Computational Barriers in Statistical Estimation

Andrea Montanari

Stanford University

July 9, 2017

Statistical estimation/Statistical learning

Class of models ($\Theta \subseteq \mathbb{R}^d$)

$$\mathcal{C}_\Theta \equiv \{\, \mathbb{P}_{\boldsymbol{\theta}} : \quad \boldsymbol{\theta} \in \Theta\}$$

Data

$$\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n \sim_{iid} \mathbb{P}_{\boldsymbol{\theta}_0}(\,\cdot\,)$$

Estimate θ_0 from data $x_1^n = (x_1, \dots, x_n)$

Statistical estimation/Statistical learning

Class of models $(\Theta \subseteq \mathbb{R}^d)$

$$\mathcal{C}_\Theta \equiv \{\, \mathbb{P}_{\boldsymbol{\theta}} : \quad \boldsymbol{\theta} \in \Theta\}$$

Data

$$\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n \sim_{iid} \mathbb{P}_{\boldsymbol{\theta}_0}(\,\cdot\,)$$

Estimate $\boldsymbol{\theta}_0$ from data $\boldsymbol{x}_1^n = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$

Minimax theory

Loss function

$$
\begin{aligned}
L : & \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R} \\
& (\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) \mapsto L(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0)
\end{aligned}
$$

Minimax risk

$$
R_n^*(\Theta) = \inf_{\hat{\boldsymbol{\theta}}(\,\cdot\,)} \sup_{\boldsymbol{\theta}_0 \in \Theta} \mathbb{E}_{\boldsymbol{\theta}_0} L(\hat{\boldsymbol{\theta}}(\boldsymbol{x}_1^n), \boldsymbol{\theta}_0)
$$

[Wald, 1950]

Kindergarten example

$$\mathbb{P}_{\boldsymbol{\theta}}(\,\cdot\,) = \mathsf{N}(\boldsymbol{\theta}, \mathrm{I}_d)$$
$$L(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}) = \|\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}\|_2^2$$
$$\Theta = \mathbb{R}^d.$$

Theorem

$$R_n^*(\Theta) = \frac{d}{n}\,.$$

- Foundation of the least squares, maximum likelihood, ...

Kindergarten example

$$\mathbb{P}_{\boldsymbol{\theta}}(\,\cdot\,) = \mathsf{N}(\boldsymbol{\theta}, \mathrm{I}_d)$$
$$L(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}) = \|\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}\|_2^2$$
$$\Theta = \mathbb{R}^d.$$

Theorem

$$R_n^*(\Theta) = \frac{d}{n}\,.$$

- Foundation of the least squares, maximum likelihood, ...

Kindergarten example

$$\mathbb{P}_{\boldsymbol{\theta}}(\,\cdot\,) = \mathsf{N}(\boldsymbol{\theta}, \mathrm{I}_d)$$
$$L(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}) = \|\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}\|_2^2$$
$$\Theta = \mathbb{R}^d.$$

Theorem

$$R_n^*(\Theta) = \frac{d}{n}\,.$$

- Foundation of the least squares, maximum likelihood, ...

A more sophisticated example

$$\mathbb{P}_{\boldsymbol{\theta}}(\,\cdot\,) = \mathsf{N}(\boldsymbol{\theta}, \mathrm{I}_d)$$
$$L(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}) = \|\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}\|_2^2$$
$$\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^d : \ \ \|\boldsymbol{\theta}\|_0 \le s_0\}\,.$$

Theorem (Donoho, Johnstone 1990s)

If $s_0/d \to 0$, then

$$R_n^*(\Theta) = \frac{2s_0}{n}\log(d/s_0)\cdot(1+o(1))\,.$$

- Key role in compressed sensing, sparse learning, ...

A more sophisticated example

$$\mathbb{P}_{\boldsymbol{\theta}}(\,\cdot\,) = \mathsf{N}(\boldsymbol{\theta}, \mathrm{I}_d)$$
$$L(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}) = \|\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}\|_2^2$$
$$\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^d : \;\; \|\boldsymbol{\theta}\|_0 \le s_0\}\,.$$

Theorem (Donoho, Johnstone 1990s)

If $s_0/d \to 0$, *then*

$$R_n^*(\Theta) = \frac{2s_0}{n}\log(d/s_0)\,\cdot\,(1 + o(1))\,.$$

- Key role in compressed sensing, sparse learning, ...

A more sophisticated example

$$\mathbb{P}_{\boldsymbol{\theta}}(\,\cdot\,) = \mathsf{N}(\boldsymbol{\theta}, \mathrm{I}_d)$$
$$L(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}) = \|\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}\|_2^2$$
$$\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^d : \ \|\boldsymbol{\theta}\|_0 \le s_0\}\,.$$

Theorem (Donoho, Johnstone 1990s)

If $s_0/d \to 0$*, then*

$$R_n^*(\Theta) = \frac{2s_0}{n} \log(d/s_0) \cdot (1 + o(1))\,.$$

- Key role in compressed sensing, sparse learning, ...

What is this talk about?

$$R_n^{\text{Poly}}(\Theta) = \inf_{\hat{\boldsymbol{\theta}}(\,\cdot\,)\in\text{Poly}} \sup_{\boldsymbol{\theta}_0\in\Theta} \mathbb{E}_{\boldsymbol{\theta}_0} L(\hat{\boldsymbol{\theta}}(x_1^n), \boldsymbol{\theta}_0)$$

Developments

- Often we expect $R_n^{\text{Poly}}(\Theta) \gneq R_n(\Theta)$
- Accurate predictions
- Convergence of fields (Statistics, CS Theory, Physics)
- New algorithms?

What is this talk about?

$$R_n^{\text{Poly}}(\Theta) = \inf_{\hat{\boldsymbol{\theta}}(\,\cdot\,)\in\text{Poly}} \sup_{\boldsymbol{\theta}_0\in\Theta} \mathbb{E}_{\boldsymbol{\theta}_0} L(\hat{\boldsymbol{\theta}}(x_1^n), \boldsymbol{\theta}_0)$$

Developments

- Often we expect $R_n^{\text{Poly}}(\Theta) \gneq R_n(\Theta)$
- Accurate predictions
- Convergence of fields (Statistics, CS Theory, Physics)
- New algorithms?

Outline

Local and message passing algorithms

Example #1: Sparse low-rank matrix

Unknowns:

$$\Theta(s_0, d) = \left\{ \boldsymbol{\theta} \in \{0,1\}^d : \|\boldsymbol{\theta}\|_0 = s_0 \right\}.$$

Loss:

$$L(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \frac{1}{d} \, \mathsf{Hamming}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})$$

Example #1: Sparse low-rank matrix

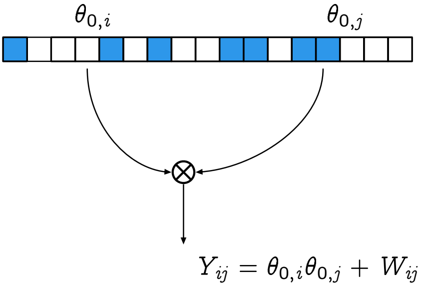
Data: $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$

$$\boldsymbol{x}_\ell = (i_\ell, j_\ell, Y_{i_\ell, j_\ell}) \in [d] \times [d] \times \mathbb{R}\,,$$

$$i_\ell, j_\ell \sim_{iid} \mathsf{Unif}([d])\,,$$

$$Y_{i_\ell, j_\ell}\big|_{i_\ell, j_\ell} \sim \mathsf{N}(\ \theta_{0,i_\ell}\theta_{0,j_\ell}\ ,\ \sigma^2\ )$$

Pictorially



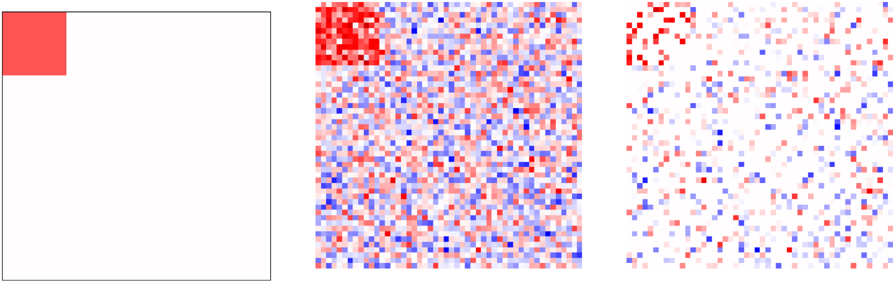
Example #1: A different formulation

Data $\boldsymbol{Y} \in \mathbb{R}^{n\times n}$

$$\boldsymbol{Y} = \mathcal{P}_E(\boldsymbol{\theta}_0\boldsymbol{\theta}_0^{\mathsf{T}} + \boldsymbol{W})$$

