STAT260 Mean Field Asymptotics in Statistical Learning Lecture 15 - 03/15/2021

Lecture 15: CGMT and Lasso asymptotics

Lecturer: Song Mei *Scriber: Wenlong Mou* *Proof reader: Zitong Yang*

In this lecture, we give a rigorous proof of the Lasso asymptotic risk based on the convex Gaussian minimax theorem. In particular, we establish a non-asymptotic result that certifies the heuristic derivation based on replica method in previous lectures. This also serves as an example of application of CGMT to penalized least square problems.

1 Statement of the theorem

We consider the standard setup of Lasso, with signal vector $\boldsymbol{x}_0 \in \mathbb{R}^d$, design matrix $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ and noise $\boldsymbol{w} \in \mathbb{R}^n$. The response is generated from $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_0 + \boldsymbol{w}$. We make the following probabilistic assumptions on the model:

$$x_{0,i} \sim_{i.i.d.} \mathbb{P}_0, \quad w_i \sim_{i.i.d.} \mathcal{N}(0, \sigma^2), \quad A_{ij} \sim_{i.i.d.} \mathcal{N}(0, 1/n).$$

We denote $\delta := n/d$, and will work with the proportional limit with $n \to +\infty$ given this ratio fixed.

The lasso estimator is then defined as:

$$\widehat{\boldsymbol{x}} := \arg\min_{\boldsymbol{x} \in \mathbb{R}^d} \left\{ \frac{1}{2n} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \frac{\lambda}{n} \|\boldsymbol{x}\|_1 \right\}. \tag{1}$$

We are interested in the joint empirical distribution of the coordinates for the ground truth $\boldsymbol{x}_0$ and the estimator $\widehat{\boldsymbol{x}}$. More precisely, we define the following random measure:

$$\widehat{\mu}_\lambda := \frac{1}{d} \sum_{j=1}^d \delta_{(x_{0,j}, \widehat{x}_j)}. \tag{2}$$

Before stating the theorem, we first define a few useful quantities. Define η to be the soft-thresholding function $\eta(x; z) := \mathrm{sign}(x) \cdot (|x| - z)_+$ for $z > 0$.

Let $\tau_\star = \tau_\star(\beta)$ to be the largest solution to the equation

$$\tau^2 = \sigma^2 + \delta^{-1} \mathbb{E}_{(X_0, G) \sim \mathbb{P}_0 \times \mathcal{N}(0,1)} \left[\left(\eta(X_0 + \tau G; \frac{\tau\lambda}{\beta}) - X_0 \right)^2 \right], \tag{3a}$$

and $\beta_\star$ is the unique non-negative solution to

$$\beta = \tau_\star(\beta) \cdot \left(1 - \delta^{-1} \mathbb{E}_{(X_0, G) \sim \mathbb{P}_0 \times \mathcal{N}(0,1)} \left[\eta'\big(X_0 + \tau_\star(\beta) G; \frac{\tau_\star(\beta)\lambda}{\beta}\big) \right] \right), \tag{3b}$$

and let $\tau_\star := \tau_\star(\beta_\star)$.

Finally, we define the joint law $\bar{\mu}_\lambda$ to be the joint law of $\Big(\eta(X_0 + \tau_\star G; \frac{\tau_\star \lambda}{\beta_\star}), X_0\Big)$, where $X_0 \sim \mathbb{P}_0$ and $G \sim \mathcal{N}(0, 1)$ are independent.

Now we are ready to state the main theorem.

Theorem 1 ([?])**.** *Under above setup, for any $0 < \lambda_{\min} \le \lambda_{\max} < +\infty$ and $B > 0$, there exist constants $c, C > 0$ depending on $(\lambda_{\min}, \lambda_{\max}, B, \delta, \sigma^2)$. The LASSO estimator (1) satisfies the following bound for any $\varepsilon > 0$:*

$$\sup_{\|x_0\|_2^2/d \le B^2} \mathbb{P}\left(\sup_{\lambda \in [\lambda_{\min}, \lambda_{\max}]} \mathcal{W}_2^2(\widehat{\mu}_\lambda, \bar{\mu}_\lambda) > \varepsilon \right) \le \frac{C}{\varepsilon^2} \exp\left(-cd\varepsilon^3 / \log^2(\varepsilon) \right). \tag{4}$$

A few remarks are in order. First, the result is stated as completely non-asymptotic, without a limit involve. A caveat, though, is that the constants (c, C) can depend on the problem parameters. Second, the result gives a strong convergence in Wasserstein distance, which implies the convergence of empirical average under any uniformly continuous test function dominated by a quadratic function. Finally, the probablistic convergence is uniform in λ. This allows for data-dependent choice of λ, which is often used in practice.

