Optimization vs Privacy in Machine Learning

Vianney Perchet

Séminaire Palaisien
Saclay, January 2020

Crest, ENSAE & Criteo AI Lab, Paris

Introduction

- **Typical** learning pb. Dataset $\mathcal{D} = \{X_1, \ldots, X_n\}$, metric $U(x, a)$

Learn/compute/optimize α^*: $\max_\alpha \mathbb{E}_{\mathcal{P}}[U(X, \alpha(X))]$

- **Assumption:** X_n iid $\sim \mathcal{P}$ unknown, $a = \alpha(X)$ is algo's **decisions**

  Examples Classification, ERM, etc. output some α^*

Learning α^* **irrelevant**, its **implementation**/rolling-out matters

- The Dataset $\mathcal{D}$ might be sensitive; **protect** it !
- Competitors/clients **do the same concurrently**.

  Implementing α^* reveals **private/valuable** information.

Differential Privacy

“Re-identification” - no privacy

The 'Re-Identification' of Governor William Weld's Medical Information: A Critical Re-Examination of Health Data Identification Risks and Privacy Protections, Then and Now

19 Pages • Posted: 4 Jun 2012 • Last revised: 3 Sep 2015

Daniel Barth-Jones
Columbia University - Mailman School of Public Health, Department of Epidemiology

Date Written: July 2012

Abstract

The 1997 re-identification of Massachusetts Governor William Weld’s medical data within an insurance data set which had been stripped of direct identifiers has had a profound impact on the development of de-identification provisions within the 2003 Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. Weld’s re-identification, purportedly achieved through the use of a voter registration list from Cambridge, MA is frequently cited as an example that computer scientists can re-identify individuals within de-identified data with “astonishing ease”. However, a careful re-examination of the population demographics in Cambridge indicates that Weld was most likely re-identifiable only because he was a public figure who experienced a highly publicized hospitalization rather than there being any certainty underlying his re-identification using the Cambridge voter data, which had missing data for a large proportion of the population.

User/Local Differential Privacy

Dataset $\mathcal{D} = \{X_1, \ldots, X_n\}$ iid Bernoulli param. p

- Construct public dataset $\widetilde{\mathcal{D}} = \{\widetilde{X}_1, \ldots, \widetilde{X}_n\}$
  - **Estimate** $\widehat{p}$ from $\mathcal{D}$ (with some accuracy)
  - without revealing X_i (with high proba)

$$\widetilde{X}_i = \begin{cases} \text{random} & \text{w.p. } 1-\varepsilon \\ X_i & \text{w.p. } \varepsilon \end{cases}$$

- Simple computations

$$\frac{1}{n}\sum_{i=1}^{n} \widetilde{X}_i = \frac{1-\varepsilon}{2} + \varepsilon p \pm \sqrt{\frac{1-\varepsilon}{n}} \pm \sqrt{\frac{p}{n}}$$

  - Accuracy: estimate p if $n \gg \frac{1}{\varepsilon^2 p^2}$
  - Privacy: $\mathbb{P}\{X_i = 1 | \widetilde{X}_i = 1\} \simeq p(1+\varepsilon)$

ε-Differential Privacy

Dataset $\mathcal{D} = \{X_1, \ldots, X_n\}$, **query** $f : \mathcal{D} \to \mathbb{R}^d$, but **privately**

- Examples of query functions
  - $f(\mathcal{D}) = (X_1, \ldots, X_n)$
    $= (X_1, \frac{X_1+X_2}{2}, \ldots, \overline{X}_n)$
- ε-diff private random query $\mathcal{A} : \mathcal{D} \to \mathbb{R}^d$

$$e^{-\varepsilon}\mathbb{P}\{\mathcal{A}(\mathcal{D}_1) \in \mathfrak{E}\} \ \leq \ \mathbb{P}\{\mathcal{A}(\mathcal{D}_0) \in \mathfrak{E}\} \ \leq \ e^{\varepsilon}\mathbb{P}\{\mathcal{A}(\mathcal{D}_1) \in \mathfrak{E}\}$$

  where $\mathcal{D}_0$ and $\mathcal{D}_1$ differ by 1 datapoint
- ”easy” solution: **Additive Laplace Noise**.
  - $\mathcal{A}(\mathcal{D}) = f(\mathcal{D}) + Y$ with Y_i independent Laplace(λ)
  - Optimal choice $\lambda = \frac{\max \|f(\mathcal{D}_0 - \mathcal{D}_1\|_1}{\varepsilon}$

Privacy vs Utility

Privacy for users

- I want to visit websites I like .... **without Google knowing** how ~~depraved~~ sophisticated I am
- I want to watch Netflix **without being classified** as white/male/(sadly in the end of his) 30's

It happens !