- $E \subseteq \binom{[d]}{2}$ uniformly random s.t. $|E| = n$
- $\mathcal{P}_E : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$: projector that zeroes entries not in E
- $(W_{ij})_{i<j} \sim_{iid} \mathsf{N}(0, \sigma^2)$, $\boldsymbol{W} = \boldsymbol{W}^{\mathsf{T}}$

$\boldsymbol{\theta}_0\boldsymbol{\theta}_0^{\mathsf{T}}$ $\quad$ $\boldsymbol{\theta}_0\boldsymbol{\theta}_0^{\mathsf{T}} + \boldsymbol{W}$ $\quad$ $\boldsymbol{Y} = \mathcal{P}_E\big(\boldsymbol{\theta}_0\boldsymbol{\theta}_0^{\mathsf{T}} + \boldsymbol{W}\big)$



Example #1: Yet another formulation

- $G = (V, E) \sim \mathcal{G}(d, n)$
  (uniform random with d vertices, n edges)

- For each $(i, j) \in E$

$$Y_{ij} = \theta_i \theta_j + W_{ij}\,, \qquad W_{ij} \sim \mathsf{N}(0, \sigma^2)$$

Problem parameters

$$\delta = \frac{n}{d} \qquad \text{(half) average graph degree}$$

$$\varepsilon = \frac{s_0}{d} \qquad \text{sparsity}$$

Asymptotics

- Dense graph: $n \asymp d^2$, $s_0 \asymp \sqrt{d}$ (∼ Planted clique problem)
- Sparse graph: $n \asymp d$, $s_0 \asymp d$ (δ, ε fixed)

Problem parameters

$$\delta = \frac{n}{d} \qquad \text{(half) average graph degree}$$

$$\varepsilon = \frac{s_0}{d} \qquad \text{sparsity}$$

Asymptotics

- Dense graph: $n \asymp d^2$, $s_0 \asymp \sqrt{d}$ (∼ Planted clique problem)
- **Sparse graph**: $n \asymp d$, $s_0 \asymp d$ (δ, ε fixed)

$$R^{\text{Poly}}(\delta, \varepsilon; d) \equiv R^{\text{Poly}}_{n=d\delta}\big(\Theta(s_0 = d\varepsilon, d)\big)$$

$$\lim_{d\to\infty} R^{\text{Poly}}(\delta, \varepsilon; d) = ?$$

We do not know, but ...

$$R^{\mathrm{Poly}}(\delta, \varepsilon; d) \equiv R^{\mathrm{Poly}}_{n=d\delta}\big(\Theta(s_0 = d\varepsilon, d)\big)$$

$$\lim_{d\to\infty} R^{\mathrm{Poly}}(\delta, \varepsilon; d) = ?$$

We do not know, but ...

First simplifications

- Worst case prior

$$\boldsymbol{\theta} \sim \mathsf{Unif}\big(\Theta(s_0, d)\big)$$

- Roughly ($\varepsilon = s_0/d$)

$$(\theta_i)_{i \le d} \sim_{iid} \mathsf{Bern}(\varepsilon)$$

First simplifications

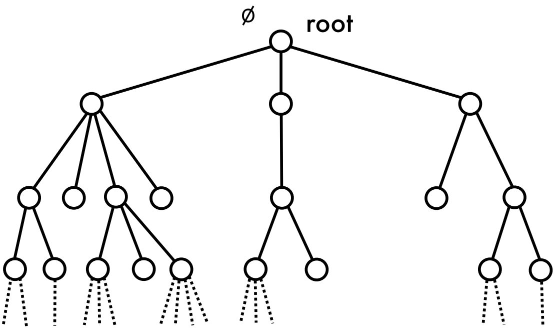
- Worst case prior

$$\boldsymbol{\theta} \sim \mathsf{Unif}\big(\Theta(s_0, d)\big)$$

- Roughly ($\varepsilon = s_0/d$)

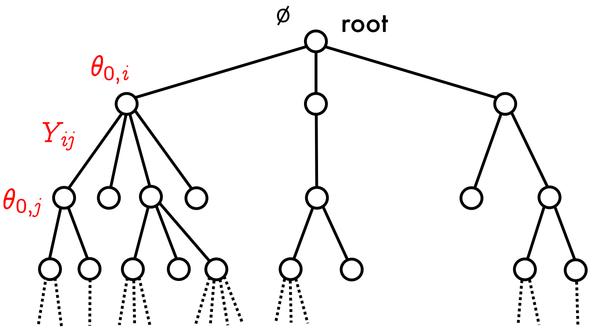
$$(\theta_i)_{i \le d} \sim_{iid} \mathsf{Bern}(\varepsilon)$$

Local weak limit: $d \to \infty$



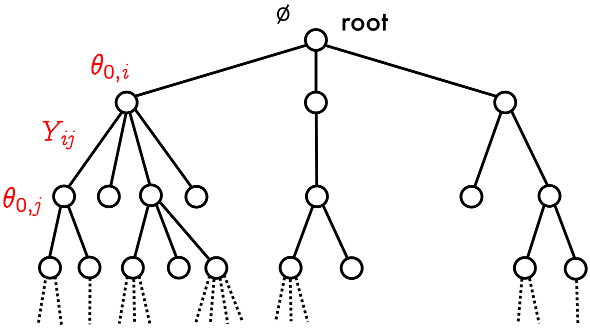
$$G \overset{lwc}{\Rightarrow} \mathsf{GW}(\mathsf{Pois}(2\delta))$$

Local weak limit: $d \to \infty$



$$(G, \boldsymbol{\theta}, \boldsymbol{Y}) \overset{lwc}{\Rightarrow} (\mathsf{GW}(\mathsf{Pois}(2\delta)), \boldsymbol{\theta}, \boldsymbol{Y})$$

First guess



$$\lim_{d\to\infty} R^{\rm Poly}(\delta,\varepsilon,d) \stackrel{?}{=} \inf_{\hat{\theta}} \mathbb{P}_{\rm Tree}\big(\hat{\theta}_{\emptyset}(\boldsymbol{Y}) \neq \theta_{0,\emptyset}\big)$$

It gets interesting

$$\lim_{d\to\infty} R^{\rm Poly}(\delta,\varepsilon,d) \stackrel{?}{=} \inf_{\hat{\boldsymbol{\theta}}} \mathbb{P}_{\rm Tree}\big(\hat{\theta}_{\emptyset}(\boldsymbol{Y}) \neq \theta_{0,\emptyset}\big)$$

How do you define the r.h.s.?

A natural idea:

$$\boldsymbol{Y}^0_\ell = \big(Y_{ij}:\ d(\emptyset,i)\le\ell\big)\,,$$

$$\hat{\theta}_{\emptyset}(\boldsymbol{Y}^0_\ell) = \arg\max_{\sigma\in\{0,1\}} \mathbb{P}\big(\theta_{\emptyset}=\sigma\big|\boldsymbol{Y}^0_\ell\big)\,.$$

It gets interesting

$$\lim_{d\to\infty} R^{\text{Poly}}(\delta,\varepsilon,d) \overset{?}{=} \inf_{\hat{\boldsymbol{\theta}}} \mathbb{P}_{\text{Tree}}\big(\hat{\theta}_{\emptyset}(\boldsymbol{Y}) \neq \theta_{0,\emptyset}\big)$$

How do you define the r.h.s.?

A natural idea:

$$\boldsymbol{Y}_{\ell}^{0} = \big(Y_{ij}:\ d(\emptyset, i) \leq \ell\big)\,,$$

$$\hat{\theta}_{\emptyset}(\boldsymbol{Y}_{\ell}^{0}) = \arg\max_{\sigma\in\{0,1\}} \mathbb{P}\big(\theta_{\emptyset} = \sigma \big| \boldsymbol{Y}_{\ell}^{0}\big)\,.$$

Risk of local algorithms

$$R^{\mathrm{loc}}(\delta, \varepsilon) = \lim_{\ell \to \infty} \mathbb{P}_{\mathrm{Tree}}\big(\hat{\theta}_{\varnothing}(\boldsymbol{Y}_{\ell}^{0}) \neq \theta_{0,\varnothing}\big)$$

Remark

Belief propagation achieves $R^{\mathrm{loc}}(\delta, \varepsilon)$

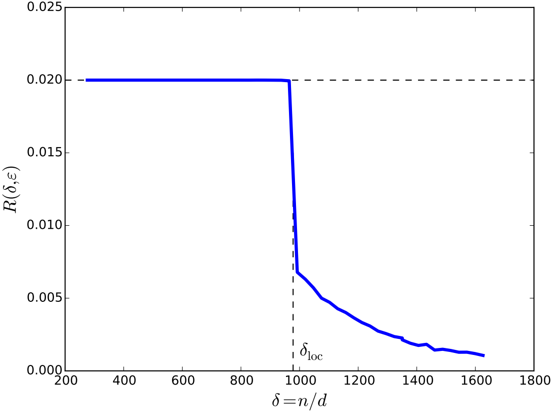
Risk of local algorithms

$$R^{\rm loc}(\delta,\varepsilon) = \lim_{\ell\to\infty} \mathbb{P}_{\rm Tree}\big(\hat{\theta}_{\emptyset}(\boldsymbol{Y}^0_{\ell}) \neq \theta_{0,\emptyset}\big)$$

