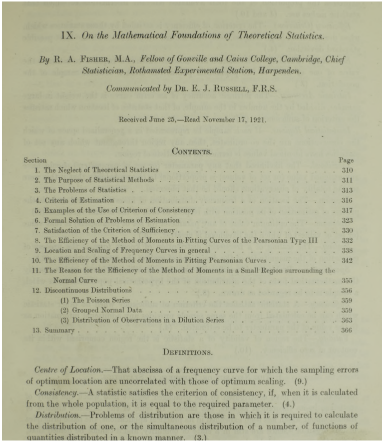
Adapting Maximum Likelihood Theory to Modern Applications

Feng Ruan

Stanford University

December 3, 2020

Ronald Fisher & Maximum Likelihood Estimation



IX. *On the Mathematical Foundations of Theoretical Statistics.*

By R. A. Fisher, M.A., *Fellow of Gonville and Caius College, Cambridge, Chief Statistician, Rothamsted Experimental Station, Harpenden.*

Communicated by Dr. E. J. Russell, F.R.S.

Received June 25,—Read November 17, 1921.

Contents.

Definitions.

Centre of Location.—That abscissa of a frequency curve for which the sampling errors of optimum location are uncorrelated with those of optimum scaling. (9.)

Consistency.—A statistic satisfies the criterion of consistency, if, when it is calculated from the whole population, it is equal to the required parameter. (4.)

Distribution.—Problems of distribution are those in which it is required to calculate the distribution of one, or the simultaneous distribution of a number, of functions of quantities distributed in a known manner. (3.)

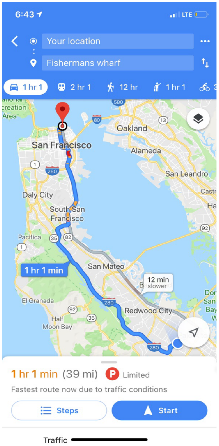
Broad Impact of Classical MLE

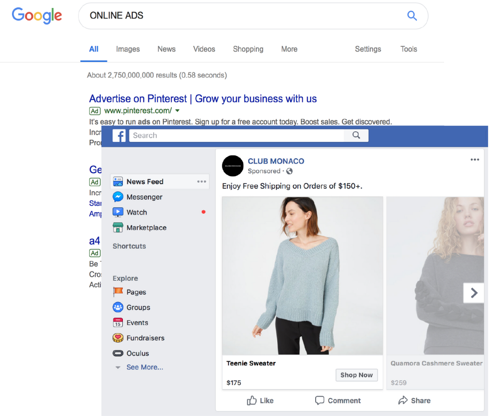
$$\text{Classical MLE:} \quad \hat{\theta}_n = \operatorname*{argmax}_{\theta} \sum_{i=1}^{n} \log p_{\theta}(X_i)$$

- Generality
  - Write down the probability model
  - Maximize the likelihood w.r.t parameter
- Optimality [Fisher, Cramér, Rao, Stein, Hájek, Le Cam, Bickel...]

Online Learning

- We need FAST algorithms to give real-time update and predictions.
  - Online advertising, Online recommendation system, Google maps...



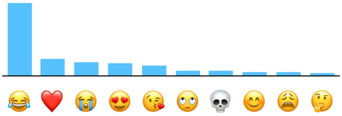


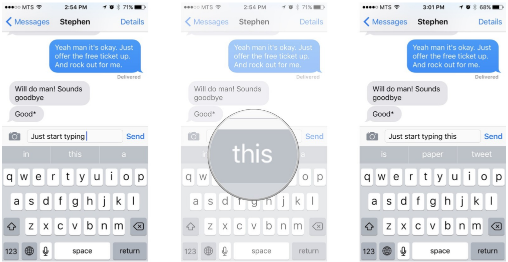
Privacy Concerns

- Many data of interest contains sensitive/personal information
  - health records, genetic data, internet browsing history, etc.

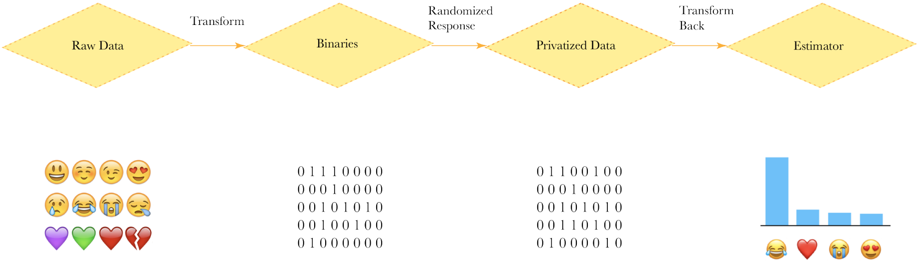
Data Privacy in Practice

- Next word prediction (auto-completion)
- Ranking emojis



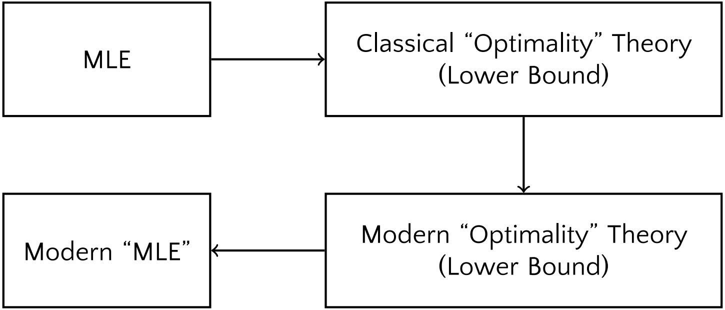


Private Algorithm



This Talk

Towards a general recipe to generate optimal procedures
for modern applications.



A First Question

What criterion should we use to define "optimality"?

Minimax Criterion

A principled way to define “optimality"—minimax criterion [Von Neumann 28, Wald 39]

$$\min_{\hat{\theta}_n} \max_{\theta \in \Theta} \mathbb{E}_\theta \left[L(\hat{\theta}_n(X_1, \ldots, X_n), \theta) \right]$$

- Θ: parameter space
- L: loss function

Criticism:

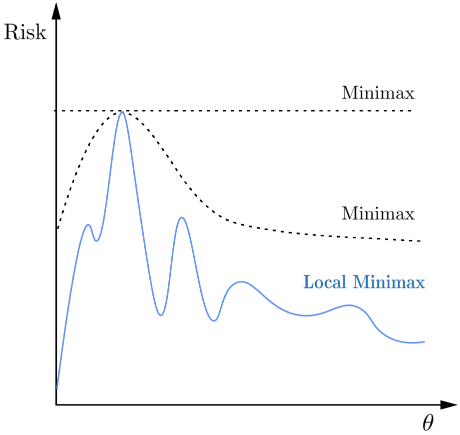
- Conservative when Θ is large.
- Statistician: I do not care worst case. I only care my problem at hand.