Film fans see red over Netflix 'targeted' posters for black viewers

The streaming service's customers say they are being duped by marketing that shows minor cast members as leading characters

▲ Set It Up is made to look like a two-hander between Taye Diggs and Lucy Liu, rather than the white couple. Photograph: Twitter Kelly Quantrill @codetrill

A concrete Example

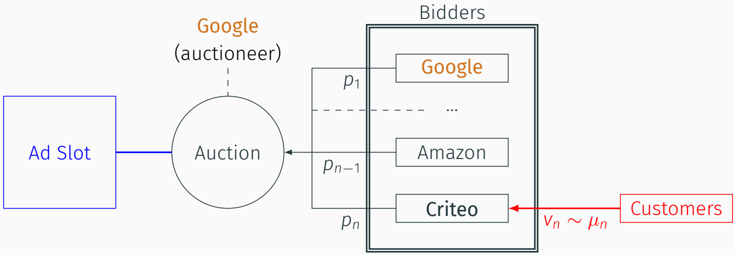

Figure 1: Online advertisement auction system

Criteo is **both** a client and a **competitor** of Google.

- Want to "exploit" good clients

  Without revealing their quality (as in poker)

A simple model

$$\max_{x \in \mathcal{X} \subset \mathbb{R}^d} x^\top c_k; \qquad c_k \in \mathbb{R}^d$$

- $k \in \{1, \ldots, K\}$ is the private **type** (known only to agent)

  public prior $\pi_0 \in \Delta_K$, i.e., $k \sim \pi_0$
- Privacy "value" is the **amount of info** leaked on k

  example. $KL(\pi_0, \pi_1)$, with π_1 posterior on k
- (Vectors $c_1, \ldots, c_K$ publicly known)

What is the posterior π_1 ?

$$\max_{x \in \mathcal{X} \subset \mathbb{R}^d} x^\top c_k; \qquad k \sim \pi_0$$

- Given $k \in \{1, \ldots, K\}$, choose $x \sim \mu_k \in \mathcal{P}(\mathcal{X})$
- $\pi_1 \in \Delta_K$ posterior knowing x

$$\pi^k_{1_{|x}} = \frac{\pi_0^k \mu_k(x)}{\sum_j \pi_0^j \mu_j(x)} \qquad \text{Bayes}$$

Private Learning Objective

$$\max_{\mu_1,\ldots,\mu_K} \sum_k \pi_0^k \mathbb{E}_{x\sim\mu_k}\left[x^\top c_k - \lambda KL(\pi_{1_{|x}}, \pi_0)\right]$$

or more general

$$\inf_{\gamma\in\mathcal{P}(\mathcal{X}\times[K]);p_1\sharp\gamma=\pi_0} \int c(x,k) + \lambda D(\pi_x, \pi_0) d\gamma(x,k)$$

The f-divergence case

f convex, $f(1) = 0$

$$D(P, Q) = \mathbb{E}_{x \sim Q}\left[f\left(\frac{p(x)}{q(x)}\right)\right]$$

- $KL(Q, P) = \mathbb{E}_{x \sim Q}\left[-\log\left(\frac{p(x)}{q(x)}\right)\right]$
- $KL(P, Q) = \mathbb{E}_{x \sim Q}\left[\frac{p(x)}{q(x)} \log\left(\frac{p(x)}{q(x)}\right)\right]$
- $TV(P, Q) = \mathbb{E}_{x \sim Q}\left[\frac{1}{2}\left|\frac{p(x)}{q(x)} - 1\right|\right]$

Convexity Result

- f convex, $f(1) = 0$

$$\inf_{\gamma \in \mathcal{P}(\mathcal{X} \times [K]); p_1 \sharp \gamma = \pi_0} \int c(x,k) d\gamma + \lambda \int \mathbb{E} f(\frac{d\pi_{1_{|x}}}{d\pi_0}) d\gamma$$

- Convex program in γ !

  $\rightarrow$ solvable in theory

- But in **infinite dimension**

  $\rightarrow$ not in **practice**

Finiteness Result

If K is finite, finiteness Theorems.

- $\forall \varepsilon > 0$, exists ε-optimal γ with finite support of size $K(K+2)$
- $\mathcal{X}$ compact and $c(\cdot, k)$ lsc true for $\varepsilon = 0$
- **Finite Reformulation**

$$\inf \sum_{i,k} \gamma_{i,k} c(x^i, k) + \lambda \sum_{i,k,j} \gamma_{i,k} \pi_0^j f\left(\frac{\gamma_{i,j}}{\pi_0^j \sum_\ell \gamma_{i,\ell}}\right)$$

where $\gamma \in \mathbb{R}^{(K+2)K}$, $X \in \mathbb{R}^{K+2}$ and the constraint $\sum_i \gamma_{i,j} = \pi_0^j$.

Finite but no longer convex !