Remark

Belief propagation achieves $R^{\rm loc}(\delta,\varepsilon)$

Can be computed $(\varepsilon = 0.02,\ \sigma = 1.5)$



$$R^{\rm loc}(\delta,\varepsilon) = \lim_{\ell\to\infty} \mathbb{P}_{\tt Tree}\big(\hat{\theta}_{\emptyset}(\boldsymbol{Y}_\ell^0) \neq \theta_{0,\emptyset}\big)$$

Phase transition for local algorithms

Theorem (∼ Deshpande, Montanari, 2015)

As $\varepsilon \to 0$ (with $\delta \to \infty$, σ fixed)

$$\frac{1}{\varepsilon} R^{\rm loc}(\delta,\varepsilon) \to \begin{cases} 1 & \text{if } \delta \le (1-o_\varepsilon(1)) \cdot \frac{\sigma^2}{2e\varepsilon^2}\,, \\ 0 & \text{if } \delta \ge (1+o_\varepsilon(1)) \cdot \frac{\sigma^2}{2e\varepsilon^2}\,, \end{cases}$$

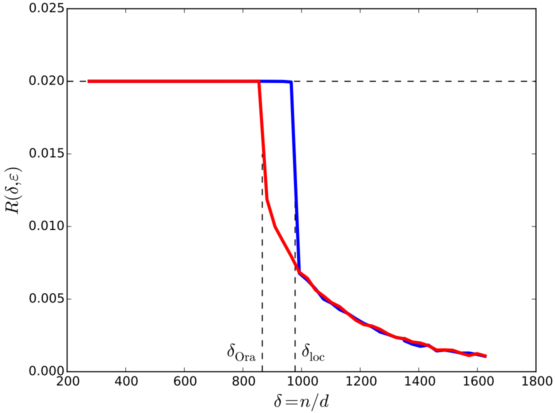
A different definition of the limit

$$
\boldsymbol{Y}_\ell^+ = \{(Y_{ij} : \; d(\emptyset, i) \le \ell); \;\; (\theta_i : \; d(\emptyset, i > \ell)\}
$$

$$
\hat{\theta}_\emptyset(\boldsymbol{Y}_\ell^+) = \arg\max_{\sigma \in \{0,1\}} \mathbb{P}(\theta_\emptyset = \sigma \big| \boldsymbol{Y}_\ell^+)
$$

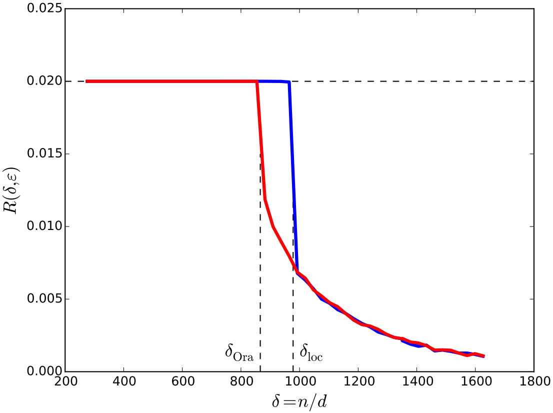
$$
R^{\rm Ora}(\delta, \varepsilon) = \lim_{\ell \to \infty} \mathbb{P}_{\rm Tree}(\hat{\theta}_\emptyset(\boldsymbol{Y}_\ell^0) \neq \theta_{0,\emptyset})
$$

It also can be computed



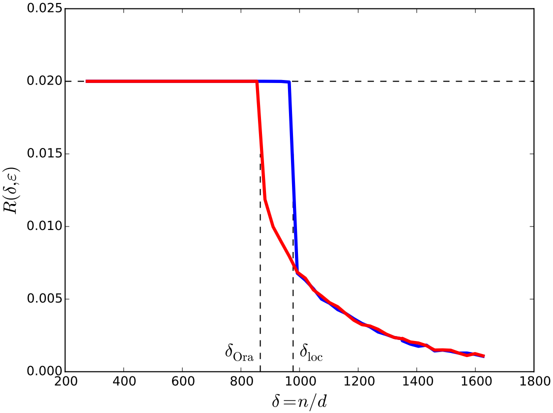
$$R^{\rm Ora}(\delta,\varepsilon) \le R^{*}(\delta,\varepsilon) \le R^{\rm Poly}(\delta,\varepsilon) \le R^{\rm loc}(\delta,\varepsilon)$$

It also can be computed



$$R^{\mathrm{Ora}}(\delta,\varepsilon) \le R^{*}(\delta,\varepsilon) \le R^{\mathrm{Poly}}(\delta,\varepsilon) \le R^{\mathrm{loc}}(\delta,\varepsilon)$$

It also can be computed



$$R^{\rm Ora}(\delta,\varepsilon) \le R^*(\delta,\varepsilon) \le R^{\rm Poly}(\delta,\varepsilon) \le R^{\rm loc}(\delta,\varepsilon)$$

Minimax risk

Theorem (∼ Montanari, 2015)

As $\varepsilon \to 0$ *(with* $\delta \to \infty$*,* σ *fixed)*

$$\frac{1}{\varepsilon} R^*(\delta, \varepsilon) \to \begin{cases} 1 & \text{if } \delta \le (1 - o_\varepsilon(1)) \cdot \frac{\sigma^2}{2\varepsilon} \log(1/\varepsilon)\,, \\ 0 & \text{if } \delta \ge (1 + o_\varepsilon(1)) \cdot \frac{\sigma^2}{2\varepsilon} \log(1/\varepsilon)\,, \end{cases}$$

- Sharp threshold (dense): Lelarge, Miolane 2017; Barbier et al. 2017
- Do not know of any polytime algorithm working for

$$1.01 \cdot \frac{\sigma^2}{2\varepsilon} \log(1/\varepsilon) < \delta < 0.99 \cdot \frac{\sigma^2}{2e\varepsilon^2}\,.$$

Minimax risk

Theorem ($\sim$ Montanari, 2015)

As $\varepsilon \to 0$ *(with* $\delta \to \infty$, σ *fixed)*

$$\frac{1}{\varepsilon} R^*(\delta, \varepsilon) \to \begin{cases} 1 & \text{if } \delta \le (1 - o_\varepsilon(1)) \cdot \frac{\sigma^2}{2\varepsilon} \log(1/\varepsilon)\,, \\ 0 & \text{if } \delta \ge (1 + o_\varepsilon(1)) \cdot \frac{\sigma^2}{2\varepsilon} \log(1/\varepsilon)\,, \end{cases}$$

- Sharp threshold (dense): Lelarge, Miolane 2017; Barbier et al. 2017
- Do not know of any polytime algorithm working for

$$1.01 \cdot \frac{\sigma^2}{2\varepsilon} \log(1/\varepsilon) < \delta < 0.99 \cdot \frac{\sigma^2}{2e\varepsilon^2}\,.$$

Minimax risk

Theorem (∼ Montanari, 2015)

As $\varepsilon \to 0$ *(with* $\delta \to \infty$*,* σ *fixed)*

$$\frac{1}{\varepsilon} R^*(\delta, \varepsilon) \to \begin{cases} 1 & \text{if } \delta \le (1 - o_\varepsilon(1)) \cdot \frac{\sigma^2}{2\varepsilon} \log(1/\varepsilon)\,, \\ 0 & \text{if } \delta \ge (1 + o_\varepsilon(1)) \cdot \frac{\sigma^2}{2\varepsilon} \log(1/\varepsilon)\,, \end{cases}$$

- Sharp threshold (dense): Lelarge, Miolane 2017; Barbier et al. 2017
- Do not know of any polytime algorithm working for

$$1.01 \cdot \frac{\sigma^2}{2\varepsilon} \log(1/\varepsilon) < \delta < 0.99 \cdot \frac{\sigma^2}{2e\varepsilon^2}\,.$$

Open problems

- Can we beat $R^{\text{loc}}(\delta, \varepsilon)$ by Gibbs sampling?
- Can we beat $R^{\text{loc}}(\delta, \varepsilon)$ by convex optimization?
- Is $R^{\text{Poly}}(\delta, \varepsilon) = R^{\text{loc}}(\delta, \varepsilon)$?

Open problems

- Can we beat $R^{\text{loc}}(\delta, \varepsilon)$ by Gibbs sampling?
- Can we beat $R^{\text{loc}}(\delta, \varepsilon)$ by convex optimization?
- Is $R^{\text{Poly}}(\delta, \varepsilon) = R^{\text{loc}}(\delta, \varepsilon)$?

Open problems

- Can we beat $R^{\rm loc}(\delta, \varepsilon)$ by Gibbs sampling?

