CSC2412: Private Gradient Descent & Empirical Risk Minimization

Sasho Nikolov

Empirical Risk Minimization

Learning: Reminder

- Known data universe $\mathcal{X}$ and an unknown probability distribution D on $\mathcal{X}$
- Known concept class C and an unknown concept $c \in C$
- We get a dataset $X = \{(x_1, c(x_1)), \ldots, (x_n, c(x_n))\}$, where each x_i is an independent sample from D.

Goal: Learn c from X.

Loss

Binary loss (data pt, label) $\ell(c', (x, y)) = \begin{cases} 1 & c'(x) \neq y \\ 0 & c'(x) = y \end{cases}$

$\mathbb{P}_{x\sim D}(c'(x) \neq c(x)) = \mathbb{E}_{x\sim D}[\underbrace{1\{c'(x) \neq c(x)\}}_{= \ell(c', (x, c(x)))}]$

$= L_{D,c}(c') = \mathbb{E}_{x\sim D}[\ell(c', (x, c(x)))] = \mathbb{P}_{x\sim D}(c'(x) \neq c(x))$

We want an algorithm $\mathcal{M}$ that outputs some $c' \in C$ and satisfies

$$\mathbb{P}(L_{D,c}(\mathcal{M}(X)) \leq \alpha) \geq 1 - \beta.$$

Agnostic learning

Maybe no concept gives 100% correct labels. → agnostic setting (as opposed to realizable)

Generally, we have a distribution D on $\mathcal{X} \times \{-1, +1\}$. distribution on labeled examples

$$L_D(c) = \mathbb{E}_{(x,y)\sim D}[\ell(c, (x, y))] = \mathbb{P}_{(x,y)\sim D}(c(x) \neq y)$$

D is unknown but we are given iid samples $X = \{(x_1, y_1), \ldots, (x_n, y_n)\}$.

$(x_i, y_i) \sim_{iid} D$

We want an algorithm $\mathcal{M}$ that outputs some $c' \in C$ and satisfies

$$\mathbb{P}(L_D(\mathcal{M}(X)) \leq \min_{c\in C} L_D(c) + \alpha) \geq 1 - \beta.$$

best possible loss achievable by C

Empirical risk minimization, again

Issue: We want to find $\arg\min_{c \in C} L_D(c)$, but we do not know D.

Solution: Instead we solve $\arg\min_{c \in C} L_X(c)$, where

$$L_X(c) = \frac{1}{n} \sum_{i=1}^{n} \ell(c, (x_i, y_i)).$$

is the empirical error.

for binary loss: $= \frac{|\{i : c(x_i) \neq y_i\}|}{n}$

Theorem (Uniform convergence)
Suppose that $n \geq \frac{\ln(|C|/\beta)}{2\alpha^2}$. *Then, with probability* $\geq 1 - \beta$,

$$\max_{c \in C} L_D(c) - L_X(c) \leq \alpha$$

Example: Linear Separators

$\langle u, v \rangle = \sum_{i=1}^{d} u_i v_i$ | $\text{sign}(z) = \begin{cases} +1 & z > 0 \\ 0 & z = 0 \\ -1 & z < 0 \end{cases}$

- $\mathcal{X} = [0,1]^d$ unit cube in $\mathbb{R}^d$
- C is all functions of the type $c_\theta(x) = \text{sign}(\langle x, \theta \rangle + \theta_0)$ for $\theta \in \mathbb{R}^d, \theta_0 \in \mathbb{R}$.

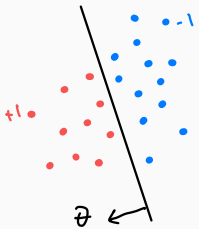
For convenience, replace x by $(x, 1) \in [0,1]^{d+1}$ and θ, θ_0 by $(\theta, \theta_0) \in \mathbb{R}^{d+1}$

$c_\theta(x) = \text{sign}(\langle x, \theta \rangle)$

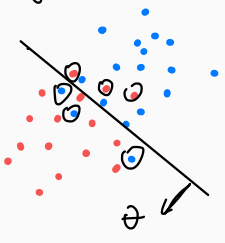
$c_\theta(x) = \begin{cases} +1 & \text{if } x \text{ "below" the plane} \\ -1 & \text{if above} \end{cases}$