Ex: the linear case ; Difference of Convex

$$c(x,k) = x^\top c_k + \beta_k, \qquad \mathcal{X} = [-1,1]^d$$

$$\inf_\gamma - \sum_i \| \sum_k \gamma_{i,k} c_k \|_1 + \sum_k \pi_0^k \beta_k + \lambda \sum_{i,k,j} \gamma_{i,k} \pi_0^j f\left(\frac{\gamma_{i,j}}{\pi_0^j \sum_\ell \gamma_{i,\ell}}\right)$$

$$= -G(\gamma) + F(\gamma)$$

can be solved with DC solver

(F and G are convex)

The special case of KL-divergence and Optimal Transport

$$\inf_{\gamma \in \mathcal{P}(\mathcal{X} \times [K]); p_1 \sharp \gamma = \nu} \int c(x,k) d\gamma + \lambda \mathbb{E} \log(\frac{d\pi_x}{d\nu}) d\gamma$$

I renamed the prior ν so that equivalent to

$$\inf_{\mu \in \mathcal{P}(\mathcal{X})} \left\{ \inf_{\pi \in T(\mu,\nu)} \int c d\pi + \lambda \int \log\left(\frac{d\pi}{d\mu d\nu}\right) d\pi \right\}$$

$$\inf_{\mu \in \mathcal{P}(\mathcal{X})} OT_{c,\lambda}(\mu, \nu)$$

Regularized Optimal Transport

$$OT_{c,\lambda}(\mu, \nu) = \inf_{\pi \in T(\mu,\nu)} \int c d\pi + \lambda \int \log \left(\frac{d\pi}{d\mu d\nu}\right) d\pi$$

1. **Solve** with Sinkhorn algo (in π)

   $\rightarrow$ highly parallelisable

   $\rightarrow$ closed form iteration

   $\Rightarrow$ works **very well** in practice

2. **Optimize** (in μ) !

Parametric Family

$$\min_{\mu} OT_{c,\lambda}(\mu, \nu)$$

1. Look for $\mu_{\theta} = \sum_{j=1}^{n} \alpha_j(\theta)\delta_{x_j(\theta)}$
2. Compute $\frac{\partial}{\partial \alpha} OT_{c,\lambda}(\mu, \nu)$ and $\frac{\partial}{\partial x} OT_{c,\lambda}(\mu, \nu)$

   either by automatic diff, or solve the dual.

Experiments

Expe 1: Toy, linear example

$c(x, y) = x^\top y$, $\mathcal{X} = [-1, 1]^d$, $K = 100$ and D=KL.

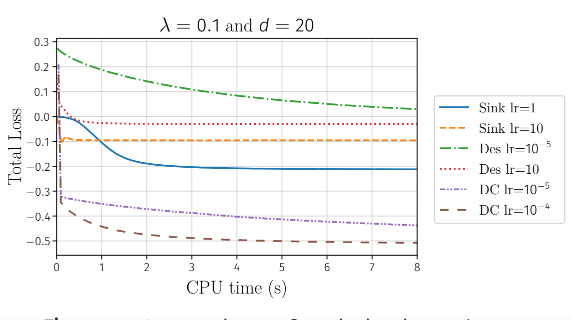

Figure 2: Comparison of optimization schemes

- DC $\succ$ Sink $\succ$ Des
- Adaptation to problem structure is primordial

Expe 2: online repeated auctions

- **Auctions:** value $v \sim \mu_{y_j} = \mathrm{Exp}(\frac{1}{y_j})$
- Bid strategy $\beta_i^j(v)$ induces fake distribution $x_i = \beta_i^j \# \mu_{y_j}$
- With $\mathrm{Exp}(\frac{1}{y_j})$ reduce to strategies $\beta_i^j(v) = \beta_i(v/y_j)$

  (those $\beta_i(\cdot)$ parametrized by a NN)
- Cost functions $c(\beta_i^j, y_j)$ can be computed [previous paper]

Expe 2: online repeated auctions

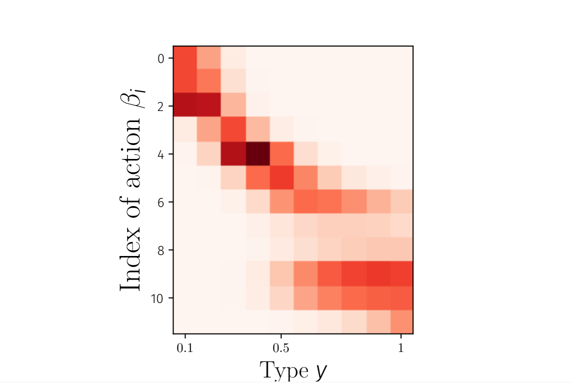

Figure 3: Joint distribution heat-map, with $\lambda = 0.01$

Expe 2: online repeated auctions

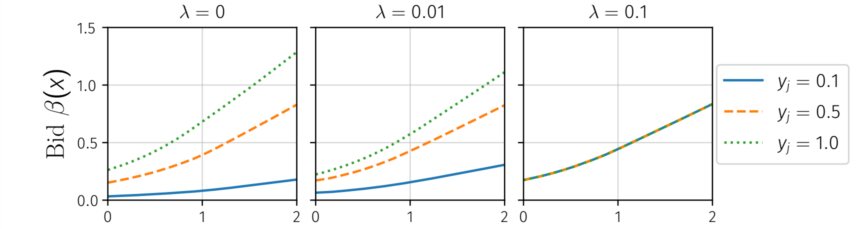

Figure 4: Evolution of the most used β_i with the type and the regularization constant

So many open questions

- Statistical guarantees
- Computational issues
- Private optimization algo (query AWS repeatedly)
- General f-divergence...

Alternative concepts/valuations of privacy, fairness ?