- Can we beat $R^{\rm loc}(\delta, \varepsilon)$ by convex optimization?

- Is $R^{\rm Poly}(\delta, \varepsilon) = R^{\rm loc}(\delta, \varepsilon)$?

Ubiquitous

Sparse principal component analysis

- Berthet, Rigollet, 2013
- Deshpande, Montanari, 2014
- Barbier et al 2016; Miolane 2017

Hidden clique problem (2 asymmetric communities)

- Jerrum, 1992
- Feige, Krauthgamer, 200
- Deshpande, Montanari, 2015; Montanari 2015

Community detection ($k \geq 5$ symmetric communities)

- Decelle, Krzakala, Moore, Zdeborova 2011
- Bordenave, Lelarge, Massoulie 2015
- Abbe, Sandon 2015

Tensor PCA

- See below

Landscapes

Empirical risk minimization/M-estimation

Class of models ($\Theta \subseteq \mathbb{R}^d$)

$$\mathcal{C}_\Theta \equiv \{ \mathbb{P}_{\boldsymbol{\theta}} : \quad \boldsymbol{\theta} \in \Theta \}$$

Data

$$\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n \sim_{iid} \mathbb{P}_{\boldsymbol{\theta}_0}(\,\cdot\,)$$

$$\text{minimize} \quad \widehat{\mathcal{L}}_n(\theta) \equiv \frac{1}{n} \sum_{i=1}^{n} \ell(\theta; x_i)$$

Empirical risk minimization/M-estimation

Class of models ($\Theta \subseteq \mathbb{R}^d$)

$$\mathcal{C}_\Theta \equiv \{ \mathbb{P}_{\boldsymbol{\theta}} : \quad \boldsymbol{\theta} \in \Theta \}$$

Data

$$\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n \sim_{iid} \mathbb{P}_{\boldsymbol{\theta}_0}(\,\cdot\,)$$

$$\text{minimize} \quad \widehat{\mathcal{L}}_n(\boldsymbol{\theta}) \equiv \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{\theta}; \boldsymbol{x}_i)$$

Empirical risk minimization/M-estimation

Rationale

$$\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \mathcal{L}(\boldsymbol{\theta})\,, \qquad \mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}\widehat{\mathcal{L}}_n(\boldsymbol{\theta})$$

- What can we say generically?

- How does complexity show up?

- What can we say generically?

- How does complexity show up?

Uniform convergence

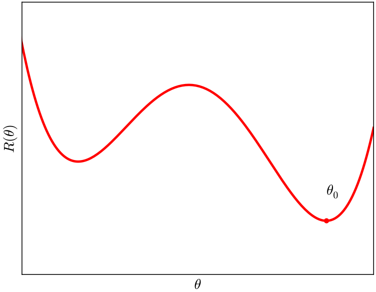
Theorem (Vapnik, Chervonenkis, 1968; ...)

Under conditions [omitted], with high probability

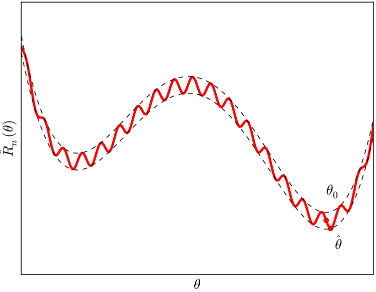
$$\sup_{\boldsymbol{\theta}\in\Theta}\left|\widehat{\mathcal{L}}_n(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta})\right| \le C\sqrt{\frac{d_*}{n}}\,.$$

($d_* =$ VC dimension; ...).

Uniform convergence



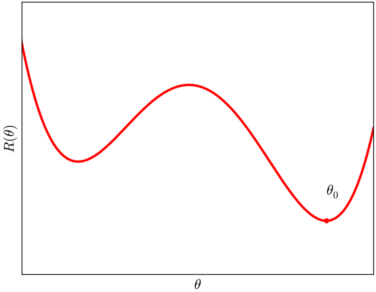
Population risk



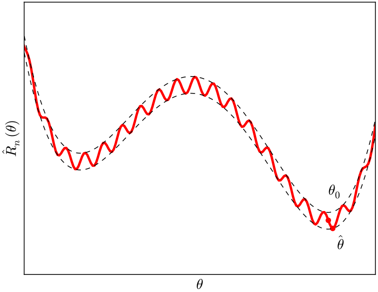
Empirical risk

Will optimization algorithms get stuck in local minima?
Landscape analysis

Uniform convergence



Population risk

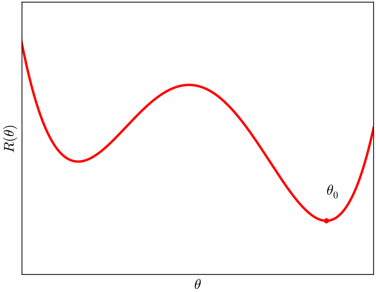


Empirical risk

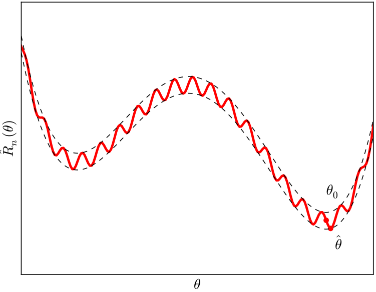
Will optimization algorithms get stuck in local minima?

Landscape analysis

Uniform convergence



Population risk



Empirical risk

Will optimization algorithms get stuck in local minima?
Landscape analysis

Assumptions

$\boldsymbol{\theta} \in \mathsf{B}^p(r) =$ Ball of radius r in $\mathbb{R}^p$

Data: $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \dots, \boldsymbol{Z}_n$ iid

A1 $\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; Z)$ is τ^2-sub-Gaussian

A2 For any $\boldsymbol{\lambda} \in \mathsf{B}^p(1)$, $\mathcal{Z}_{\boldsymbol{\lambda}} \equiv \langle \boldsymbol{\lambda}, \nabla^2 \ell(\boldsymbol{\theta}; Z) \boldsymbol{\lambda} \rangle$ is τ^2-sub-Exponential.

A3 The Hessian of the population risk at $\mathbf{0}$ is bounded by a polynomial

$$\|\nabla^2 \mathcal{L}(\mathbf{0})\|_{\rm op} \le \tau^2 d^C \,.$$

A4 The Hessian of the loss is Lipschitz continuous with integrable constant

$$\mathbb{E}\{\|\nabla^2 \ell(\,\cdot\,; Z)\|_{\rm Lip}\} \le \tau^3\, d^C \,.$$

'Under mild assumptions'

Assumptions

$\boldsymbol{\theta} \in \mathsf{B}^p(r) =$ Ball of radius r in $\mathbb{R}^p$
Data: $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \dots, \boldsymbol{Z}_n$ iid

- A1 $\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \boldsymbol{Z})$ is τ^2-sub-Gaussian
- A2 For any $\boldsymbol{\lambda} \in \mathsf{B}^p(1)$, $\mathcal{Z}_{\boldsymbol{\lambda}} \equiv \langle \boldsymbol{\lambda}, \nabla^2 \ell(\boldsymbol{\theta}; \boldsymbol{Z}) \boldsymbol{\lambda} \rangle$ is τ^2-sub-Exponential.
- A3 The Hessian of the population risk at $\mathbf{0}$ is bounded by a polynomial

$$\|\nabla^2 \mathcal{L}(\mathbf{0})\|_{\rm op} \le \tau^2 d^C \, .$$

- A4 The Hessian of the loss is Lipschitz continuous with integrable constant

$$\mathbb{E}\{\|\nabla^2 \ell(\,\cdot\,; \boldsymbol{Z})\|_{\rm Lip}\} \le \tau^3 \, d^C \, .$$

'Under mild assumptions'

Assumptions

$\boldsymbol{\theta} \in \mathsf{B}^p(r) =$ Ball of radius r in $\mathbb{R}^p$
Data: $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \dots, \boldsymbol{Z}_n$ iid

- A1 $\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \boldsymbol{Z})$ is τ^2-sub-Gaussian
- A2 For any $\boldsymbol{\lambda} \in \mathsf{B}^p(1)$, $\mathcal{Z}_{\boldsymbol{\lambda}} \equiv \langle \boldsymbol{\lambda}, \nabla^2 \ell(\boldsymbol{\theta}; \boldsymbol{Z}) \boldsymbol{\lambda} \rangle$ is τ^2-sub-Exponential.
- A3 The Hessian of the population risk at $\mathbf{0}$ is bounded by a polynomial

$$\|\nabla^2 \mathcal{L}(\mathbf{0})\|_{\mathrm{op}} \le \tau^2 d^C \,.$$

- A4 The Hessian of the loss is Lipschitz continuous with integrable constant

$$\mathbb{E}\{\|\nabla^2 \ell(\,\cdot\,; \boldsymbol{Z})\|_{\mathrm{Lip}}\} \le \tau^3\, d^C \,.$$

'Under mild assumptions'

Assumptions

$\boldsymbol{\theta} \in \mathsf{B}^p(r) =$ Ball of radius r in $\mathbb{R}^p$
Data: $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \dots, \boldsymbol{Z}_n$ iid

A1 $\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \boldsymbol{Z})$ is τ^2-sub-Gaussian

A2 For any $\boldsymbol{\lambda} \in \mathsf{B}^p(1)$, $\mathcal{Z}_{\boldsymbol{\lambda}} \equiv \langle \boldsymbol{\lambda}, \nabla^2 \ell(\boldsymbol{\theta}; Z)\boldsymbol{\lambda} \rangle$ is τ^2-sub-Exponential.