The rest of this lecture is devoted to a proof sketch of Theorem 1. The key technical tool is convex Gaussian minimax theorem introduced from last lecture, which is recalled as follows:

Theorem 2 ([**?**]). *Let $S_u \subseteq \mathbb{R}^m$ and $S_v \subseteq \mathbb{R}^n$ be compact sets, consider a continuous function ψ on $S_u \times S_v$. Define the following functions:*

$$\Phi(\boldsymbol{G}) := \min_{\boldsymbol{u}\in S_u} \max_{\boldsymbol{v}\in S_v} \langle \boldsymbol{u}, \boldsymbol{G}\boldsymbol{v}\rangle + \psi(\boldsymbol{u}, \boldsymbol{v}), \quad \textit{and} \tag{5a}$$

$$\phi(\boldsymbol{g}, \boldsymbol{h}) := \min_{\boldsymbol{u}\in S_u} \max_{\boldsymbol{v}\in S_v} \|\boldsymbol{u}\|_2 \langle \boldsymbol{g}, \boldsymbol{v}\rangle + \|\boldsymbol{v}\|_2 \langle \boldsymbol{h}, \boldsymbol{u}\rangle + \psi(\boldsymbol{u}, \boldsymbol{v}) \tag{5b}$$

For $G_{i,j}, g_i, h_j \sim_{i.i.d.} \mathcal{N}(0, 1)$, the following results hold true:

(a) For any $\tau \in \mathbb{R}$, we have:

$$\mathbb{P}\big(\Phi(\boldsymbol{G}) \leq \tau\big) \leq 2\mathbb{P}\big(\phi(\boldsymbol{g}, \boldsymbol{h}) \leq \tau\big). \tag{6a}$$

(b) Assume in addition that S_u and S_v are both convex, and the function ψ is convex-concave on $S_u \times S_v$, for any $\tau \in \mathbb{R}$, there is:

$$\mathbb{P}\big(\Phi(\boldsymbol{G}) \geq \tau\big) \leq 2\mathbb{P}\big(\phi(\boldsymbol{g}, \boldsymbol{h}) \geq \tau\big). \tag{6b}$$

See Lecture 14 for the proof of this theorem.

2 Derivation of the limit and proof sketch

In this section, we derive the limiting empirical distribution (2) using Theorem 2, and provide a sketch of the proof for Theorem 1. Note that the purpose of this section is just to give an overview of the method of derivation based on CGMT. Several steps in the derivation are not fully justified in this note. In the next section, we will discuss how to make the proof rigorous, highlighting the key techniques. The readers are encouraged to read the paper [**?**] for a complete proof.

In order to prove the convergence in Wasserstein distance, we take a sufficiently smooth test function $\psi : \mathbb{R}^2 \to \mathbb{R}$, and define the perturbed objective:

$$f(\gamma) := \lim_{d\to+\infty} \min_{\boldsymbol{x}} \left\{ \frac{1}{2d}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \frac{\lambda}{d}\|\boldsymbol{x}\|_1 + \frac{\gamma}{d}\sum_{i=1}^{d} \psi(x_i, x_{0,i}) \right\}. \tag{7}$$

Assuming that $f(\gamma)$ concentrates around its expectation uniformly over γ, and assume the regularity conditions that allow exchanging limit with expectation and derivatives, one would hope to get:

$$f'(0) = \lim_{d\to+\infty} \mathbb{E}\left[\frac{1}{d}\sum_{i=1}^{d} \psi(\widehat{x}_i, x_{0,i})\right]. \tag{8}$$

