Mean Field Asymptotics in Statistical Learning.

Mar 15.

Lecture 15. CGMT and LASSO Asymptotics.

[Stojnic, 2013]
[Thrampoulidis, Oymak, Hassibi, 2015].

① Convex Gaussian minimax theorem.

Thm (CGMT). Let $S_u \subseteq \mathbb{R}^m$, $S_v \subseteq \mathbb{R}^n$ be compact sets, ψ be continous on $S_u \times S_v$, and $G \in \mathbb{R}^{m\times n}$, $g \in \mathbb{R}^n$, $h \in \mathbb{R}^m$, with $G_{ij}, g_j, h_i \sim \text{iid } N(0,1)$.

Define
$$\Phi(G) = \min_{u \in S_u} \max_{v \in S_v} \langle u, Gv \rangle + \psi(u,v) \qquad \text{(PO)}$$
$$\phi(g,h) = \min_{u \in S_u} \max_{v \in S_v} \|u\|_2 \langle g, v\rangle + \|v\|_2 \langle h, u\rangle + \psi(u,v) \qquad \text{(AO)}$$

(a) $\forall\, \tau \in \mathbb{R}$
$$\mathbb{P}(\Phi(G) \le \tau) \le 2\,\mathbb{P}(\phi(g,h) \le \tau).$$

(b) Further assume that S_u, S_v are convex set and ψ is convex-concave on $S_u \times S_v$ (Strong duality hold)

Then for all $\tau \in \mathbb{R}$.
$$\mathbb{P}(\Phi(G) \ge \tau) \le 2\,\mathbb{P}(\phi(g,h) \ge \tau).$$

In particular, $\forall\, \mu \in \mathbb{R}$, $t > 0$,
$$\mathbb{P}(|\Phi(G) - \mu| > t) \le 2 \cdot \mathbb{P}(|\phi(g,h) - \mu| \ge t).$$

Steps to apply CGMT.

(A) Figure out how to transform the original problem to primary problem.

(B) Write down the auxiliary optimization problem.

(C) Analyze the auxiliary problem.

(D) Extract information from the auxiliary problem. ← Derivation of formula.

(E) Proof.

② The LASSO risk

Signal: $x_0 \in \mathbb{R}^d$, Noise: $w \in \mathbb{R}^n$, Sensing matrix: $A \in \mathbb{R}^{n\times d}$.

Response: $y = Ax_0 + w \in \mathbb{R}^n$. $A_{ij} \sim_{iid} N(0, \frac{1}{n})$, $w_i \sim_{iid} N(0, \sigma^2)$, $x_{0j} \sim_{iid} \mathbb{P}_0$

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

We have seen in the replica derivation: $\psi : \mathbb{R}^2 \to \mathbb{R}$

$$\lim_{\substack{d\to\infty \\ n/d\to\delta}} \mathbb{E}\left[\frac{1}{d}\sum_{j=1}^d \psi(\hat{x}_j, x_{0,j})\right] = \mathbb{E}_{(x_0, G)\sim \mathbb{P}_0\times N(0,1)}\left[\psi(\eta(x_0 + \tau_* G; \alpha_* \tau_*), x_0)\right].$$

where (τ_*, α_*) solves some self-consistent equations.

Denote $\hat{\mu}_\lambda = \frac{1}{d}\sum_{j=1}^d \delta_{(\hat{x}_{j,\lambda}, x_{0,j})}$, $\bar{\mu}_\lambda$ be the distribution of $(\eta(x_0 + \tau_* G; \alpha_* \tau_*), x_0)$ when $(x_0, G) \sim \mathbb{P}_0 \times N(0,1)$.

Then $\hat{\mu}_\lambda \to \bar{\mu}_\lambda$ in some weak sense.

Definition: (Wasserstein-2 distance). $\mu, \nu \in \mathcal{P}(\mathbb{R}^k)$.

$$W_2(\mu, \nu) = \left[\inf_{r\in \mathcal{C}(\mu,\nu)} \int \|x-y\|_2^2\, r(dx dy)\right]^{\frac{1}{2}}.$$

Remark: ① $\lim_{d\to\infty} W_2(\mu_d, \mu_*) = 0 \iff \lim_{d\to\infty} \int f(x)\mu_n(dx) = \int f(x)\mu_*(dx)$
$\forall\ |f(x) - f(y)| \le L\cdot(1 + |x| + |y|)\,|x-y|$.