A3 The Hessian of the population risk at $\mathbf{0}$ is bounded by a polynomial

$$\|\nabla^2 \mathcal{L}(\mathbf{0})\|_{\rm op} \le \tau^2 d^C\,.$$

A4 The Hessian of the loss is Lipschitz continuous with integrable constant

$$\mathbb{E}\{\|\nabla^2 \ell(\,\cdot\,; Z)\|_{\rm Lip}\} \le \tau^3\, d^C\,.$$

'Under mild assumptions'

Assumptions

$\boldsymbol{\theta} \in \mathsf{B}^p(r) =$ Ball of radius r in $\mathbb{R}^p$

Data: $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \dots, \boldsymbol{Z}_n$ iid

- A1 $\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \boldsymbol{Z})$ is τ^2-sub-Gaussian
- A2 For any $\boldsymbol{\lambda} \in \mathsf{B}^p(1)$, $\mathcal{Z}_{\boldsymbol{\lambda}} \equiv \langle \boldsymbol{\lambda}, \nabla^2 \ell(\boldsymbol{\theta}; \boldsymbol{Z}) \boldsymbol{\lambda} \rangle$ is τ^2-sub-Exponential.
- A3 The Hessian of the population risk at $\mathbf{0}$ is bounded by a polynomial

$$\|\nabla^2 \mathcal{L}(\mathbf{0})\|_{\text{op}} \le \tau^2 d^C .$$

- A4 The Hessian of the loss is Lipschitz continuous with integrable constant

$$\mathbb{E}\{\|\nabla^2 \ell(\,\cdot\,; \boldsymbol{Z})\|_{\text{Lip}}\} \le \tau^3 \, d^C .$$

'Under mild assumptions'

Assumptions

$\boldsymbol{\theta} \in \mathsf{B}^p(r) =$ Ball of radius r in $\mathbb{R}^p$
Data: $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \dots, \boldsymbol{Z}_n$ iid

- A1 $\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}; \boldsymbol{Z})$ is τ^2-sub-Gaussian
- A2 For any $\boldsymbol{\lambda} \in \mathsf{B}^p(1)$, $\mathcal{Z}_{\boldsymbol{\lambda}} \equiv \langle \boldsymbol{\lambda}, \nabla^2\ell(\boldsymbol{\theta}; \boldsymbol{Z})\boldsymbol{\lambda}\rangle$ is τ^2-sub-Exponential.
- A3 The Hessian of the population risk at $\mathbf{0}$ is bounded by a polynomial

$$\|\nabla^2\mathcal{L}(\mathbf{0})\|_{\rm op} \le \tau^2 d^C\,.$$

- A4 The Hessian of the loss is Lipschitz continuous with integrable constant

$$\mathbb{E}\{\|\nabla^2\ell(\,\cdot\,; \boldsymbol{Z})\|_{\rm Lip}\} \le \tau^3\, d^C\,.$$

'Under mild assumptions'

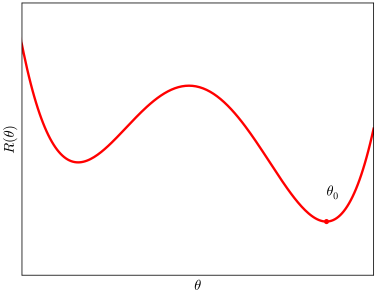
Lemma

Under assumptions A1, A2, A3, A4, *if* $n \ge Cp \log p$, *then with probability at least* $1-\delta$, *the following hold:*

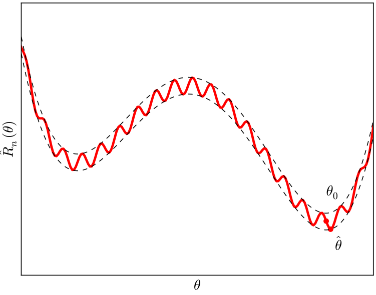
$$\sup_{\boldsymbol{\theta} \in \mathsf{B}^p(r)} \left\| \nabla \widehat{\mathcal{L}}_n(\boldsymbol{\theta}) - \nabla \mathcal{L}(\boldsymbol{\theta}) \right\|_2 \le \tau \sqrt{\frac{Cd \log n}{n}},$$

$$\sup_{\boldsymbol{\theta} \in \mathsf{B}^p(r)} \left\| \nabla^2 \widehat{\mathcal{L}}_n(\boldsymbol{\theta}) - \nabla^2 \mathcal{L}(\boldsymbol{\theta}) \right\|_{\mathrm{op}} \le \tau^2 \sqrt{\frac{Cd \log n}{n}}.$$

This cannot happen!

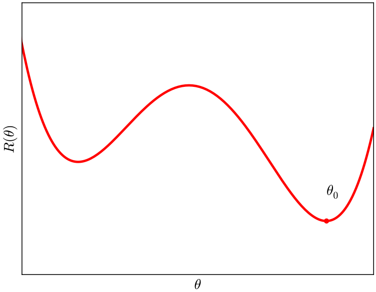


Population risk

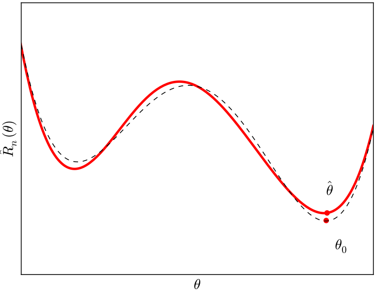


Empirical risk

This can happen!



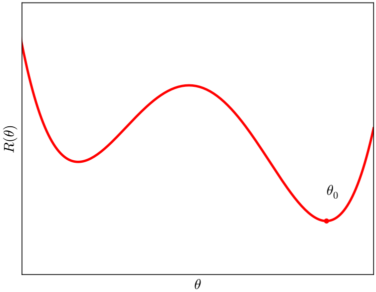
Population risk



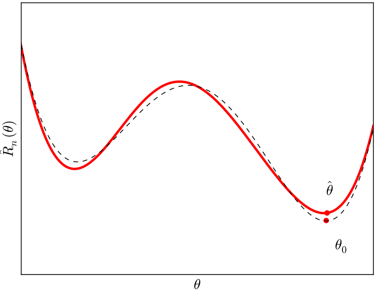
Empirical risk

Nice population risk $\Rightarrow$ Nice empirical risk

This can happen!



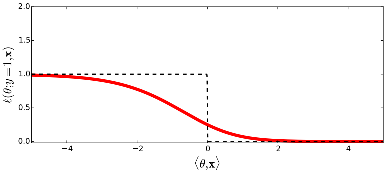
Population risk



Empirical risk

Nice population risk $\Rightarrow$ Nice empirical risk

Example: Binary classification



$$\boldsymbol{z}_i = (y_i, \boldsymbol{x}_i),\ y_i \in \{0,1\},\ \boldsymbol{x}_i \in \mathbb{R}^d$$

$$\mathbb{P}(y_i = 1 | \boldsymbol{x}_i) = \sigma(\langle \boldsymbol{\theta}_0, \boldsymbol{x}_i \rangle)$$

$$\widehat{\mathcal{L}}_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \sigma(\langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle)\right)^2 .$$

- Rosenblatt 1958 (Perceptron); ... many extensions
- More robust than logistic regression

Sample application: Binary classification

Theorem (Mei, Bai, Montanari 2017)

Assume $\boldsymbol{X}_i$ *to be centered, sub-Gaussian, with* $\mathbb{E}\{\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}}\} \succeq \delta \mathrm{I}_d$.
For nice[a] *functions* σ, *whp:*

1. *The population risk has a unique critical point* $\hat{\boldsymbol{\theta}}_n$.
2. *Gradient descent converges exponentially fast to* $\hat{\boldsymbol{\theta}}_n$.
3. *The estimation error is* $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|_2 \le C\sqrt{(d \log n)/n}$.

[a] $\sigma'(x) > 0$, $\|\sigma'\|_\infty, \|\sigma''\|_\infty, \|\sigma'''\|_\infty \le C$.