$\langle x, \theta \rangle + \theta_0 = \left\langle \begin{pmatrix} x \\ 1 \end{pmatrix}, \begin{pmatrix} \theta \\ \theta_0 \end{pmatrix} \right\rangle$

Realizable



Agnostic

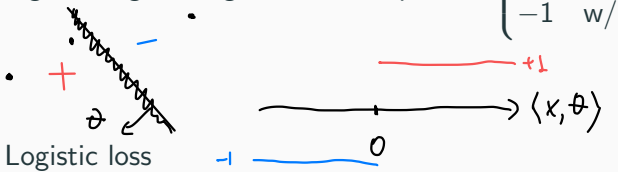


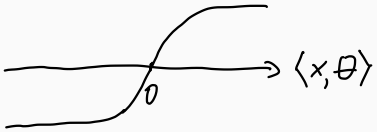
From now, will ignore θ_0

→ Finding best seperator is generally computationally hard

Logistic Regression

Sign for sigmoid: given θ and x predict $\begin{cases} +1 & \text{w/ prob. } \frac{1}{1+e^{-\langle x,\theta\rangle}} \\ -1 & \text{w/ prob. } \frac{1}{1+e^{\langle x,\theta\rangle}} \end{cases}$





Logistic loss

$$\ell(\theta,(x,y)) = \log\left(\frac{1}{\mathbb{P}(\text{predict } y \text{ from } \langle x,\theta\rangle)}\right) = \log(1+e^{-y\cdot\langle x,\theta\rangle}).$$

c_θ

Logistic regression: Given $X = \{(x_1,y_1),\dots,(x_n,y_n)\}$ solve

$$\arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \log(1+e^{-y\cdot\langle x,\theta\rangle})$$

Can be efficiently solved

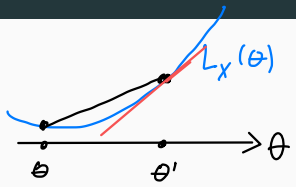
Empirical loss

Connection to population loss in another class

(Private) Gradient Descent

Convex loss

The function $L_X(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log(1 + e^{-y \cdot \langle x, \theta \rangle})$ is **convex** in θ.



$$\forall \theta, \theta' \in \mathbb{R}^{d+1} : \quad L_X\left(\frac{\theta+\theta'}{2}\right) \le \frac{1}{2} L_X(\theta) + \frac{1}{2} L_X(\theta')$$

$$\Updownarrow \qquad \forall \lambda \in [0,1] \quad L_X(\lambda\theta + (1-\lambda)\theta') \le \lambda L_X(\theta) + (1-\lambda) L_X(\theta')$$

$$\forall \theta, \theta' : \quad L_X(\theta) \ge L_X(\theta') + \langle \nabla L_X(\theta'), \theta - \theta' \rangle$$

$$\nabla L_X(\theta) = \begin{pmatrix} \frac{\partial L_X(\theta)}{\partial \theta_0} \\ \vdots \\ \frac{\partial L_X(\theta)}{\partial \theta_d} \end{pmatrix}$$

Convex functions can be minimized efficiently.

- for non-convex, it's complicated

Gradient descent

Logistic regression

$$\arg\min_{\theta \in B_2^{d+1}(R)} \frac{1}{n}\sum_{i=1}^{n} \log(1 + e^{-y_i \cdot \langle x_i, \theta \rangle})$$

$L_X(\theta)$

$B_2^{d+1}(R) = \{\theta : \|\theta\|_2 \le R\}$, $\|\theta\|_2 = \sqrt{\sum \theta_i^2}$

parameter

$$\nabla L_X(\theta) = \begin{pmatrix} \frac{\partial L_X(\theta)}{\partial \theta_0} \\ \vdots \\ \frac{\partial L_X(\theta)}{\partial \theta_d} \end{pmatrix}$$

