t = \Pi_{\mathcal{C}}\left(\theta_{t-1} - \eta(t)\left(n \cdot \nabla\ell(\theta_{t-1}, x_i, y_i) + b_{t-1}\right)\right)$, where $b_{t-1} \sim N(0, \sigma^2 \cdot I_{d\times d})$
end
return $\hat{\theta} = \theta_T$

Utility

$$G_t(\theta_t) = n \cdot \nabla \ell(\theta_t, x_i, y_i) + b_t$$

$$E[G_t(\theta_t)] = \nabla \mathcal{L}(\theta_t, D) \cancel{+0}$$

$$E[\|G_t\|_2^2] = n^2 E[\|\nabla\ell\|_2^2] + 2nE[\langle \nabla\ell, b_t\rangle] + E[\|b_t\|_2^2]$$
$$\leq n^2L^2 + 0 + d\sigma^2$$

$$E[\mathcal{L}(\theta_T, D) - \mathcal{L}(\theta^*, D)] \leq \tilde{O}\left(\frac{\|C\|_2\sqrt{n^2L^2 + d\sigma^2}}{\sqrt{n^2}}\right)$$
$$= \tilde{O}\left(\frac{\|C\|_2 L\sqrt{d\log(1/\delta)}}{\varepsilon}\right)$$

Privacy

1. Gaussian Mech
2. Amp. by subsamp.
3. Adv. comp.

Algorithm 3: Private Stochastic Projected Gradien

Define $\sigma^2 \leftarrow \frac{32L^2 n^2 \log(n/\delta)\log(1/\delta)}{\varepsilon^2}$

Set $\theta_0 \in \mathcal{C}$ arbitrarily

for $t = 1$ *to* n^2 **do**

Select $i \in [n]$ uniformly at random

Compute $\theta_t = \Pi_{\mathcal{C}}\left(\theta_{t-1} - \eta(t)\left(n \cdot \nabla \ell(\theta_{t-1}, x_i, y\right.\right.$

end

return $\hat{\theta} = \theta_T$

1. $n \cdot \nabla \ell(\theta_{t-1}, x_i, y_i)$ $\quad \|\nabla \ell\|_2 \le L$

ℓ_2-Sens $\le 2nL$

Assume i is fixed $\quad \forall t = 1$ to n^2

G.M. $\Rightarrow$ priv loss RV for $G_t(\theta_t) \le \frac{\varepsilon}{2\sqrt{\log(1/\delta)}}$ w.p $1 - \frac{\delta}{2}$. ✓

2. Lemma: If A is $\varepsilon' < 1$ DP, if executed on a size γn subsample ~~on~~ from a dataset of size n, then result is $2\gamma\varepsilon'$-DP.

$\gamma = \frac{1}{n}$, $\quad \forall t = 1$ to n^2

$\varepsilon' = \frac{\varepsilon}{2\sqrt{\log(1/\delta)}} \rightarrow \frac{\varepsilon}{n\sqrt{\log(1/\delta)}}$-DP w.p $1 - \frac{\delta}{2}$.

3. n^2 iterations

Basic comp $\left(\varepsilon n, \frac{\delta}{2}\right)$-DP

Advanced (ε, δ)-DP