Similar results for robust regression, one-bit compressed sensing, . . .
[see paper]

Sample application: Binary classification

Theorem (Mei, Bai, Montanari 2017)

Assume $\boldsymbol{X}_i$ to be centered, sub-Gaussian, with $\mathbb{E}\{\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}}\} \succeq \delta \mathrm{I}_d$. For nice[a] functions σ, whp:

1. *The population risk has a unique critical point $\hat{\boldsymbol{\theta}}_n$.*
2. *Gradient descent converges exponentially fast to $\hat{\theta}_n$.*
3. *The estimation error is $\|\hat{\theta}_n - \theta_0\|_2 \le C\sqrt{(d \log n)/n}$.*

[a] $\sigma'(x) > 0$, $\|\sigma'\|_\infty, \|\sigma''\|_\infty, \|\sigma'''\|_\infty \le C$.

Similar results for robust regression, one-bit compressed sensing, ...

[see paper]

Sample application: Binary classification

Theorem (Mei, Bai, Montanari 2017)

Assume $\boldsymbol{X}_i$ to be centered, sub-Gaussian, with $\mathbb{E}\{\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}}\} \succeq \delta \mathrm{I}_d$. For nice[a] functions σ, whp:

1. *The population risk has a unique critical point $\hat{\boldsymbol{\theta}}_n$.*
2. *Gradient descent converges exponentially fast to $\hat{\boldsymbol{\theta}}_n$.*
3. *The estimation error is $\|\hat{\theta}_n - \theta_0\|_2 \le C\sqrt{(d \log n)/n}$.*

[a] $\sigma'(x) > 0$, $\|\sigma'\|_\infty, \|\sigma''\|_\infty, \|\sigma'''\|_\infty \le C$.

Similar results for robust regression, one-bit compressed sensing, ...

[see paper]

Sample application: Binary classification

Theorem (Mei, Bai, Montanari 2017)

Assume $\boldsymbol{X}_i$ *to be centered, sub-Gaussian, with* $\mathbb{E}\{\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}}\} \succeq \delta \mathrm{I}_d$.
For nice[a] *functions* σ, *whp:*

1. *The population risk has a unique critical point* $\hat{\boldsymbol{\theta}}_n$.
2. *Gradient descent converges exponentially fast to* $\hat{\boldsymbol{\theta}}_n$.
3. *The estimation error is* $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|_2 \le C\sqrt{(d \log n)/n}$.

[a] $\sigma'(x) > 0$, $\|\sigma'\|_\infty, \|\sigma''\|_\infty, \|\sigma'''\|_\infty \le C$.

Similar results for robust regression, one-bit compressed sensing, ...
[see paper]

Sample application: Binary classification

Theorem (Mei, Bai, Montanari 2017)

Assume $\boldsymbol{X}_i$ to be centered, sub-Gaussian, with $\mathbb{E}\{\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}}\} \succeq \delta \mathrm{I}_d$. For nice[a] functions σ, whp:

1. *The population risk has a unique critical point $\hat{\boldsymbol{\theta}}_n$.*
2. *Gradient descent converges exponentially fast to $\hat{\boldsymbol{\theta}}_n$.*
3. *The estimation error is $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|_2 \le C\sqrt{(d \log n)/n}$.*

[a] $\sigma'(x) > 0$, $\|\sigma'\|_\infty, \|\sigma''\|_\infty, \|\sigma'''\|_\infty \le C$.

Similar results for robust regression, one-bit compressed sensing, ...

[see paper]

Non-convex literature

Convergence to 'statistical neighborhood'

- Loh, Wainwright, 2012
- Loh Wainwright, 2013
- Yang, Wang, Liu, Eldar, Zhang, 2015
- ...

Smart initialization

- Keshavan, Montanari, Oh, 2009
- Chen, Candés, 2015
- Anandkumar, Ge, Jenzamin, 2015
- ...

Unique local minimum

- Loh, 2015 [High-dim. regression, $n \gtrsim s_0^2$]
- Sun, Qu, Wright, 2016 [Phase retrieval]
- Ge, Lee, Ma, 2016 [Matrix completion]

Non-convex literature

Convergence to 'statistical neighborhood'

- Loh, Wainwright, 2012
- Loh Wainwright, 2013
- Yang, Wang, Liu, Eldar, Zhang, 2015
- ...

Smart initialization

- Keshavan, Montanari, Oh, 2009
- Chen, Candés, 2015
- Anandkumar, Ge, Jenzamin, 2015
- ...

Unique local minimum

- Loh, 2015 [High-dim. regression, $n \gtrsim s_0^2$]
- Sun, Qu, Wright, 2016 [Phase retrieval]
- Ge, Lee, Ma, 2016 [Matrix completion]

- What can we say generically?

- How does complexity show up?

Intuition

Population risk very flat $\Leftrightarrow$ Many local minima

[Close to where local algorithms fail?]

Simplest example: Spiked tensor model

- Unknown parameter $\boldsymbol{\theta}_0 \in \mathbb{R}^n$, $\|\boldsymbol{\theta}_0\|_2 = 1$
- **Data**[1]

$$\boldsymbol{Y} = \lambda \boldsymbol{\theta}_0^{\otimes k} + \boldsymbol{W}$$

- $\boldsymbol{W}$ = symmetric Gaussian noise tensor.
  $(W_{i_1,\dots,i_k})_{i_1<\dots<i_k} \sim_{iid} \mathsf{N}(0, 1/n)$

[1] Equivalently, $Y_{i_1,\dots,i_k} = \lambda\theta_{0,i_1}\cdots\theta_{0,i_k} + W_{i_1,\dots,i_k}$.

Spiked tensor model: What do we know?

$$Y = \lambda \theta_0^{\otimes k} + W$$

Theorem (Montanari, Richard, 2014; Hopkins, Shi, Steurer, 2015)

For any $\varepsilon > 0$, there exist constants $\lambda_{\rm IT}$, $\lambda_{\rm ML}(\varepsilon)$, $C(\varepsilon)$, such that:

- *If $\lambda > \lambda_{\rm ML}(\varepsilon)$, then $\mathbb{E}\{|\langle \hat{\theta}^{\rm ML}, \theta_0 \rangle|\} \geq 1 - \varepsilon$.*
- *No estimator can achieve $\mathbb{E}\{|\langle \hat{\theta}, \theta_0 \rangle|\} \geq \varepsilon$ unless $\lambda > \lambda_{\rm IT}$.*
- *There exists poly-time estimator achieving $\mathbb{E}\{|\langle \hat{\theta}^{\rm Poly}, \theta_0 \rangle|\} \geq 1 - \varepsilon$, provided $\lambda \geq C(\varepsilon) n^{(k-2)/4}$.*

No efficient estimator is known for $1 \ll \lambda \ll n^{(k-2)/4}$!

Spiked tensor model: What do we know?

$$\boldsymbol{Y} = \lambda \boldsymbol{\theta}_0^{\otimes k} + \boldsymbol{W}$$

Theorem (Montanari, Richard, 2014; Hopkins, Shi, Steurer, 2015)

For any $\varepsilon > 0$, there exist constants $\lambda_{\rm IT}$, $\lambda_{\rm ML}(\varepsilon)$, $C(\varepsilon)$, such that:

- *If $\lambda > \lambda_{\rm ML}(\varepsilon)$, then $\mathbb{E}\{|\langle \hat{\boldsymbol{\theta}}^{\rm ML}, \boldsymbol{\theta}_0 \rangle|\} \ge 1 - \varepsilon$.*
- *No estimator can achieve $\mathbb{E}\{|\langle \hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0 \rangle|\} \ge \varepsilon$ unless $\lambda > \lambda_{\rm IT}$.*
- *There exists poly-time estimator achieving $\mathbb{E}\{|\langle \hat{\boldsymbol{\theta}}^{\rm Poly}, \boldsymbol{\theta}_0 \rangle|\} \ge 1 - \varepsilon$, provided $\lambda \ge C(\varepsilon) n^{(k-2)/4}$.*

No efficient estimator is known for $1 \ll \lambda \ll n^{(k-2)/4}$!

Spiked tensor model: What do we know?

$$\boldsymbol{Y} = \lambda \boldsymbol{\theta}_0^{\otimes k} + \boldsymbol{W}$$

Theorem (Montanari, Richard, 2014; Hopkins, Shi, Steurer, 2015)

For any $\varepsilon > 0$, there exist constants $\lambda_{\rm IT}$, $\lambda_{\rm ML}(\varepsilon)$, $C(\varepsilon)$, such that:

- *If $\lambda > \lambda_{\rm ML}(\varepsilon)$, then $\mathbb{E}\{|\langle \hat{\boldsymbol{\theta}}^{\rm ML}, \boldsymbol{\theta}_0 \rangle|\} \geq 1 - \varepsilon$.*
- *No estimator can achieve $\mathbb{E}\{|\langle \hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0 \rangle|\} \geq \varepsilon$ unless $\lambda > \lambda_{\rm IT}$.*
- *There exists poly-time estimator achieving $\mathbb{E}\{|\langle \hat{\boldsymbol{\theta}}^{\rm Poly}, \boldsymbol{\theta}_0 \rangle|\} \geq 1 - \varepsilon$, provided $\lambda \geq C(\varepsilon) n^{(k-2)/4}$.*

No efficient estimator is known for $1 \ll \lambda \ll n^{(k-2)/4}$!

