Mean Field Asymptotics in Statistical Learning. Jan 20th.

Lecture 1: Introduction to the mean field asymptotics.

Today's agenda:

A. Syllabus and logistics. (and questions).

1. Schedule. Office hours.
2. Zoom policy: don't share the Zoom link. Recordings.
3. Assignments (Coding, Proof, Latex). Projects. No exam. Final score.
4. Homework proof reader. Grader (4 times).
5. Scribe notes (volunteer for the first lecture).
6. Prerequisites. Difficulty adjustment.
7. No textbook. Reference see online syllabus.

B. Introduction to mean field asymptotics.

1. A motivating example: the LASSO problem.
2. Non-asymptotic theory.
3. Asymptotic theory.
4. Comparisons.
5. Why mean field? Why Statistical Physics?
6. Topics that will be covered.
7. Active research directions.
8. Level of rigorous.

① Motivating example: The LASSO problem.

Let $x_0 \in \mathbb{R}^d$, $A \in \mathbb{R}^{n\times d}$, $w \in \mathbb{R}^n$, $y = Ax_0 + w \in \mathbb{R}^n$.

(x_0 will be k-sparse, $k << n << d$).

We are given (A, y). Estimate x_0.

LASSO: $\hat{x} = \arg\min_x \frac{1}{2n}\|y - Ax\|_2^2 + \frac{\lambda}{n}\|x\|_1$

Goal: quantify / bound $\|\hat{x} - x_0\|_2^2 / \|x_0\|_2^2$.

② Non-asymptotic theory

[Negabhan, Ravikumar, Wainwright, Yu, 2012].

Definition (Restricted Strong Convexity).

We say $A \in \mathbb{R}^{n\times d}$ satisfy RSC if

$\exists\ c_1, c_2$, s.t. $\forall\ v \in \mathbb{R}^d$, we have

$$\underbrace{\frac{\|Av\|_2^2}{n} \geq c_1 \cdot \|v\|_2^2} - c_2 \frac{\log d}{n}\|v\|_1^2$$

$f(x) = \frac{1}{2n}\|y - Ax\|_2^2$, $\nabla^2 f(x) = \frac{A^T A}{n}$

SC: $\nabla^2 f(x) \succeq c_1 \cdot I. \Rightarrow \frac{\|Av\|_2^2}{n} \geq c_1 \cdot \|v\|_2^2$.

Thm ([NRWY12]). (Simplified)

$\forall$ A satisfy RSC w/ const c_1, c_2.

$\exists\ C < \infty$. s.t. as long as $\lambda \geq 2 \cdot \|A^T w\|_\infty$.

$\forall\ x_0 \in \mathbb{R}^d$ with support $S \subseteq \{1, 2, \cdots, d\}$

$|S| \leq \frac{n}{c \cdot \log d}$,

$\|\hat{x} - x_0\|_2^2 \leq C \cdot \frac{\lambda^2 |S|}{n^2}$.

Prop: We take $A_{ij} \sim$ i.i.d. $N(0, 1)$.

Then A satisfy RSC with high probability.

Corollary: Let $A \in \mathbb{R}^{n\times d}$ with $A_{ij} \sim N(0, \frac{1}{\|x_0\|_2^2})$. $|S| \leq k$

Let x_0 be a k-sparse vector w/ support of x_0 to be S,

Let $w_i \sim$ i.i.d. $N(0, \sigma^2)$. Then for any $\delta > 0$, $\exists\ C(\delta)$,

s.t. as long as $n \geq C(\delta) \cdot k \log d$, choose $\lambda = C(\delta) \cdot \sigma \sqrt{n \log d}$.

Then with prob. at least $1 - \delta$,

$$\|\hat{x} - x_0\|_2^2 / \|x_0\|_2^2 \leq \frac{C(\delta)\, \sigma^2 k \log d}{n}.$$

Implication: $n \gg \sigma^2 k \log d$. LASSO consistent.

① Everything is explicit, no limiting statement.

② Very weak assumption.

③ High-dimensional asymptotics of LASSO.

Thm ([Bayati, Montanari, 2012]).

Consider the asymptotic regime $n/d \to \delta \in (0, \infty)$ as $d \to \infty$. Let $A_{ij} \sim N(0, \frac{1}{n})$. Let $x_0 \in \mathbb{R}^d$, $x_{0,i} \sim_{i.i.d.} \mathbb{P}_0$. Let $w_i \sim_{i.i.d.} N(0, \sigma^2)$. $y = Ax_0 + w$.

$$\hat{x} = \arg\min_x \frac{1}{2n} \|y - Ax\|_2^2 + \frac{\lambda}{n}\|x\|_1.$$

$$\lim_{\substack{n/d \to \infty \\ d \to \infty}} \|\hat{x} - x_0\|_2^2 / \|x_0\|_2^2$$

$$= \mathbb{E}_{(X_0, Z) \sim \mathbb{P}_0 \times N(0,1)} \left[\left(\eta(X_0 + \tau_* Z ; \tau_* \alpha_*) - X_0 \right)^2 \right].$$

$$\eta(x; \theta) = \arg\min_v \frac{1}{2}(x - v)^2 + \theta |v|^2 = \mathrm{sign}(x) \cdot (|x| - \theta)_+$$

(τ_*, α_*) satisfy some explicit non linear equation.