② $W_2\left(\frac{1}{d}\sum_{j=1}^d \delta_{x_j}, \frac{1}{d}\sum_{j=1}^d \delta_{y_j}\right) \le \|x - y\|_2^2/d$.

A control of Euclidean distance gives a control of W_2.

Theorem ([Miolane, Montanari, 2018])

$0 < \lambda_{min} < \lambda_{max} < \infty$, $N/n = \delta$. $B < \infty$,

$\exists\ c, C$ constants depending on $(\delta, \sigma^2, B, \lambda_{min}, \lambda_{max})$, s.t.

$$\sup_{\frac{\|x_0\|_2^2}{d}\le B} \mathbb{P}\left(\sup_{\lambda\in[\lambda_{min}, \lambda_{max}]} W_2(\hat{\mu}_\lambda, \bar{\mu}_\lambda)^2 \ge \varepsilon\right) \le \frac{C}{\varepsilon^2}\exp\left[-c\, d\, \varepsilon^3/\log^2(\varepsilon)\right].$$

Benefit: Non-asymptotic, uniform in x_0 and λ.

Asymptotic formula is still correct when λ is adaptively chosen.

③ Derivation of the LASSO risk (Calculation instead of proof).

Goal: derive $m_* \equiv \lim_{d\to\infty} \mathbb{E}\left[\frac{1}{d}\sum_{i=1}^{d}\psi(\hat{x}_i, x_{0,i})\right]$. when $n/d \to \delta$.

Free energy approach ($\beta=\infty$): the perturbed LASSO objective.

$$f(h) \equiv \lim_{d\to\infty} \mathbb{E}\left[\min_x \left\{ \frac{1}{2d}\|y-Ax\|_2^2 + \frac{\lambda}{d}\|x\|_1 + h\frac{1}{d}\sum_{i=1}^{d}\psi(x_i, x_{0i}) \right\}\right].$$

(need to argue why limit and derivative can be exchanged).

$$\Rightarrow f'(0) = \lim_{d\to\infty} \mathbb{E}\left[\frac{1}{d}\sum_{i=1}^{d}\psi(\hat{x}_i, x_{0,i})\right].$$

Derivation:

$$f(h)\times d \overset{①}{=} \min_{u\in\mathbb{R}^d} \left[\frac{1}{2}\|Au+w\|_2^2 + \lambda\|x_0+u\|_1 + h\sum_{j=1}^{d}\psi(x_j, x_{0j})\right]$$

① : take $u = x - x_0$.

$$\overset{②}{=} \min_{u\in\mathbb{R}^d}\max_{v\in\mathbb{R}^n}\Big[\underbrace{\langle v, Au\rangle + \langle v, w\rangle - \tfrac{1}{2}\|v\|_2^2 + \lambda\|x_0+u\|_1 + h\sum_{j=1}^{d}\psi(x_j,x_{0j})}_{Q(u,v) \text{ in CGMT}}\Big]$$

② : $\frac{1}{2}\|z\|_2^2 = \sup_v\left[\langle v, z\rangle - \frac{1}{2}\|v\|_2^2\right]$, $z = Au+w$

$$\overset{③}{\approx} \min_{u\in\mathbb{R}^d}\max_{v\in\mathbb{R}^n}\Big[\|u\|_2\langle g, v\rangle/\sqrt{n} + \|v\|_2\langle h, u\rangle/\sqrt{n} + \langle v, w\rangle - \tfrac{1}{2}\|v\|_2^2 + \lambda\|x_0+u\|_1 + h\cdot M\Big]$$

$M = \sum_{j=1}^{d}\psi(x_j, x_{0j})$

③ : CGMT with $G = \sqrt{n}\times A$, $\psi(u,v) = \sqrt{n}\times Q(u,v)$. Need truncation argu.

$$\overset{④}{=} \min_{u\in\mathbb{R}^d}\max_{\beta\geq 0}\Big[\big(\big\|\,\|u\|_2 g + \sqrt{n}\cdot\sigma\bar{w}\big\|_2 + \langle h, u\rangle\big)\times\beta - \tfrac{n}{2}\beta^2 + \lambda\|x_0+u\|_1 + hM\Big].$$

