Lecture 2: Uniform Convergence and Optimization

Ruhr-University Bochum

April 29, 2019

Concentration inequalities: Nuts and bolts

- Markov:

$$\mathbb{P}\Big(|X| \geq \varepsilon\Big) \leq \frac{\mathbb{E}|X|}{\varepsilon}$$

Concentration inequalities: Nuts and bolts

- Markov:

$$\mathbb{P}\Big(|X| \geq \varepsilon\Big) \leq \frac{\mathbb{E}|X|}{\varepsilon}$$

- Chebychev:

$$\mathbb{P}\Big(\frac{|X - \mathbb{E}X|}{\sqrt{var(X)}} \geq \varepsilon\Big) \leq \frac{1}{\varepsilon^2}$$

Concentration inequalities: Nuts and bolts

- Markov:

$$\mathbb{P}\Big(|X| \geq \varepsilon\Big) \leq \frac{\mathbb{E}|X|}{\varepsilon}$$

- Chebychev:

$$\mathbb{P}\Big(\frac{|X - \mathbb{E}X|}{\sqrt{var(X)}} \geq \varepsilon\Big) \leq \frac{1}{\varepsilon^2}$$

- Chernov: X_i iid, in $[0, 1]$

$$\mathbb{P}\Big(\sum_{i=1}^{n} X_i - \mathbb{E}X_i \geq \varepsilon\Big) \leq \exp(-2\varepsilon^2/(2n))$$

Statistical queries

- A **statistical query** is a (measurable) function $\phi : \mathcal{X} \to [0, 1]$.
- Notation: $\mathcal{D}$ is some probability distribution on $\mathcal{X}$.
  Population mean:

$$\phi(\mathcal{D}) := \mathbb{E}_{x \sim \mathcal{D}} \phi(x)$$

  $\mathcal{S} \sim \mathcal{D}^n$ iid data sample consisting of $X_1, ..., X_n$.
  Empirical mean:

$$\phi(\mathcal{S}) = \frac{1}{n} \sum_{i=1}^{n} \phi(X_i)$$

M, the mediating mechanism

How do we model the situation of interest?

- A's aim: Find out features of $\mathcal{D}$: $\phi_1(\mathcal{D}), ..., \phi_K(\mathcal{D})$.
- Problem: Neither $\mathcal{D}$ is known, nor is $\mathcal{S}$ directly accessible.
- A interacts with a mechanism M, which returns answers $a_1, ..., a_K$ to his queries.

M, the mediating mechanism

How do we model the situation of interest?

- A's aim: Find out features of $\mathcal{D}$: $\phi_1(\mathcal{D}), ..., \phi_K(\mathcal{D})$.
- Problem: Neither $\mathcal{D}$ is known, nor is $\mathcal{S}$ directly accessible.
- A interacts with a mechanism M, which returns answers $a_1, ..., a_K$ to his queries.

If M is in charge of some data base with sensitive information, A will not always get all the information.

M, the mediating mechanism

How do we model the situation of interest?

- A's aim: Find out features of $\mathcal{D}$: $\phi_1(\mathcal{D}), ..., \phi_K(\mathcal{D})$.
- Problem: Neither $\mathcal{D}$ is known, nor is $\mathcal{S}$ directly accessible.
- A interacts with a mechanism M, which returns answers $a_1, ..., a_K$ to his queries.

If M is in charge of some data base with sensitive information, A will not always get all the information.
Examples of M:

1. Empirical mechanism: Returns $\phi_j(\mathcal{S})$ for ϕ_j.
2. Privatizing mechanism: Returns $\phi_j(\mathcal{S}) + L_j$ for ϕ_j. L_j is some Laplace noise.

Which question? Whose answer?

A theoretical measure of performance is given by

$$err_{\mathcal{S}}(M, A) = \max_{j=1,\dots,K} |\phi_j(\mathcal{D}) - a_j|$$

We say M is (α, β)-**accurate** for K queries on iid data for every analyst A if

$$\mathbb{P}\big(err_{\mathcal{S}}(M, A) \leq \alpha\big) \geq 1 - \beta.$$

Sometimes we look at expected errors such as

$$\sup_{\mathcal{D}} \sup_{A} \mathbb{E} err_{\mathcal{S}}(M, A).$$

or even

$$\inf_{M} \sup_{\mathcal{D}} \sup_{A} \mathbb{E} err_{\mathcal{S}}(M, A).$$

Example for expected errors

The next Theorem demonstrates that a high accuracy is feasable via the empirical mechanism M_{emp}.