Spiked tensor model: What do we know?

$$\boldsymbol{Y} = \lambda \boldsymbol{\theta}_0^{\otimes k} + \boldsymbol{W}$$

Theorem (Montanari, Richard, 2014; Hopkins, Shi, Steurer, 2015)

For any $\varepsilon > 0$, there exist constants $\lambda_{\rm IT}$, $\lambda_{\rm ML}(\varepsilon)$, $C(\varepsilon)$, such that:

- *If $\lambda > \lambda_{\rm ML}(\varepsilon)$, then $\mathbb{E}\{|\langle \hat{\boldsymbol{\theta}}^{\rm ML}, \boldsymbol{\theta}_0\rangle|\} \geq 1 - \varepsilon$.*
- *No estimator can achieve $\mathbb{E}\{|\langle \hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0\rangle|\} \geq \varepsilon$ unless $\lambda > \lambda_{\rm IT}$.*
- *There exists poly-time estimator achieving $\mathbb{E}\{|\langle \hat{\boldsymbol{\theta}}^{\rm Poly}, \boldsymbol{\theta}_0\rangle|\} \geq 1 - \varepsilon$, provided $\lambda \geq C(\varepsilon) n^{(k-2)/4}$.*

No efficient estimator is known for $1 \ll \lambda \ll n^{(k-2)/4}$!

Spiked tensor model: What do we know?

$$\boldsymbol{Y} = \lambda \boldsymbol{\theta}_0^{\otimes k} + \boldsymbol{W}$$

Theorem (Montanari, Richard, 2014; Hopkins, Shi, Steurer, 2015)

For any $\varepsilon > 0$, there exist constants $\lambda_{\rm IT}$, $\lambda_{\rm ML}(\varepsilon)$, $C(\varepsilon)$, such that:

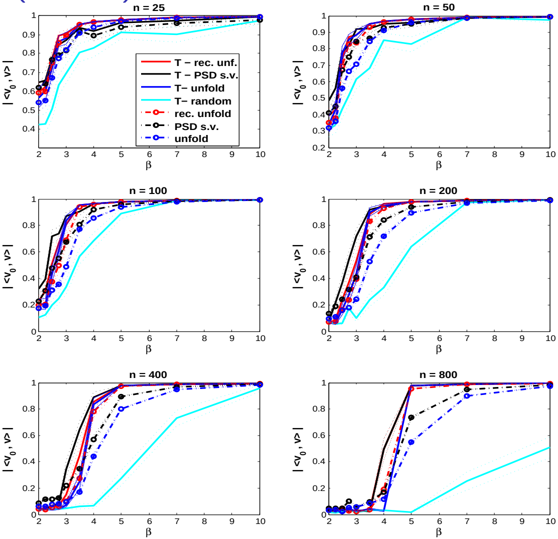
- *If $\lambda > \lambda_{\rm ML}(\varepsilon)$, then $\mathbb{E}\{|\langle \hat{\boldsymbol{\theta}}^{\rm ML}, \boldsymbol{\theta}_0\rangle|\} \geq 1 - \varepsilon$.*
- *No estimator can achieve $\mathbb{E}\{|\langle \hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0\rangle|\} \geq \varepsilon$ unless $\lambda > \lambda_{\rm IT}$.*
- *There exists poly-time estimator achieving $\mathbb{E}\{|\langle \hat{\boldsymbol{\theta}}^{\rm Poly}, \boldsymbol{\theta}_0\rangle|\} \geq 1 - \varepsilon$, provided $\lambda \geq C(\varepsilon) n^{(k-2)/4}$.*

No efficient estimator is known for $1 \ll \lambda \ll n^{(k-2)/4}$!

More precise results

- Montanari, Reichman, Zeitouni, 2015
- Bandeira, Perry, Wein, 2017
- Krzakala, Lelarge, Miolane, Zdeborova, 2017
- ...

In practice ($k = 3$)



What does the landscape look like?

Maximum likelihood

$$\begin{aligned} \text{minimize} \quad & \widehat{\mathcal{L}}_n(\boldsymbol{\theta}) = \left\| \boldsymbol{Y} - \lambda \boldsymbol{\theta}^{\otimes k} \right\|_F \\ \text{subject to} \quad & \|\boldsymbol{\theta}\|_2 = 1 \end{aligned}$$

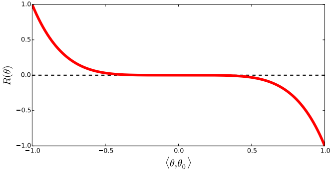
$$\widehat{\mathcal{L}}_n(\boldsymbol{\theta}) = \text{const.} - 2\lambda \langle \boldsymbol{Y}, \boldsymbol{\theta}^{\otimes k} \rangle$$

Maximum likelihood

$$\begin{aligned} \text{minimize} \quad & \widehat{\mathcal{L}}_n(\boldsymbol{\theta}) = \big\| \boldsymbol{Y} - \lambda \boldsymbol{\theta}^{\otimes k} \big\|_F \\ \text{subject to} \quad & \|\boldsymbol{\theta}\|_2 = 1 \end{aligned}$$

$$\widehat{\mathcal{L}}_n(\boldsymbol{\theta}) = \text{const.} - 2\lambda \langle \boldsymbol{Y}, \boldsymbol{\theta}^{\otimes k} \rangle$$

Risk



Maximum likelihood

$$\begin{aligned} \text{minimize} \quad & \widehat{\mathcal{L}}_n(\boldsymbol{\theta}) = -\langle \boldsymbol{Y}, \boldsymbol{\theta}^{\otimes k} \rangle \\ \text{subject to} \quad & \|\boldsymbol{\theta}\|_2 = 1 \end{aligned}$$

'Population' risk

$$\mathcal{L}(\boldsymbol{\theta}) = -\lambda \langle \boldsymbol{\theta}_0, \boldsymbol{\theta} \rangle^k$$

Back-of-the-envelope

Expected gradient

$$\nabla \mathcal{L}(\boldsymbol{\theta}) = -k\lambda \langle \boldsymbol{\theta}, \boldsymbol{\theta}_0 \rangle^{k-1} \boldsymbol{\theta}_0$$

Random initialization $\langle \boldsymbol{\theta}, \boldsymbol{\theta}_0 \rangle = \Theta(n^{-1/2})$:

$$\begin{aligned} \langle \boldsymbol{\theta}_0, \nabla \widehat{\mathcal{L}}_n(\boldsymbol{\theta}) \rangle &= -k\lambda \langle \boldsymbol{\theta}, \boldsymbol{\theta}_0 \rangle^{k-1} - k \langle \boldsymbol{W}, \boldsymbol{\theta}_0 \otimes \boldsymbol{\theta}^{\otimes(k-1)} \rangle \\ &= -\lambda \Theta(n^{-(k-1)/2}) + \Theta(n^{-1/2}) \end{aligned}$$

- Convergence: $\lambda \gg n^{(k-2)/2}$

Back-of-the-envelope

Expected gradient

$$\nabla \mathcal{L}(\boldsymbol{\theta}) = -k\lambda \langle \boldsymbol{\theta}, \boldsymbol{\theta}_0 \rangle^{k-1} \boldsymbol{\theta}_0$$

Random initialization $\langle \boldsymbol{\theta}, \boldsymbol{\theta}_0 \rangle = \Theta(n^{-1/2})$:

$$\begin{aligned}\langle \boldsymbol{\theta}_0, \nabla \widehat{\mathcal{L}}_n(\boldsymbol{\theta}) \rangle &= -k\lambda \langle \boldsymbol{\theta}, \boldsymbol{\theta}_0 \rangle^{k-1} - k \langle \boldsymbol{W}, \boldsymbol{\theta}_0 \otimes \boldsymbol{\theta}^{\otimes (k-1)} \rangle \\ &= -\lambda \Theta(n^{-(k-1)/2}) + \Theta(n^{-1/2})\end{aligned}$$

- Convergence: $\lambda \gg n^{(k-2)/2}$

Back-of-the-envelope

Expected gradient

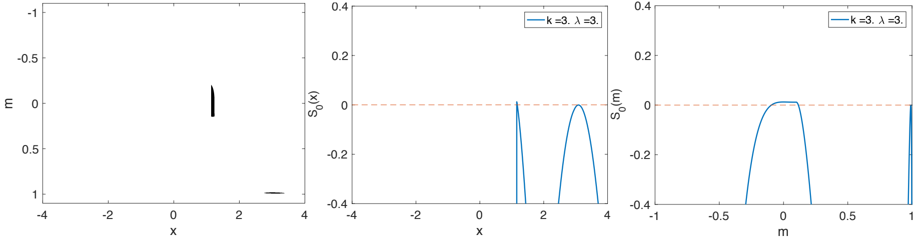
$$\nabla \mathcal{L}(\boldsymbol{\theta}) = -k\lambda \langle \boldsymbol{\theta}, \boldsymbol{\theta}_0 \rangle^{k-1} \boldsymbol{\theta}_0$$