$w = \sigma\bar{w}$

$\bar{w} \sim N(0, I_d)$

④ : $\max_v f(v) = \max_{\beta\geq 0}\max_{\|v\|_2=\beta} f(v)$

$$\overset{⑤}{\approx} \min_{u\in\mathbb{R}^d} \max_{\beta\geq 0} \left[\sqrt{\frac{\|u\|_2^2}{n}+\sigma^2} \times n + \langle h, u\rangle \right] \times \beta - \frac{n}{2}\beta^2 + \lambda \|x_0+u\|_1 + hM.$$

⑤ : $\left\| \frac{\|u\|_2}{\sqrt{n}} g + \sigma \bar{w} \right\|_2 \overset{d}{=} \left\| N\left(0, \left(\frac{\|u\|_2^2}{n}+\sigma^2\right) I_n\right) \right\|_2 \approx \sqrt{n}.$

$$\overset{⑥}{=} \min_{u\in\mathbb{R}^d} \max_{\beta\geq 0} \min_{\tau\geq 0} \left\{ n \times \left[\frac{\tau}{2} + \frac{1}{2\tau}\left(\frac{\|u\|_2^2}{n}+\sigma^2\right) \right] + \langle h, u\rangle \right\} \times \beta - \frac{n}{2}\beta^2 + \lambda \|x_0+u\|_1 + hM$$

⑥ : $\sqrt{a} = \min_{\tau>0} \left\{ \frac{1}{2}\tau + \frac{a}{2\tau} \right\}$, $a = \frac{\|u\|_2^2}{n} + \sigma^2$

Why this transformation? We hope the objective function to be a separable function of u, so that the high dim opt. can be done coordinatewisely. Like delta identity formula + saddle point approx

Alternative equivalent approach: Legendre transform

$$f\left(\sum_{j=1}^d \phi(u_j)\right) = \mathrm{ext}_{\mu, m} \left[f(m) + \mu\left(\sum_{j=1}^d \phi(u_j) - m\right) \right] = \mathrm{ext}_{\mu} \left[\mu \sum_{j=1}^d \phi(u_j) - f^*(\mu) \right]$$

$$\overset{⑦}{=} d \times \max_{\beta\geq 0} \min_{\tau} \left\{ \left(\frac{\sigma^2}{\tau}+\tau\right)\frac{\beta\delta}{2} - \frac{\delta}{2}\beta^2 + \min_{u} \left[\frac{1}{d}\sum_{j=1}^d \left(\frac{\beta}{2\tau}u_j^2 + \beta h_j u_j + \lambda |u_j + x_{0,j}| + h\psi(u_j + x_{0,j}, x_{0,j}) \right) \right] \right\}$$

⑦ : Exchange "$\min_u$" with "$\max_{\beta\geq 0} \min_{\tau\geq 0}$" by strong duality.

Thm B.2 in [Miolane, Montanari 2018].

Don't bother changing "min" "max" during derivation!

Just think about them as "ext"

Explaining why is the task during proof, not derivation.

$$\overset{⑧}{=} d \times \max_{\beta\geq 0} \min_{\tau} \left\{ \left(\frac{\sigma^2}{\tau}+\tau\right)\frac{\beta\delta}{2} - \frac{\delta}{2}\beta^2 + \mathbb{E}_{(X_0, G)\sim \mathbb{P}_0\times N(0,1)} \left[\min_{u\in\mathbb{R}} \left(\frac{\beta}{2\tau}u^2 - \beta G u + \lambda |u + X_0| + h\psi(u+X_0, X_0) \right) \right] \right\}.$$

⑧: Uniform law of large number

$$\frac{1}{d}\sum_{j=1}^{d}\phi(x_{0,j}, h_j) \to \mathbb{E}[\phi(X_0, G)]$$

$$\phi(x,h) = \min_{u\in\mathbb{R}}\left(\frac{\beta}{2\tau}u^2 + \beta h u + \lambda|u+x_0| + h\psi(u+x_0, x_0)\right).$$

Uniform over β and τ.