Theorem 3:

Let $\mathcal{D}$ be any probability distribution on $\mathcal{X}$,$\phi_1, ..., \phi_K$ (data independent) statistical queries of the analyst A and $\mathcal{S} \sim \mathcal{D}^n$ iid data. Then with probability $\geq 1 - \delta$

$$err_{\mathcal{S}}(M_{emp}, A) \leq \sqrt{\frac{\log(2K/\delta)}{2n}}.$$

Proof:
Last time we have seen that with probability $\geq 1-\alpha$ for any query

$$|\phi(\mathcal{S})-\phi(\mathcal{D})| \leq \sqrt{\frac{\log(2/\alpha)}{2n}}.$$

Choosing $\alpha = \delta/K$ we see that

$$\mathbb{P}\left(\exists j \in \{1,\ldots,K\} : |\phi_j(\mathcal{S})-\phi_j(\mathcal{D})| > \sqrt{\frac{\log(2K/\delta)}{2n}}\right)$$
$$\leq \sum_{j=1}^{K} \mathbb{P}\left(|\phi_j(\mathcal{S})-\phi_j(\mathcal{D})| > \sqrt{\frac{\log(2K/\delta)}{2n}}\right) \leq \delta.$$

$\square$

Theorem 3:

Corollary:

Under the Assumptions of Theorem 3:

$$\mathbb{E}err_{\mathcal{S}}(M_{emp}, A) = \mathcal{O}\left(\sqrt{\frac{\log(2K)}{2n}}\right)$$

Stochastic optimization

Statistical models are often determined by a parameter $w \in \Theta \subset \mathbb{R}^d$, which can be expressed as the minimizer of an averaged loss function, i.e.

$$w^* = argmin_{w \in \Theta} \mathbb{E}_{x \sim \mathcal{D}} \ell(w, x).$$

How do we find out w^*? A typical estimator is the empirical minimizer

$$\hat{w}_{emp} = argmin_{w \in \Theta} \sum_{i=1}^{n} \ell(w, X_i).$$

This srategy can only be successful if ℓ is "well behaved" in some fashion.

Stochastic optimization

Assumptions:

- $\Theta \subset \{u : \|u\| \leq R\}$.
- For all $x \in \mathcal{X}$

$$|\ell(u;x) - \ell(v;x)| \leq |u - v|C \quad u, v \in \Theta.$$

Theorem 5:

Under the above assumption for iid data $\mathcal{S} \sim \mathcal{D}^n$ it holds with probability $\geq 1 - \delta$:

$$\sup_{u \in \Theta} |\sum_{i=1}^{n} \ell(u;X_i) - \mathbb{E}_{x \sim \mathcal{D}} \ell(u;x)| \leq 6RC\sqrt{\frac{d \log(n/\delta)}{n}}.$$

Convex optimization

Theorem 5 suggests minimizing the empirical loss

$$\hat{w}^* = \text{argmin}_w \sum_{i=1}^{n} \ell(w; X_i).$$

How difficult is minimizing this?
For the broad class of convex optimization problems this task is efficiently solvable.

Subdifferential:

The subdifferential of a function $f : \mathbb{R}^d \supset \Theta \to \mathbb{R}$ in a point x is defined as

$$\partial f(x) := \{g \in \mathbb{R}^d : f(x) + < g, y - x > \leq f(y) \forall y \in \Theta\}.$$

Convex optimization

Convexity:

1. A set $\Theta \subset \mathbb{R}^d$ is called convex if it equals a (possibly infinite) intersection of halfspaces.
2. Let $\Theta \subset \mathbb{R}^d$ be a convex set. A function $f : \Theta \to \mathbb{R}$ is convex on Θ iff $\partial f(x) \neq \emptyset$ for all $x \in \Theta$.

Remarks:

Let Θ be closed and convex. The projection

$$\Pi_\Theta(x) := \text{argmin}_{w \in \Theta} \|x - w\|$$

is well defined for all $x \in \mathbb{R}^d$. Furthermore for all $w \in \Theta$ and $y \in \mathbb{R}^d$ projection reduces distances, i.e.

$$\|\Pi_\Theta(y) - w\| \leq \|y - w\|$$

Projected Gradient Descent

Let $f : \Theta \to \mathbb{R}$ be a function and Θ a convex subset of $\mathbb{R}^d$. The method of PGD is defined as follows:

1. Choose some $x_0 \in \Theta$, $\eta > 0$ and $T \in \mathbb{N}$.
2. Set $y_{t+1} = x_t - \eta g_t$, where $g_t \in \partial f(x_t)$
3. Set $x_{t+1} = \Pi_\Theta(y_{t+1})$.
4. If $t + 1 = T$ stop and output x_T . Else set $t = t + 1$ and repeat 2.

Theorem 10:

Let $\Theta \subset \{u : \|u\| \leq R\}$ be closed, convex and $f : \Theta \to \mathbb{R}$ be convex and C-lipschitz. If we run PGD T times with $\eta = R/(C\sqrt{T})$, then

$$f\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) - f(x^*) \leq RC/\sqrt{T},$$

where $x^* \in \Theta$ is the minimizer of f.