"Minimax optimal" for all of machine learning

- Goal: Predict binary label Y from X.
- Worst case: X independent of Y.
- Random guess is minimax optimal!

What optimality criterion characterizes MLE?

- Classical MLE is not just simply optimal for the worst-case problem.
- It is optimal for problems of all difficulties.

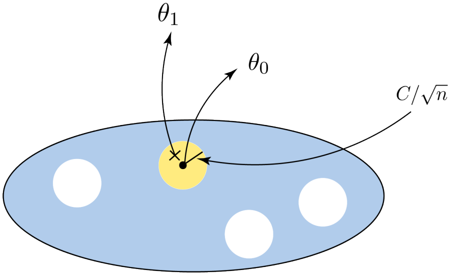


Definition: Local Minimax Risk

$$\mathfrak{M}_n := \inf_{\hat{\theta}_n} \sup_{\theta \in \Theta} \mathbb{E}_\theta \left[L(\hat{\theta}_n, \theta) \right].$$

$$\mathfrak{M}_n^{\mathrm{loc}}(\theta_0) := \sup_{\theta_1 \in \Theta} \inf_{\hat{\theta}_n} \max_{\theta \in \{\theta_0, \theta_1\}} \mathbb{E}_\theta \left[L(\hat{\theta}_n, \theta) \right]$$

$$\approx \inf_{\hat{\theta}_n} \sup_{\|\theta - \theta_0\| \le \frac{C}{\sqrt{n}}} \mathbb{E}_\theta \left[L(\hat{\theta}_n, \theta) \right].$$



- A localized quantity characterizing the difficulty of estimation for $\theta = \theta_0$.
  [Stein 56', Donoho & Liu 87', 91', Cai & Low 15']

Local Minimax Criterion–the Right Criterion

- Classical local minimax theory: [Stein 56', Donoho & Liu 87', 91', Cai & Low 15']

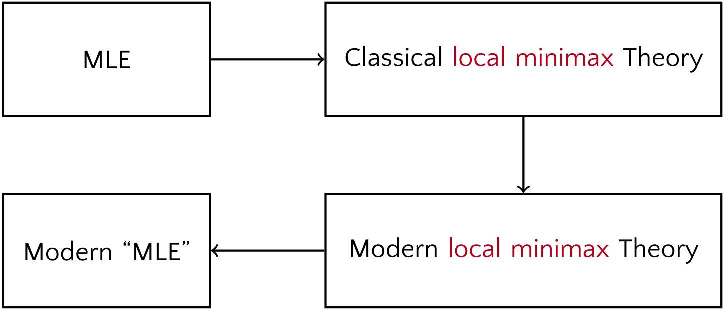
$$\mathfrak{M}_n^{\mathrm{loc}}(\theta_0) \asymp \mathbb{E} L\left(\theta_0 + \frac{1}{\sqrt{n}} W, \theta_0\right) \text{ where } W \sim \mathsf{N}(0, I_{\theta_0}^{-1}).$$

- Optimality of MLE:

$$\sqrt{n}\left(\hat{\theta}_n^{\mathrm{mle}} - \theta_0\right) \xrightarrow{d} \mathsf{N}(0, I_{\theta_0}^{-1}).$$

An Overview of the General Strategy

- Classical local minimax theory $\Rightarrow$ optimality of classical MLE.
- Outline:
  1. Private local minimax theory $\Rightarrow$ Private "MLE"
  2. Online local minimax theory $\Rightarrow$ Online "MLE"

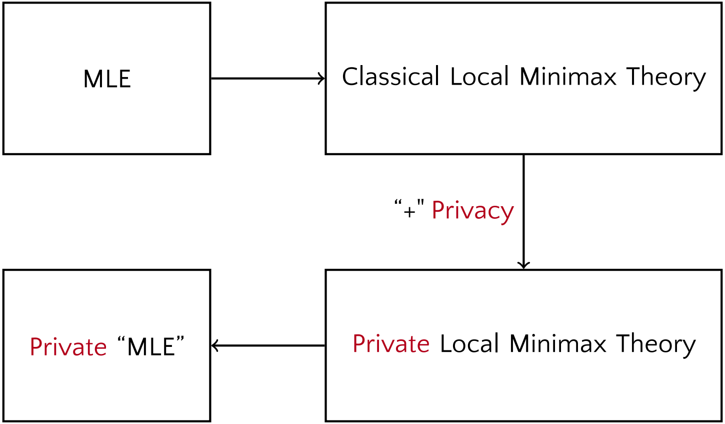


Outline

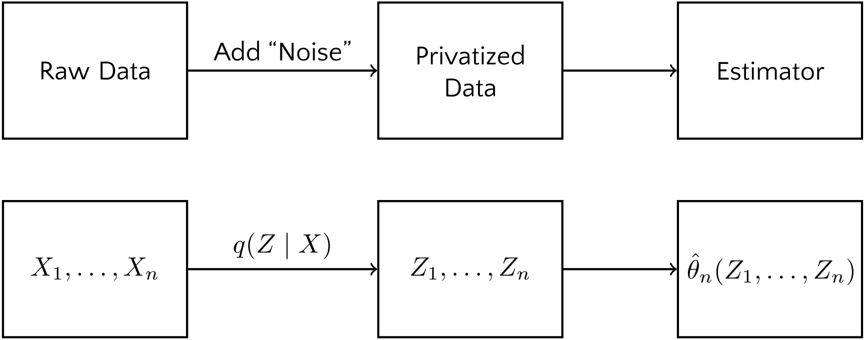
- Privacy
- Online learning
- Future direction

Privacy

An overview of the strategy



Private Data Analysis



- Valid $q \in \mathcal{Q}$: $Z_1, \ldots, Z_n$ preserve privacy of $X_1, \ldots, X_n$.