$-\nabla L_X(\theta)$: direction in which $L_X(\theta)$ decreases the fastest

$\theta_0 = 0$

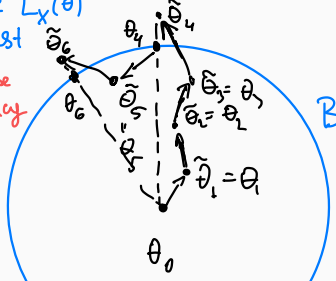
for $t = 1 \dots T-1$ **do**

$\tilde{\theta}_t = \theta_{t-1} - \eta(\nabla L_X(\theta_{t-1}) + Z_t)$ — Gauss noise for privacy

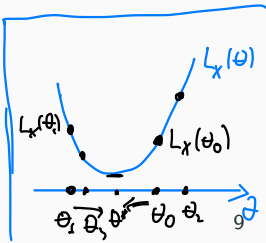
$\theta_t = \tilde{\theta}_t / \max\{1, \|\tilde{\theta}_t\|_2 / R\}$ — projection into $B_2^{d+1}(R)$

end for

output $\frac{1}{T}\sum_{t=0}^{T-1} \theta_t$



$B_2(R)$



Advanced composition - warmup

Publish k functions $f_1, \ldots, f_k : \mathcal{X}^n \to \mathbb{R}^d$ with (ε, δ)-DP, where $\forall i \;\; \Delta_2 f_i \le C$

Want to release $f_1(X), f_2(X), \ldots, f_k(X)$. Can achieve noise / function $\simeq C\sqrt{k \log k}$

f_i could be adaptive: f_i depends on $f_1(X)+Z_1, \ldots, f_{i-1}(X)+Z_{i-1}$

1) Apply Gaussian mechanism k times, use composition; $\Rightarrow$ noise $\simeq Ck\log k$

Release $f_i(X) + Z_i \quad Z_i \sim N\left(0, \frac{C^2}{\rho} I\right)$ for $\rho \simeq \frac{\varepsilon^2/k^2}{\log(k/\delta)}$, to achieve $(\frac{\varepsilon}{k}, \frac{\delta}{k})$-DP

By composition thm $(f_1(X)+Z_1, \ldots, f_k(X)+Z_k)$ is (ε, δ)-DP

2) $g(X) = \begin{pmatrix} f_1(X) \\ f_2(X) \\ \vdots \\ f_k(X) \end{pmatrix} \in \mathbb{R}^{dk}$; Use the Gaussian noise mechanism for g. $\Rightarrow$ noise $\simeq C\sqrt{k}$

$(\Delta_2 g)^2 = \max_{X \sim X'} \|g(X) - g(X')\|_2^2 = \max_{X \sim X'} \sum_{i=1}^{k} \|f_i(X) - f_i(X')\|_2^2 \le kC^2$

Advanced composition (for Gaussian noise)

$f_i : \mathcal{X}^n \to \mathbb{R}^d$

Suppose we realease $Y_1 = f_1(X) + Z_1, \ldots, Y_k = f_k(X) + Z_k$ where $f_i : \mathcal{X}^n \to \mathbb{R}^d$ depends also on $Y_1, \ldots, Y_{i-1}$.

$Z_i \in \mathbb{R}^d \sim \mathrm{N}\left(0, \frac{(\Delta_2 f)^2}{\rho} \cdot I\right)$

Then the output $(Y_1, \ldots, Y_k)$ satisfies (ε, δ)-DP for $\varepsilon = k\rho + \sqrt{2k\rho \ln(1/\delta)}$.

$\Rightarrow$ for $\rho \simeq \frac{\varepsilon^2}{k \log(1/\delta)}$ to achieve (ε, δ)-DP

Same as if queries were not adaptive

$\Rightarrow$ noise per query f_i is $\simeq \frac{\Delta_2 f \cdot \sqrt{k} \cdot \sqrt{\log 1/\delta}}{\varepsilon}$

Sensitivity of gradients

$X = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ $\quad \nabla L_X(\theta) = \frac{1}{n} \sum_{i=1}^{n} \nabla \log\left(1 + e^{-y_i \cdot \langle x_i, \theta \rangle}\right)$

$X' = \{(x_1, y_1), \ldots, (x_i', y_i'), \ldots, (x_n, y_n)\}$

Suppose $\mathcal{X} = [-1, +1]^{d+1}$.