Need concentration inequality and truncation argument

$$\Rightarrow f(h) \to \max_{\beta\geq 0}\min_{\tau}\Big\{\left(\frac{\sigma^2}{\tau}+\tau\right)\frac{\beta\delta}{2} - \frac{\delta}{2}\beta^2$$

$$+ \mathbb{E}_{(X_0,G)\sim\mathbb{P}_0\times N(0,1)}\left[\min_u\left(\frac{\beta}{2\tau}u^2 - \beta G u + \lambda|u+X_0| + h\psi(u+X_0, X_0)\right)\right]\Big\}.$$

$$\mathbb{E}\left[\frac{1}{d}\sum_{j=1}^{d}\psi(\hat{x}_j, x_{0,j})\right] \overset{⑨}{\approx} f'(0) = \mathbb{E}_{(X_0,G)}[\psi(\hat{X}, X_0)]$$

$$\hat{X} = \arg\min_u\left(\frac{\beta}{2\tau}u^2 - \beta G u + \lambda|u+X_0|\right) + X_0$$

$$= \eta\left(X_0 + \tau_* G;\ \frac{\tau_*\lambda}{\beta_*}\right).$$

τ_*, β_* solves

$$\begin{cases} \tau^2 = \sigma^2 + \delta^{-1}\mathbb{E}\left[\left(\eta\left(X_0+\tau G;\ \tau\frac{\lambda}{\beta}\right) - X_0\right)^2\right]. \\ \beta = \tau\left(1 - \delta^{-1}\mathbb{E}\left[\eta'\left(X_0+\tau G;\ \tau\frac{\lambda}{\beta}\right)\right]\right). \end{cases}$$

Coincide with the replica prediction.

Remark: 1) CGMT only handles optimization problem where Replica method can handle Bayes inference problem.

2) CGMT analysis is rigorous, and non-asymptotic.

④ W_2 convergence.

③, ⑤, ⑦, ⑧, ⑨.

$\underbrace{③, ⑤, ⑦, ⑧}_{\text{Straight-forward}}$, $\underbrace{⑨}_{\text{Need careful treatment.}}$

Problem of ⑨:

We need W_2 convergence (uniform for all psedo-Lipschitz test) instead of looking at a specific test function ψ.

A few steps in showing W_2 convergence:

Denote

Original problem: $\Phi(u) = \frac{1}{2d} \|Au+w\|_2^2 + \frac{\lambda}{d}\|u+x_0\|_1$

Gordon problem after simplification: $\Gamma(u) = \left[\frac{\delta}{2}\left(\tau_* + \frac{\sigma^2}{\tau_*}\right) + \frac{\|u\|_2^2}{2\tau_* d} + \frac{\langle h, u\rangle}{d}\right]\beta_* - \frac{\delta}{2}\beta_*^2 + \lambda \|x_0+u\|_1$

Φ_* = asymptotic LASSO risk.

CGMT gives

1) $\forall S$ (almost compact)

$$\mathbb{P}\left(\min_{u\in S} \Phi(u) \le \Phi_* + \varepsilon\right) \lesssim \mathbb{P}\left(\min_{u\in S} \Gamma(u) \le \Phi_* + \varepsilon\right)$$

2) When T is convex (almost compact)

$$\mathbb{P}\left(\left|\min_{u\in T} \Phi(u) - \Phi_*\right| \ge \varepsilon\right) \lesssim \mathbb{P}\left(\left|\min_{u\in T} \Gamma(u) - \Phi_*\right| \ge \varepsilon\right).$$

☆ Take $T = \mathbb{R}^d$, $S = S_\delta = \mathbb{R}^d \setminus \{W_2(\hat{\mu}_{(x_0+u, x_0)}, \bar{\mu}_\lambda) \ge \delta\}$

Suppose we show $\mathbb{P}\left(\min_{u\in S_\delta} \Gamma(u) \le \Phi_* + \varepsilon\right) \to 0$. *

then $\mathbb{P}\left(\min_{u:\, W_2(\hat{\mu}_{(u+x_0, x_0)}, \bar{\mu}_\lambda) \ge \varepsilon} \Phi(u) \le \min_{u\in\mathbb{R}^d} \Phi(u) + \varepsilon\right) \to 0$

$\Rightarrow$ For $\hat{u} = \arg\min_{u\in\mathbb{R}^d} \Phi(u)$ $\Rightarrow$ $W_2(\hat{\mu}_{(x_0+\hat{u}, x_0)}, \bar{\mu}_\lambda) \le \delta$ w.h.p

To show *, suffice to show the strong convexity of $I^*(u)$ near global min, since

$$W_2(\hat{\mu}_{(u+x_0, x_0)}, \bar{\mu}_\lambda) \lesssim \|u - \hat{u}_I\|_2^2 / d$$

where $\hat{u}_I = \underset{u}{\operatorname{argmin}} \; I(u)$.