Random initialization $\langle \boldsymbol{\theta}, \boldsymbol{\theta}_0 \rangle = \Theta(n^{-1/2})$:

$$\begin{aligned} \langle \boldsymbol{\theta}_0, \nabla \widehat{\mathcal{L}}_n(\boldsymbol{\theta}) \rangle &= -k\lambda \langle \boldsymbol{\theta}, \boldsymbol{\theta}_0 \rangle^{k-1} - k \langle \boldsymbol{W}, \boldsymbol{\theta}_0 \otimes \boldsymbol{\theta}^{\otimes(k-1)} \rangle \\ &= -\lambda \Theta(n^{-(k-1)/2}) + \Theta(n^{-1/2}) \end{aligned}$$

- Convergence: $\lambda \gg n^{(k-2)/2}$

Expected number of local minima: $k = 3$, $\lambda = 3$



- Exponential in black region $\quad (m = \langle \boldsymbol{\theta}, \boldsymbol{\theta}_0 \rangle,\ x = \langle \boldsymbol{Y}, \boldsymbol{\theta}^{\otimes k} \rangle)$
- N = number of local minima

$$\mathbb{E}N(m, x) = e^{nS_0(m,x)+o(n)}$$

[Ben Arous, Mei, Montanari, Nica, 2017]

Complexity of landscape $\overset{?}{\leftrightarrow}$ Complexity for local algorithms

SDP relaxations

An emerging dichotomy

In several statistical estimation problems

- Either local (or message-passing) algorithms work. . .
- . . . or SDP hierarchies do not work

Why?

An emerging dichotomy

In several statistical estimation problems

- Either local (or message-passing) algorithms work…
- …or SDP hierarchies do not work

Why?

An emerging dichotomy

In several statistical estimation problems

- Either local (or message-passing) algorithms work...
- ...or SDP hierarchies do not work

Why?

A possible explanation

Perhaps SDPs on random instances can be solved by local algorithms. . .

Simplest example

Centered adjacency matrix of $G = (V, E)$ $\quad$ (d = average degree)

$$A^{\rm cen}_{ij} = \begin{cases} 1 - \frac{d}{n} & \text{if } (i,j) \in E, \\ -\frac{d}{n} & \text{otherwise.} \end{cases}$$

SDP($\boldsymbol{A}^{\rm cen}$):

$$\begin{aligned} \text{maximize} \quad & \langle \boldsymbol{A}^{\rm cen}, \boldsymbol{X} \rangle , \\ \text{subject to} \quad & \boldsymbol{X} \in \mathbb{R}^{n \times n}, \ \boldsymbol{X} \succeq 0 , \\ & X_{ii} = 1 . \end{aligned}$$

- Graph clustering, embedding, testing latent structure,. . .

What does it mean?

Input: Graph $G_n = (V_n, E_n)$

1. Generate $\boldsymbol{z} = (z(i))_{i \in V} \sim_{iid} \mathsf{N}(0,1)$
2. Compute, for each $v \in V_n$, $\xi_v = F(\mathsf{B}_\ell(v; G_n), \boldsymbol{z}|_{\mathsf{B}_\ell(v;G_n)})$
3. Output $\boldsymbol{X} = \mathbb{E}_{\boldsymbol{z}}\{\boldsymbol{\xi}\boldsymbol{\xi}^{\mathsf{T}}\} + \cdots \in \mathbb{R}^{n \times n}$

Can this achieve $\langle \boldsymbol{A}^{\text{cen}}, \boldsymbol{X} \rangle \geq (1 - o_n(1))\mathsf{SDP}(\boldsymbol{A}^{\text{cen}})$?

Erdős-Renyi random graphs

Theorem (Fan, Montanari, 2016)

Let $G \sim \mathcal{G}(n, d/n)$ *and* $\boldsymbol{A}^{\text{cen}} = \boldsymbol{A}^{\text{cen}}_G$. *Then, a.s.,*

$$2\sqrt{d}\left(1 - \frac{1}{d+1}\right) \le \liminf_{n\to\infty} \frac{1}{n}\mathsf{SDP}(\boldsymbol{A}^{\text{cen}}) \le$$
$$\le \limsup_{n\to\infty} \frac{1}{n}\mathsf{SDP}(\boldsymbol{A}^{\text{cen}}) \le 2\sqrt{d}\left(1 - \frac{1}{2d}\right).$$

Further, the lower bound is achieved by local algorithms.

- A local algorithm achieves 8/9 of $\mathsf{SDP}(\boldsymbol{A}^{\text{cen}})$.

Erdős-Renyi random graphs

Theorem (Fan, Montanari, 2016)

Let $G \sim \mathcal{G}(n, d/n)$ *and* $\boldsymbol{A}^{\rm cen} = \boldsymbol{A}^{\rm cen}_G$. *Then, a.s.,*

$$2\sqrt{d}\left(1 - \frac{1}{d+1}\right) \le \lim\inf_{n\to\infty} \frac{1}{n}\mathsf{SDP}(\boldsymbol{A}^{\rm cen}) \le$$

$$\le \lim\sup_{n\to\infty} \frac{1}{n}\mathsf{SDP}(\boldsymbol{A}^{\rm cen}) \le 2\sqrt{d}\left(1 - \frac{1}{2d}\right).$$

Further, the lower bound is achieved by local algorithms.

- A local algorithm achieves 8/9 of $\mathsf{SDP}(\boldsymbol{A}^{\rm cen})$.

Erdős-Renyi random graphs

Theorem (Fan, Montanari, 2016)

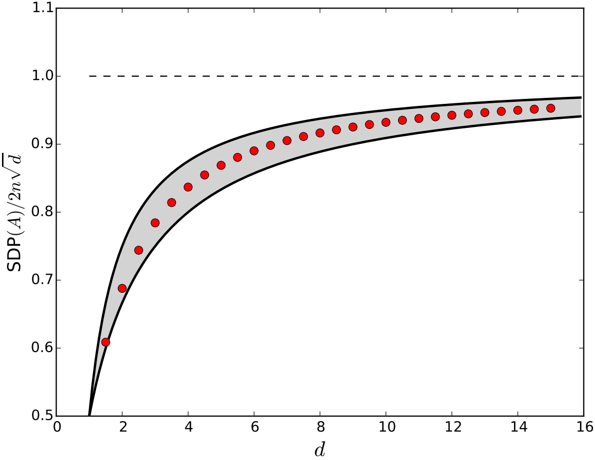
Let $G \sim \mathcal{G}(n, d/n)$ and $\boldsymbol{A}^{\text{cen}} = \boldsymbol{A}_G^{\text{cen}}$. Then, a.s.,

$$2\sqrt{d}\left(1 - \frac{1}{d+1}\right) \leq \liminf_{n\to\infty} \frac{1}{n}\mathsf{SDP}(\boldsymbol{A}^{\text{cen}}) \leq$$
$$\leq \limsup_{n\to\infty} \frac{1}{n}\mathsf{SDP}(\boldsymbol{A}^{\text{cen}}) \leq 2\sqrt{d}\left(1 - \frac{1}{2d}\right).$$

Further, the lower bound is achieved by local algorithms.

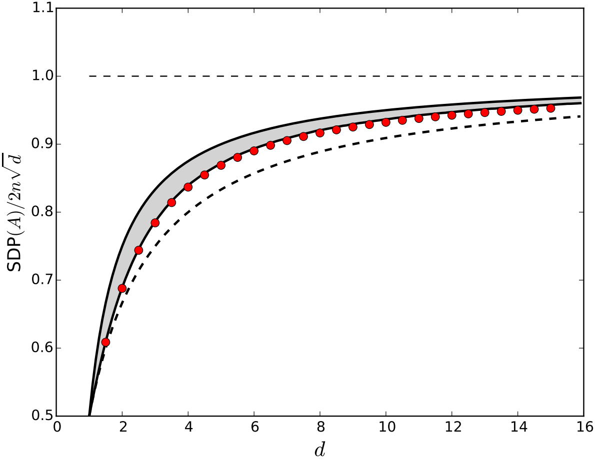
- A local algorithm achieves 8/9 of $\mathsf{SDP}(\boldsymbol{A}^{\text{cen}})$.

Bounds vs numerical simulations



[Related results for planted partition model, regular graphs, ...]

A better local algorithm (see paper)



[Related results for planted partition model, regular graphs, ...]

Conclusion

Conclusion

- Which statistical problems are tractable?
- Multiple points of view:
  - Local-message passing algorithms
  - Landscape analysis
  - SDP hierarchies

Thanks!

Conclusion

- Which statistical problems are tractable?
- Multiple points of view:
  - Local-message passing algorithms
  - Landscape analysis
  - SDP hierarchies

Thanks!