$$\Delta_2 \nabla L_X(\theta) = \max_{X \sim X'} \left\| \nabla L_X(\theta) - \nabla L_{X'}(\theta) \right\|_2 = \max_{\substack{(x,y) \\ (x',y')}} \frac{1}{n} \left\| \nabla \log\left(1 + e^{-y \cdot \langle x, \theta \rangle}\right) - \nabla \log\left(1 + e^{-y' \langle x', \theta \rangle}\right) \right\|_2$$

$$\leq \frac{2}{n} \cdot \max_{(x,y)} \left\| \nabla \log\left(1 + e^{-y \langle x, \theta \rangle}\right) \right\|_2$$

$$\leq \frac{2\sqrt{d+1}}{n}$$

$$\nabla \log(1 + e^{-y\langle x, \theta \rangle}) = -\frac{1}{1 + e^{y\langle x, \theta \rangle}} yx$$

$y \in \{\pm 1\}$ $\quad x \in [-1, +1]^{d+1}$

$$\left\| \nabla \log\left(1 + e^{-y\langle x, \theta \rangle}\right) \right\|_2 = \left\| \frac{1}{1 + e^{y\langle x, \theta \rangle}} \cdot yx \right\|_2 \leq \|x\|_2 \leq \sqrt{d+1}$$

Private gradient descent

Think of $\nabla L_X(\theta_0), \nabla L_X(\theta_1), \ldots$ as the adaptively chosen functions $f_1, f_2, \ldots$,

$\theta_0 = 0$

for $t = 1 \ldots T-1$ **do**

$\tilde{\theta}_t = \theta_{t-1} - \eta(\nabla L_X(\theta_{t-1}) + Z_t)$

$\theta_t = \tilde{\theta}_t / \max\{1, \|\tilde{\theta}_t\|_2 / R\}$

end for

output $\frac{1}{T}\sum_{t=0}^{T-1} \theta_t$

$\forall_t$

$$Z_t \sim N\left(0, \frac{4d+1}{n^2\rho} I\right) \quad \| \sigma^2$$

$$\rho \approx \frac{\varepsilon^2}{T \log(1/\delta)}$$

$$\sigma^2 \approx \frac{dT \cdot \log(1/\delta)}{\varepsilon^2 n^2}$$

(ε, δ)-DP by advanced composition + post-processing

Accuracy analysis

Proof in the notes (optional)

Theorem
Suppose $\mathbb{E}\|\nabla L_X(\theta_t) + Z_t\|_2^2 \le B^2$ *for all* t. *For* $\eta = \frac{R}{B\sqrt{T}}$ *we have*

$$\mathbb{E}\left[L_X\left(\frac{1}{T}\sum_{t=0}^{T-1}\theta_t\right)\right] \le \min_{\theta \in B_2^{d+1}(R)} L_X(\theta) + \frac{RB}{T^{1/2}}$$

optimal value

goes to 0 as $T \to \infty$

Plugging in

$$\mathbb{E}\|\nabla L_X(\theta_t) + Z_t\|_2^2 = \mathbb{E}\|\nabla L_X(\theta_t)\|_2^2 + \underbrace{\mathbb{E}\|Z_t\|_2^2}_{\mathbb{E}\sum_{i=0}^{d}(Z_t)_i^2 = \sigma^2(d+1)} \leq \cancel{d+1} + \sigma^2(d+1) = B^2$$

$$\sigma^2 \approx \frac{d T \cdot \log(1/\delta)}{\varepsilon^2 n^2}$$

$$B^2 \approx d + \frac{d^2 T \log(1/\delta)}{\varepsilon^2 n^2}$$

For any θ $\|\nabla L_X(\theta)\|_2 \leq \sqrt{d+1}$

$$\|\nabla L_X(\theta)\|_2 = \left\| \frac{1}{n} \sum \nabla \log\left(1 + e^{-y_i \langle x_i, \theta \rangle}\right) \right\|_2$$

$$\leq \frac{1}{n} \sum \left\| \nabla \log\left(1 + e^{-y_i \langle x_i, \theta \rangle}\right) \right\|_2 \leq \sqrt{d+1}$$

$$\text{error} = \frac{RB}{T^{1/2}} \approx \frac{R\sqrt{d}}{T^{1/2}} + \frac{Rd\sqrt{\log 1/\delta}}{\varepsilon n}$$