$\tau_* = \tau_*(\alpha_*)$ is the largest sol. of

$$\tau^2 = \sigma^2 + \delta^{-1} \mathbb{E}_{(X_0, Z) \sim \mathbb{P}_0 \times N(0,1)} \{ [\eta(X_0 + \tau Z; \alpha_* \tau) - X_0]^2 \}.$$

and α_* is the unique non-negative sol. of

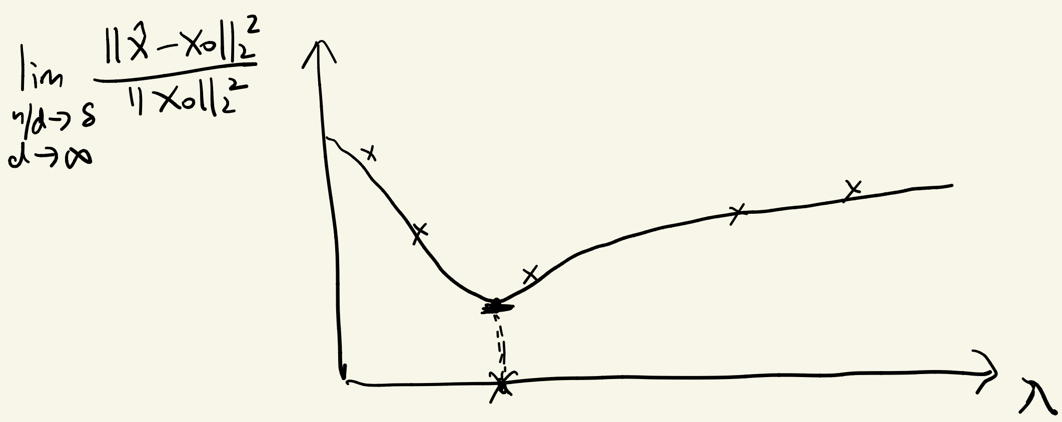
$$\lambda = \alpha \tau_*(\alpha) \cdot \left[1 - \delta^{-1} \mathbb{E}[\eta'(X_0 + \tau_*(\alpha) Z; \alpha \tau_*(\alpha))] \right].$$

Estimation error of a one dimensional problem.

$X_0 \sim \mathbb{P}_0$, $Z \sim N(0,1)$, $Y = X_0 + \tau_* Z$.

$$\hat{X} = \arg\min_v (Y - v)^2 + \tau_* \alpha_* |v|$$

$\mathbb{E}\left[(\hat{X} - X_0)^2 \right]$ estimation error.



① Strong assumptions.

② Sharp result.

③ Weaker SNR regime.

④ Comparison.

	Non-asymptotic	Asymptotic
Typical regime	Strong SNR	Weaker SNR.
Advantage	Less model assumption.	precise formula.
Limitations.	A $\Theta(1)$ gap between upper and lower bound.	More model assumptions.
When useful :	General assumption.	Pin down phase transition.
Example :	Statical learning theory. uniform converge bound	Phase transition of compressed sensing ; double-descent.

⑤ Mean-field.

Mean field : In physics and probability theory, mean-field theory studies the behavior of high-dimensional random models by studying a simpler (low-dimensional) model that approximates the original one by averaging over the degree of freedom.

Expressing through empirical distribution of coordinates.

Why statistical physics.

next 3 lectures.

Optimization
Bayes inference. $\longleftrightarrow$ Gibbs measure of spin glass model. (Sherrington-Kirkpatrick model). 1975

applications:

Coding theory
random combinatorial optimization.
Statistics learning

Heuristic ~~tool~~ tool. 1980s - 1990s.

Replica method
Cavity method
TAP approach
Kac-Rice formula
Dynamical MF theory.

Rigorous methods:

CGMT
AMP
leave-one-out

⑥ Topics.

a) Connection between statistical decision theory & statistical physics.

b). Concentration inequalities.

c). Replica method.

d). CGMT.

e). Random metrices.

f). AMP.

⑦ Active reserach direction.

a). Neural networks.

b). Active learning.

c) Adversial robustness

d) Unsupervised / semi-supervised.

e). ....

⑧ Level of rigorous.

No measure thoretic issue.

Sometimes: assume differentiability.
exchange of limit
exchange of limit and differentiation.