Local Minimax Framework for Privacy

$$\mathfrak{M}_n^{\mathrm{loc}}(\theta_0) := \sup_{\theta_1 \in \Theta} \inf_{\hat{\theta}_n} \max_{\theta \in \{\theta_0, \theta_1\}} \mathbb{E}_\theta \left[L(\hat{\theta}_n, \theta) \right].$$

$$\boxed{X_1, \ldots, X_n} \xrightarrow{\mathrm{q}(Z \mid X)} \boxed{Z_1, \ldots, Z_n} \xrightarrow{\hat{\theta}_n} \boxed{\hat{\theta}_n(Z_1, \ldots, Z_n)}$$

- Valid $q \in \mathcal{Q}$: $Z_1, \ldots, Z_n$ preserve privacy of $X_1, \ldots, X_n$.

Local Minimax Framework for Privacy

$$\mathfrak{M}_n^{\mathrm{loc}}(\theta_0) := \sup_{\theta_1 \in \Theta} \inf_{\hat{\theta}_n} \max_{\theta \in \{\theta_0, \theta_1\}} \mathbb{E}_\theta \left[L(\hat{\theta}_n, \theta) \right].$$

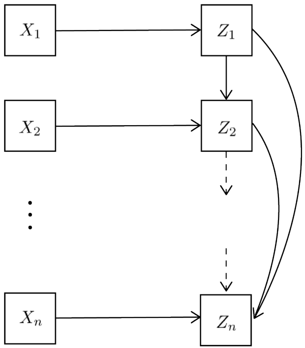
$$\mathfrak{M}_{n,\mathrm{priv}}^{\mathrm{loc}}(\theta_0) := \sup_{\theta_1 \in \Theta} \inf_{\mathrm{q} \in \mathcal{Q}} \inf_{\hat{\theta}_n} \max_{\theta \in \{\theta_0, \theta_1\}} \mathbb{E}_\theta \left[L(\hat{\theta}_n, \theta) \right].$$

$$\boxed{X_1, \ldots, X_n} \xrightarrow{\mathrm{q}(Z \mid X)} \boxed{Z_1, \ldots, Z_n} \xrightarrow{\hat{\theta}_n} \boxed{\hat{\theta}_n(Z_1, \ldots, Z_n)}$$

- Valid $q \in \mathcal{Q}$: $Z_1, \ldots, Z_n$ preserve privacy of $X_1, \ldots, X_n$.

Definition of Privacy

$$q(Z \mid X) = \prod_{i=1}^{n} q(z_i \mid x_i, z_{1:(i-1)})$$



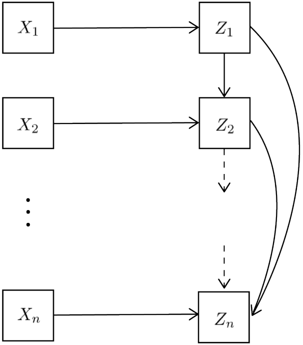
- ϵ-Differential Privacy [Dwork, McSherry, Nissim, Smith, 06]: for $x_i, x_i', z_{1:(i-1)}$,

$$\frac{q(z_i \mid x_i, z_{1:(i-1)})}{q(z_i \mid x_i', z_{1:(i-1)})} \leq \exp(\epsilon).$$

- One can't distinguish x_i and x_i' by looking at z_i, conditioning on $z_{1:(i-1)}$.

Definition of Privacy

$$q(Z \mid X) = \prod_{i=1}^{n} q(z_i \mid x_i, z_{1:(i-1)})$$



- ϵ-Differential Privacy [Dwork, McSherry, Nissim, Smith, 06]: for $x_i, x_i', z_{1:(i-1)}$,

$$\frac{q(z_i \mid x_i, z_{1:(i-1)})}{q(z_i \mid x_i', z_{1:(i-1)})} \le \exp(\epsilon).$$

- One can't distinguish x_i and x_i' by looking at z_i, conditioning on $z_{1:(i-1)}$.

- Average ϵ-Privacy [Mironov 17, Duchi & R. 18]: for $x_i, x_i', z_{1:(i-1)}$,

$$\mathbb{E}_{q(z_i \mid x_i, z_{1:(i-1)})}\left[\frac{q(z_i \mid x_i, z_{1:(i-1)})}{q(z_i \mid x_i', z_{1:(i-1)})}\right] \le \exp(\epsilon).$$

Private Information $I_{\theta_0,\mathrm{priv}} \neq I_{\theta_0}$

Consider a 1-dim $\mathcal{P} = \{P_\theta\}_{\theta\in\mathbb{R}}$, and the log likelihood $\ell_\theta = \log p_\theta$.

Theorem (Classical local minimax theory: Donoho & Liu 87, 91)

$$\mathfrak{M}_n^{\mathrm{loc}}(\theta_0) \asymp \mathbb{E}[L(\theta_0 + n^{-1/2}W, \theta_0)] \ \textit{for}\ W \sim \mathsf{N}(0, I_{\theta_0}^{-1}),$$

where $I_{\theta_0} = \mathbb{E}[|\dot{\ell}_{\theta_0}|^2]$ *is the classical Fisher information.*

Private Information $I_{\theta_0,\mathrm{priv}} \neq I_{\theta_0}$

Consider a 1-dim $\mathcal{P} = \{P_\theta\}_{\theta \in \mathbb{R}}$, and the log likelihood $\ell_\theta = \log p_\theta$.

Theorem (Classical local minimax theory: Donoho & Liu 87, 91)

$$\mathfrak{M}_n^{\mathrm{loc}}(\theta_0) \asymp \mathbb{E}[L(\theta_0 + n^{-1/2} W, \theta_0)] \;\; \textit{for} \;\; W \sim \mathsf{N}(0, I_{\theta_0}^{-1}),$$

where $I_{\theta_0} = \mathbb{E}[|\dot{\ell}_{\theta_0}|^2]$ *is the classical Fisher information.*

Theorem (Private local minimax theory: Duchi & R. 18)

$$\mathfrak{M}_{n,\mathrm{priv}}^{\mathrm{loc}}(\theta_0) \asymp \mathbb{E}[L(\theta_0 + (n\epsilon^2)^{-1/2} W, \theta_0)] \;\; \textit{for} \;\; W \sim \mathsf{N}(0, I_{\theta_0,\mathrm{priv}}^{-1}).$$

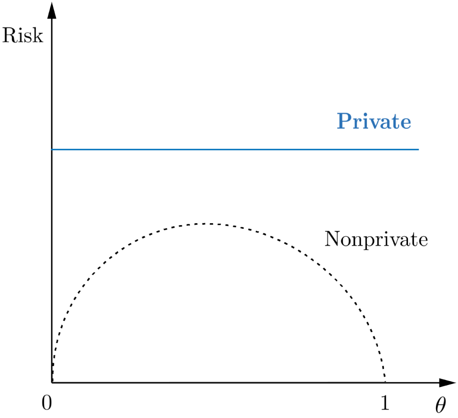
where $I_{\theta_0,\mathrm{priv}} = (\mathbb{E}[|\dot{\ell}_{\theta_0}|])^2$ *is defined to be the private information.*

Remark:

- Our private "MLE" achieves the private information lower bound (Later).
- Superefficiency result [Duchi & R. 18'] (not discussed in this talk).

The cost of privacy: Bernoulli estimation

Consider $X_i \overset{\text{iid}}{\sim} \mathrm{Ber}(\theta)$. ($\mathbb{P}(X = 1) = 1 - \mathbb{P}(X = 0) = \theta$).



Non-private:

Classical Fisher Information:

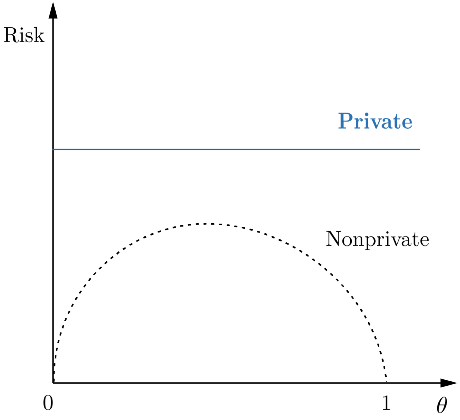
$$I_\theta = (\theta(1-\theta))^{-1}.$$

Classical MLE $\bar{X}_n$:

$$\sqrt{n}\left(\bar{X}_n - \theta\right) \xrightarrow{d} \mathsf{N}(0, \theta(1-\theta)).$$

The cost of privacy: Bernoulli estimation

Consider $X_i \overset{\text{iid}}{\sim} \mathrm{Ber}(\theta)$. ($\mathbb{P}(X = 1) = 1 - \mathbb{P}(X = 0) = \theta$).



Private:

Private Information:

$$I_{\theta,\mathrm{priv}} = 1$$

Randomized Response [Warner 65']:

$$Z_i = \epsilon^{-1} \cdot \begin{cases} X_i & \text{w.p. } \frac{1+\epsilon/2}{2} \\ 1 - X_i & \text{w.p. } \frac{1-\epsilon/2}{2} \end{cases}$$

Private "MLE" estimator $\bar{Z}_n$:

$$\sqrt{n\epsilon^2}\left(\bar{Z}_n - \theta\right) \xrightarrow{d} \mathsf{N}(0,1).$$

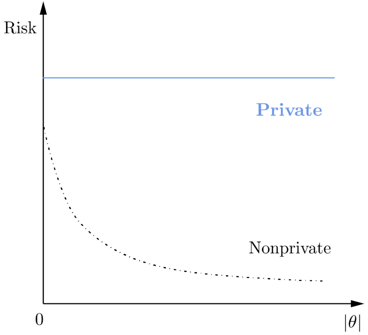
The cost of privacy: binary logistic regression

Let (X_i, Y_i) i.i.d satisfy the logistic model:

$$\mathbb{P}(Y = 1 \mid X = x) = \frac{1}{1 + \exp(-\theta x)}.$$

Prediction error loss:

$$L(\theta, \theta_0) = \mathbb{E}_{(X,Y)\sim P_{\theta_0}} \left[|\mathbb{P}_\theta(Y \mid X) - \mathbb{P}_{\theta_0}(Y \mid X)| \right]$$



Non-private:

Local risk of classical MLE:

$$\mathfrak{M}_n^{\mathrm{loc}}(\theta_0) \asymp \frac{1}{\sqrt{n}} \exp(-|\theta_0|/2)$$

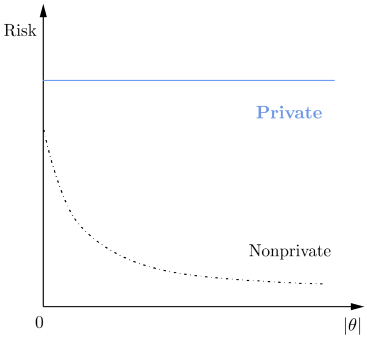
The cost of privacy: binary logistic regression

Let (X_i, Y_i) i.i.d satisfy the logistic model:

$$\mathbb{P}(Y = 1 \mid X = x) = \frac{1}{1 + \exp(-\theta x)}.$$

Prediction error loss:

$$L(\theta, \theta_0) = \mathbb{E}_{(X,Y)\sim P_{\theta_0}} \left[|\mathbb{P}_\theta(Y \mid X) - \mathbb{P}_{\theta_0}(Y \mid X)| \right]$$



Private:

Local risk of Private "MLE":

$$\mathfrak{M}^{\text{loc}}_{n,\text{priv}}(\theta_0) \asymp \frac{1}{\sqrt{n\epsilon^2}}$$

Challenge: how to construct private "MLE"?

Question: For one-dimensional family $\{P_\theta\}_{\theta \in \mathbb{R}}$, how do we construct $\hat{\theta}_{n,\mathrm{priv}}$

$$\sqrt{n\epsilon^2}(\hat{\theta}_{n,\mathrm{priv}} - \theta_0) \xrightarrow{d} \mathsf{N}(0, I_{\theta_0,\mathrm{priv}}^{-1}) \text{ where } I_{\theta_0,\mathrm{priv}} = (\mathbb{E}|\dot{\ell}_{\theta_0}|)^2.$$

Theorem (Private local minimax theory: Duchi & R. 18)

$$\mathfrak{M}^{\mathrm{loc}}_{n,\mathrm{priv}}(\theta_0) \asymp \mathbb{E}[L(\theta_0 + (n\epsilon^2)^{-1/2}W, \theta_0)] \text{ for } W \sim \mathsf{N}(0, I_{\theta_0,\mathrm{priv}}^{-1}).$$

where $I_{\theta_0,\mathrm{priv}} = (\mathbb{E}[|\dot{\ell}_{\theta_0}|])^2$ *is the private information.*

Challenge: how to construct private "MLE"?

Bernoulli: $X_i \overset{\text{iid}}{\sim} \mathrm{Ber}(\theta)$.

Private MLE:

- Randomized Response [Warner 65']:

$$Z_i = \epsilon^{-1} \cdot \begin{cases} X_i & \text{w.p. } \frac{1+\epsilon/2}{2} \\ 1 - X_i & \text{w.p. } \frac{1-\epsilon/2}{2} \end{cases}$$

- Estimation procedure:

$$\sqrt{n\epsilon^2}\left(\bar{Z}_n - \theta\right) \xrightarrow{d} \mathsf{N}(0,1).$$

Challenge: how to construct private “MLE”?

$$\text{General: } X_i \overset{\text{iid}}{\sim} P_\theta. \quad \sqrt{n\epsilon^2}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{d} \mathsf{N}(0, I_{\theta_0,\text{priv}}^{-1})$$

Idea: reduction to Bernoulli case:

$$\begin{tikzcd} X \arrow[r, "g"] & \underset{\in \{0,1\}}{g(X)} \arrow[r, "\text{RR}"] & Z \arrow[r] & \hat{\theta}_{n,\text{priv}} \end{tikzcd}$$

Challenge: how to construct private "MLE"?

General: $X_i \overset{\text{iid}}{\sim} P_\theta$. $\sqrt{n\epsilon^2}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{d} \mathsf{N}(0, I_{\theta_0,\text{priv}}^{-1})$

Idea: reduction to Bernoulli case:

$$X \xrightarrow{g} \underset{\in \{0,1\}}{g(X)} \xrightarrow{\text{RR}} Z \longrightarrow \hat{\theta}_{n,\text{priv}}$$

Estimating Equation:

$$\hat{\theta}_{n,\text{priv}} = \text{invert}_\theta \left\{ \mathbb{E}_\theta[g(X)] = \bar{Z}_n \right\}.$$

Challenge: how to construct private "MLE"?

$$\text{General: } X_i \overset{\text{iid}}{\sim} P_\theta. \quad \sqrt{n\epsilon^2}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{d} \mathsf{N}(0, I_{\theta_0,\text{priv}}^{-1})$$

Idea: reduction to Bernoulli case:

$$X \xrightarrow{g} \underset{\in\{0,1\}}{g(X)} \xrightarrow{\text{RR}} Z \longrightarrow \hat{\theta}_{n,\text{priv}}$$

Delta method:

$$\mathrm{Var}(\hat{\theta}_{n,\text{priv}}) = \left(\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}_\theta[g(X)]\right)^{-2} \cdot \mathrm{Var}(\bar{Z}_n)$$

Estimating Equation:

$$\hat{\theta}_{n,\text{priv}} = \mathrm{invert}_\theta\left\{\mathbb{E}_\theta[g(X)] = \bar{Z}_n\right\}.$$

Challenge: how to construct private "MLE"?

General: $X_i \overset{\text{iid}}{\sim} P_\theta$. $\sqrt{n\epsilon^2}\left(\hat{\theta}_n - \theta_0\right) \overset{d}{\to} \mathsf{N}(0, I_{\theta_0,\text{priv}}^{-1})$

Idea: reduction to Bernoulli case:

$$X \xrightarrow{g} \underset{\in \{0,1\}}{g(X)} \xrightarrow{\text{RR}} Z \longrightarrow \hat{\theta}_{n,\text{priv}}$$

Estimating Equation:

$$\hat{\theta}_{n,\text{priv}} = \text{invert}_\theta \left\{\mathbb{E}_\theta[g(X)] = \bar{Z}_n\right\}.$$

Delta method:

$$\text{Var}(\hat{\theta}_{n,\text{priv}}) = \left(\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}_\theta[g(X)]\right)^{-2} \cdot \text{Var}(\bar{Z}_n)$$

Requirement on $g(\cdot) \in \{0,1\}$:

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}_\theta[g(X)] \mid_{\theta=\theta_0} = I_{\theta_0,\text{priv}}^{1/2}.$$

Challenge: how to construct private "MLE"?

General: $X_i \overset{\text{iid}}{\sim} P_\theta$. $\sqrt{n\epsilon^2}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{d} \mathsf{N}(0, I^{-1}_{\theta_0,\text{priv}})$

Idea: reduction to Bernoulli case:

$$X \xrightarrow{g} \underset{\in \{0,1\}}{g(X)} \xrightarrow{\text{RR}} Z \longrightarrow \hat{\theta}_{n,\text{priv}}$$

Estimating Equation:

$$\hat{\theta}_{n,\text{priv}} = \text{invert}_\theta \left\{ \mathbb{E}_\theta[g(X)] = \bar{Z}_n \right\}.$$

Delta method:

$$\text{Var}(\hat{\theta}_{n,\text{priv}}) = \left(\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}_\theta[g(X)]\right)^{-2} \cdot \text{Var}(\bar{Z}_n)$$

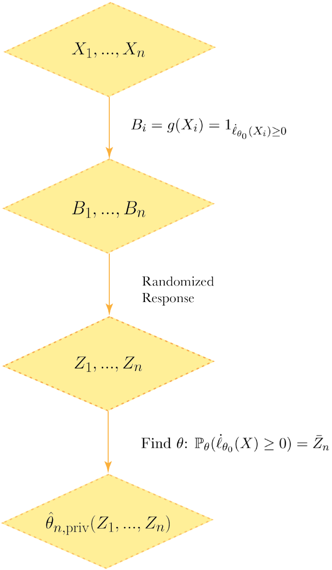
Requirement on $g(\cdot) \in \{0,1\}$:

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}_\theta[g(X)] \, |_{\theta=\theta_0} = I^{1/2}_{\theta_0,\text{priv}}.$$

Fact

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}_\theta\left[1_{\dot{\ell}_{\theta_0}(X)\geq 0}\right] |_{\theta=\theta_0} = \mathbb{E}_{\theta_0}|\dot{\ell}_{\theta_0}| = I^{1/2}_{\theta_0,\text{priv}}.$$

Algorithm: private MLE



Divide the data into two groups.

- Privatize the first group of data, get an initializer:

$$\hat{\theta}_{n,\text{init}} \xrightarrow{p} \theta_0.$$

- Transform the second group of data $X_1, \ldots, X_n$ into binaries $B_1, \ldots, B_n$:

$$B_i = \mathbf{1}\left\{\dot{\ell}_{\hat{\theta}_{n,\text{init}}}(X_i) \geq 0\right\}$$

- Privatize $B_1, \ldots, B_n$ with randomized response:

$$Z_i = \epsilon^{-1} \cdot \begin{cases} W_i & \text{w.p. } \frac{1+\epsilon/2}{2} \\ 1 - W_i & \text{w.p. } \frac{1-\epsilon/2}{2} \end{cases}$$

- Construct the final estimator $\hat{\theta}_{n,\text{priv}}$:

$$\hat{\theta}_{n,\text{priv}} = \text{invert}_\theta \left\{\mathbb{P}_\theta(\dot{\ell}_{\hat{\theta}_{n,\text{init}}}(X) \geq 0) = \bar{Z}_n\right\}.$$

Extension

- Functionals for high dimensional parametric models $\{P_\theta\}_{\theta \in \mathbb{R}^p}$.
- Model misspecification: true distribution $P \notin \{P_\theta\}_{\theta \in \Theta}$.
  - Ex: find the best linear predictor θ without assuming a linear model on P.

Extension

- Functionals for high dimensional parametric models $\{P_\theta\}_{\theta \in \mathbb{R}^p}$.
- Model misspecification: true distribution $P \notin \{P_\theta\}_{\theta \in \Theta}$.
  - Ex: find the best linear predictor θ without assuming a linear model on P.
- Private local minimax theory $\Rightarrow$ private information $\Rightarrow$ private MLE!

Simulation: Flow Cytometry Experiment

- Goal: predicting network structure linking the proteins using a real flow cytometry dataset [Hastie, Tibshirani & Friedman 09]
- Logistic regression:
$$\log \frac{\mathbb{P}(Y = 1 \mid X)}{\mathbb{P}(Y = 0 \mid X)} = \theta^T X.$$
where Y is the link prediction and X is the gene expression.
- Treat the raw data as population.
- Run vanilla logistic regression on the population (raw data) and get $\theta^{\star}$.
- Simulate new data from the population.
- Target: estimate $\theta^{\star} = (\theta_1^{\star}, \theta_2^{\star}, \dots, \theta_p^{\star})$.
- Compare non-private MLE, private local and private global minimax [Duchi, Jordan & Wainwright 16'] estimators for estimating $\theta^{\star} = (\theta_1^{\star}, \theta_2^{\star}, \dots, \theta_p^{\star})$.

Flow Cytometry Experiment

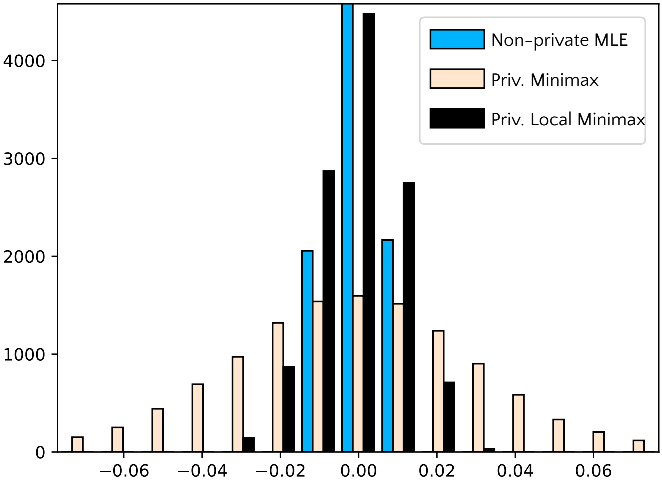


Figure: Histograms of errors across $T = 1000$ simulation experiments in estimation of the coordinates $\theta_1^\star, \theta_2^\star, \ldots, \theta_p^\star$ (privacy: $\epsilon = 1$)

Flow Cytometry Experiment

$\frac{\lVert \hat{\theta}_n - \theta^\star \rVert_2}{\lVert \theta^\star \rVert_2}$	Non-private	Private-Local	Private-Global
n/p = 5	42.2%	95%	>100%
n/p = 20	28.1%	64.8%	82.3%
n/p = 40	19%	42.5%	69.5%
n/p = 80	14.3%	30.2%	60%
n/p = 320	6.8%	13.8%	38.6%
n/p = 1280	3.4%	6.5%	20.2%

Table: Relative error of $\hat{\theta}_n$ across $T = 1000$ simulation experiments (privacy: $\epsilon = 1$).

Insights & Follow-ups

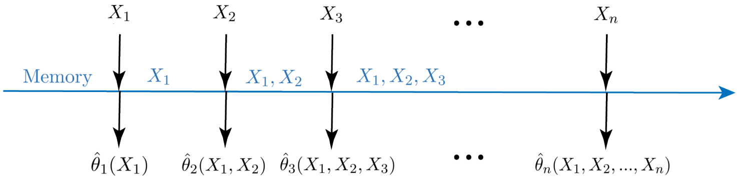
- Private local minimax procedure does lead to improvement.
- Privacy learning is challenging in high dimension or ϵ is low.
- Impacted Apple's privacy [Bhowmick et al '18].

Online Learning

Offline vs. Online Algorithm

$$\text{minimize}_\theta \ R(\theta) := \mathbb{E}_P[\ell(\theta; X)]$$

Offline algorithm:



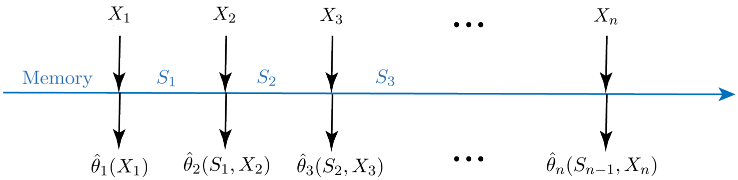
Example: empirical risk minimization (ERM)

$$\hat{\theta}_n = \underset{\theta}{\text{argmin}} \ \frac{1}{n} \sum_{i=1}^{n} \ell(\theta; X_i)$$

Offline vs. Online Algorithm

$$\text{minimize}_\theta \; R(\theta) := \mathbb{E}_P[\ell(\theta; X)]$$

Online algorithm:



Example: stochastic gradient descent ($S_t = \{\hat{\theta}_t\}$)

$$\hat{\theta}_{t+1} = \hat{\theta}_t - \alpha_t \nabla_\theta \ell(\hat{\theta}_t; X_t)$$

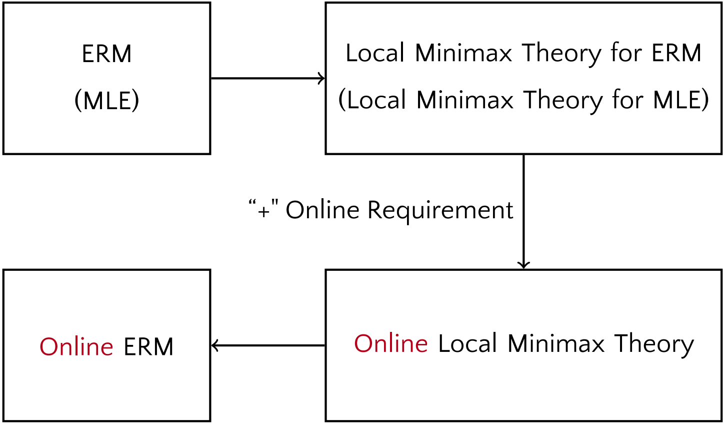
Problem

Find optimal online algorithm to solve convex and smooth problem:

$$\begin{aligned} \text{minimize} \quad & R(\theta) := \mathbb{E}_P[\ell(\theta; X)] \\ \text{subject to} \quad & \theta \in \Theta = \{c_i(\theta) \leq 0 : i = 1, 2, \ldots, m\}. \end{aligned}$$

Ex: Nonnegative least squares, Ridge, Lasso, (Regularized) Portfolio optimization...

Solution


ERM
(MLE)
Local Minimax Theory for ERM
(Local Minimax Theory for MLE)
“+" Online Requirement
Online ERM
Online Local Minimax Theory

Local Minimax Theory for Online Optimization

- Loss: $L\left(\sqrt{n}\left(\hat{\theta}_n - \theta_0\right)\right)$.
- $\mathfrak{M}^{\mathrm{loc}}_{\infty,\mathrm{on}}(\mathcal{P}_0)$: online local asymptotic minimax risk [Duchi & R. 18']
- $\mathfrak{M}^{\mathrm{loc}}_{\infty,\mathrm{off}}(\mathcal{P}_0)$: offline local asymptotic minimax risk [Hájek & Le Cam 70', 72', Levit 76', Bickel, Klassen, Ritov Wellner 93', Duchi & R. 18']

Theorem (Duchi & R. 18)

Assume regularity conditions on L. Then

$$\mathfrak{M}^{\mathrm{loc}}_{\infty,\mathrm{on}}(\mathcal{P}_0) = \mathfrak{M}^{\mathrm{loc}}_{\infty,\mathrm{off}}(\mathcal{P}_0) = \mathbb{E}[L(W)] \ \textit{for}\ W \sim \mathsf{N}\left(0, I^{\dagger}_{\mathcal{P}_0}\right)$$

Takehome Message:

$$I_{\mathcal{P}_0} = I_{\mathcal{P}_0,\mathrm{on}} = I_{\mathcal{P}_0,\mathrm{off}} = H\Sigma^{\dagger}H.$$

The upper bound

How do we construct the optimal online “ERM”?

$$\sqrt{n}\left(\hat{\theta}_{n,\mathrm{on}} - \theta_0\right) \to \mathsf{N}(0, I_{\mathcal{P}_0}^{\dagger}).$$

Unconstrained Problem

Population objective (no constraints):

$$\mathsf{minimize}_\theta \; R(\theta) = \mathbb{E}_{P_0}[\ell(\theta; X)]$$

Theorem (Polyak & Judisky 92 + Duchi & R. 18)

SGD averaging is optimal for unconstrained optimization:

$$\sqrt{n}\left(\hat{\theta}_{n,\mathrm{on}} - \theta^\star\right) \xrightarrow{d} \mathsf{N}(0, I^\dagger_{\mathcal{P}_0}).$$

Stochastic gradient descent (SGD):

$$\theta_{t+1} = \theta_t - \alpha_t \nabla \ell(\theta_t; X_t)$$

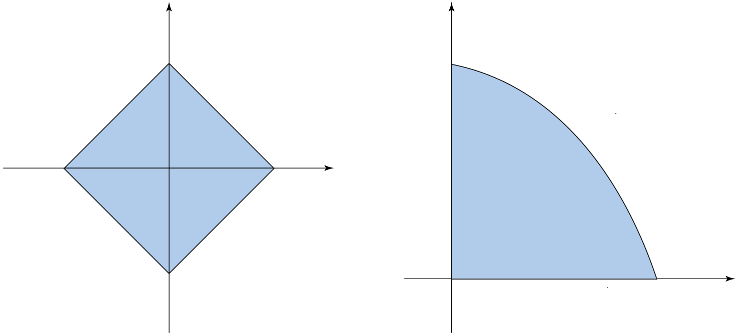
Keep track of the running average:

$$\hat{\theta}_{n,\mathrm{on}} = \bar{\theta}_n$$

Challenge

What about constrained optimization problems?

$$\begin{aligned} &\text{minimize}_\theta \; R(\theta) = \mathbb{E}_{P_0}[\ell(\theta; X)] \\ &\text{subject to } c_i(\theta) \leq 0, \; i = 1, \ldots, m. \end{aligned}$$

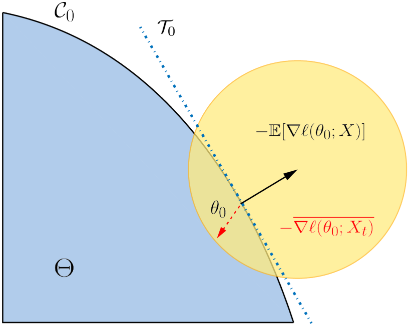


A Surprise: Projected-SGD fails

A Surprise: Projected-SGD fails

Projected stochastic gradient descent (PSGD):

$$\theta_{t+1} = \Pi_\Theta(\theta_t - \alpha_t \nabla \ell(\theta_t; X_t))$$



Failure: [Duchi & R. 18]

$$I_{\mathcal{P}_0} = H \Sigma^\dagger H$$

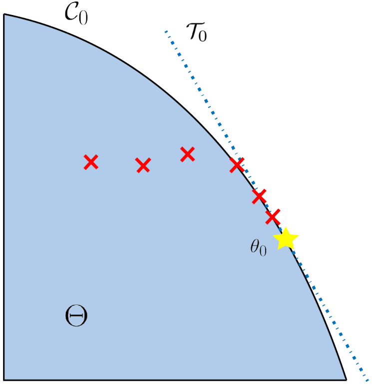
where

$$\Sigma = \Pi_{\mathcal{T}_0} \mathrm{Cov}_{P_0}(\nabla \ell(\theta_0; X)) \Pi_{\mathcal{T}_0}$$

A Surprise: Projected-SGD fails

Projected stochastic gradient descent (PSGD):

$$\theta_{t+1} = \Pi_\Theta(\theta_t - \alpha_t \nabla \ell(\theta_t; X_t))$$



Insight: need identify $\mathcal{C}_0$

$$I_{\mathcal{P}_0} = H \Sigma^\dagger H$$

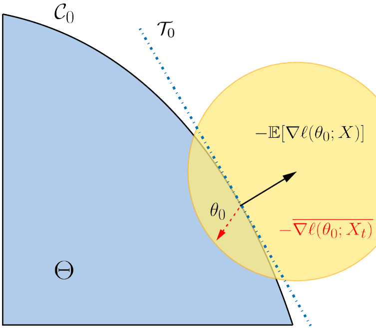
where

$$\Sigma = \Pi_{\mathcal{T}_0} \mathrm{Cov}_{P_0}(\nabla \ell(\theta_0; X)) \Pi_{\mathcal{T}_0}$$

A Fix: Dual Averaging?

Dual Averaging (DA) [Nesterov 07']:

$$z_t = -\frac{1}{t}\sum_{k=1}^{t} \nabla \ell(\theta_k; X_k) \text{ and } \theta_t = \Pi_\Theta(\alpha_t z_t)$$



Insight: averaging stabilizes noise

$$z_t \approx -\mathbb{E}[\nabla \ell(\theta_0; X)] + \overline{\text{noise}}_t.$$

Theorem (Duchi & R. 18)

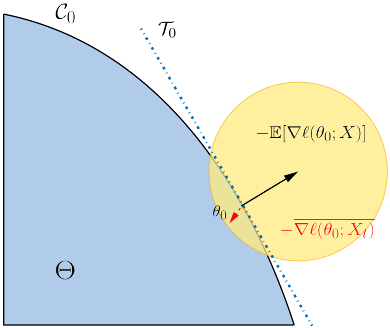
DA identifies the active constraints, i.e.,

$$\theta_t \in \mathcal{C}_0 \text{ eventually.}$$

A Fix: Dual Averaging?

Dual Averaging (DA) [Nesterov 07']:

$$z_t = -\frac{1}{t}\sum_{k=1}^{t} \nabla \ell(\theta_k; X_k) \text{ and } \theta_t = \Pi_\Theta(\alpha_t z_t)$$



Insight: averaging stabilizes noise

$$z_t \approx -\mathbb{E}[\nabla l(\theta_0; X)] + \overline{\text{noise}}_t.$$

Theorem (Duchi & R. 18)

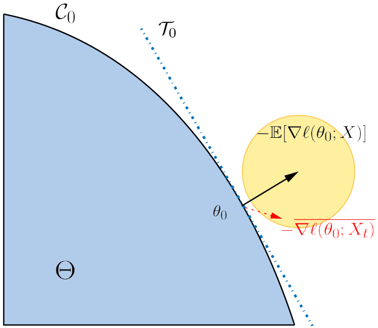
DA identifies the active manifold, i.e.,

$$\theta_t \in \mathcal{C}_0 \text{ eventually.}$$

A Fix: Dual Averaging?

Dual Averaging (DA) [Nesterov 07']:

$$z_t = -\frac{1}{t}\sum_{k=1}^{t} \nabla \ell(\theta_k; X_k) \text{ and } \theta_t = \Pi_\Theta(\alpha_t z_t)$$



Insight: averaging stabilizes noise

$$z_t \approx -\mathbb{E}[\nabla l(\theta_0; X)] + \overline{\text{noise}}_t.$$

Theorem (Duchi & R. 18)

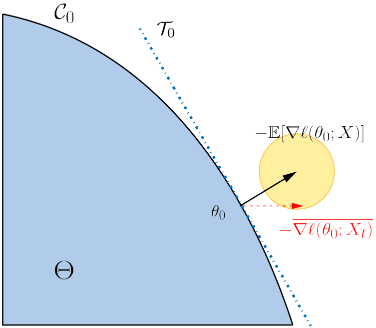
DA identifies the active manifold, i.e.,

$$\theta_t \in \mathcal{C}_0 \text{ eventually.}$$

A Fix: Dual Averaging?

Dual Averaging (DA) [Nesterov 07']:

$$z_t = -\frac{1}{t}\sum_{k=1}^{t} \nabla \ell(\theta_k; X_k) \text{ and } \theta_t = \Pi_\Theta(\alpha_t z_t)$$



Insight: averaging stabilizes noise

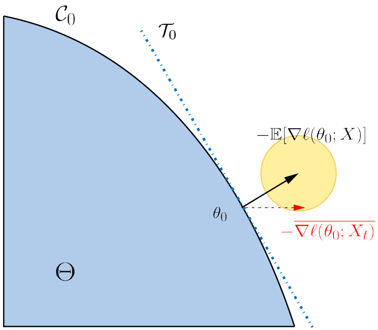
$$z_t \approx -\mathbb{E}[\nabla l(\theta_0; X)] + \overline{\text{noise}}_t.$$

Theorem (Duchi & R. 18)

DA identifies the active manifold, i.e.,

$$\theta_t \in \mathcal{C}_0 \text{ eventually.}$$

A Fix: Dual Averaging?



Observation [Duchi & R. 18]:

DA does not adapt to curvature.

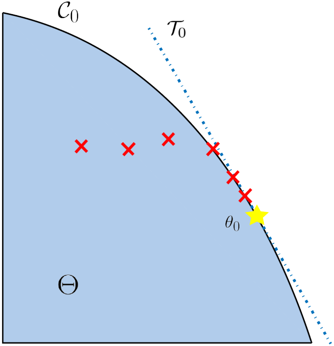
Failure:

$$I_{\mathcal{P}_0} = H \Sigma^{\dagger} H.$$

New Algorithm: Riemannian dual averaging

High Level Idea: Alternate between (variants of) DA and Riemannian SGD.

(see [Duchi & R. 18'] for details of the algorithm)



Theorem (Duchi & R. 18)

$$\sqrt{n}\left(\hat{\theta}_{n,\mathrm{RDA}} - \theta^\star\right) \xrightarrow{d} \mathsf{N}(0, I_{P_0}^\dagger).$$

Summary

Online information:

$$I_{P_0} = H\Sigma^{\dagger}H.$$

Algorithm	Adapt to Σ (identify constraints)	Adapt to H
Projected-SGD	✗	✗
Dual Averaging	✓	✗
RDA	✓	✓

Theorem (Duchi & R. 18)

$$\sqrt{n}\left(\hat{\theta}_{n,\mathrm{RDA}} - \theta^{\star}\right) \xrightarrow{d} \mathsf{N}(0, I_{P_0}^{\dagger}).$$

Conclusion

Takehome Message

- Towards a general recipe to optimal procedures for modern applications