Taking $\boldsymbol{u} = \boldsymbol{x} - \boldsymbol{x}_0$, we start with the following derivation:

$$\begin{aligned}
& f(\gamma) \times d \\
&= \min_{\boldsymbol{u}} \max_{\boldsymbol{v}} \left\{ \langle \boldsymbol{v}, \boldsymbol{A}\boldsymbol{u} \rangle + \langle \boldsymbol{v}, \boldsymbol{w} \rangle - \frac{1}{2}\|\boldsymbol{v}\|_2^2 + \lambda \|\boldsymbol{u} + \boldsymbol{x}_0\|_1 + \gamma \sum_{j=1}^d \psi(u_j + x_{0,j}, x_{0,j}) \right\} && \text{(9a)} \\
&\approx \min_{\boldsymbol{u}} \max_{\boldsymbol{v}} \left\{ \|\boldsymbol{u}\|_2 \langle \boldsymbol{g}, \boldsymbol{v} \rangle / \sqrt{n} + \|\boldsymbol{v}\|_2 \langle \boldsymbol{h}, \boldsymbol{u} \rangle / \sqrt{n} + \langle \boldsymbol{v}, \boldsymbol{w} \rangle - \frac{1}{2}\|\boldsymbol{v}\|_2^2 + \lambda \|\boldsymbol{u} + \boldsymbol{x}_0\|_1 + \gamma \sum_{j=1}^d \psi(u_j + x_{0,j}, x_{0,j}) \right\} && \text{(9b)} \\
&= \min_{\boldsymbol{u}} \max_{\beta \geq 0} \left\{ \beta \cdot \left(\|\|\boldsymbol{u}\|_2 \boldsymbol{g} + \sqrt{n} \cdot \boldsymbol{w}\|_2 + \langle \boldsymbol{h}, \boldsymbol{u} \rangle \right) - \frac{n}{2}\beta^2 + \lambda \|\boldsymbol{u} + \boldsymbol{x}_0\|_1 + \gamma \sum_{j=1}^d \psi(u_j + x_{0,j}, x_{0,j}) \right\} && \text{(9c)} \\
&\approx \min_{\boldsymbol{u}} \max_{\beta > 0} \left\{ \beta \cdot \left(n\sqrt{\sigma^2 + \frac{\|\boldsymbol{u}\|_2^2}{n}} + \langle \boldsymbol{h}, \boldsymbol{u} \rangle \right) - \frac{n}{2}\beta^2 + \lambda \|\boldsymbol{u} + \boldsymbol{x}_0\|_1 + \gamma \sum_{j=1}^d \psi(u_j + x_{0,j}, x_{0,j}) \right\} && \text{(9d)}
\end{aligned}$$

In step (9a), we use the variational representation of squared norm

$$\|\boldsymbol{x}\|_2^2 = \max_{\boldsymbol{z}} \langle \boldsymbol{z}, \boldsymbol{x} \rangle - \frac{1}{2}\|\boldsymbol{z}\|_2^2.$$

In step (9b), we apply Theorem 2. Assuming the concentration of the optimal value $\phi(\boldsymbol{g}, \boldsymbol{h})$ of the auxiliary minimax optimization problem, we can deduce the concentration of the optimal value $\Phi(\boldsymbol{G})$ of the original problem. However, there are two issues applying this result: first, the theorem requires compact domain, while minimax optimization problem in Eq (9a) is searching over the entire Euclidean space; second, the function $(\boldsymbol{u}, \boldsymbol{v}) \mapsto \langle \boldsymbol{v}, \boldsymbol{w} \rangle - \frac{1}{2}\|\boldsymbol{v}\|_2^2 + \lambda \|\boldsymbol{u} + \boldsymbol{x}_0\|_1 + \gamma \cdot \sum_{j=1}^d \psi(u_j + x_{0,j}, x_{0,j})$ is required to be convex-concave. The first issue can be resolved by taking a large bounded domain with radius $\omega(\sqrt{d})$. As $d \to +\infty$, the probability that the optimal solution lies outside the domain diminishes. To resolve the second issue, one method is to note that the condition is satisfied when the function ψ is always convex with respect to its first argument when $\gamma > 0$, and concave when $\gamma < 0$. We can then represent a sufficiently smooth univariate function by difference of convex and smooth functions, and obtain the result for arbitrary test function ψ. Another method is via an application of CGMT within a carefully-constructed domain, which is discussed in the next section.

In step (9c), we take the change-of-variable $\boldsymbol{v} = \beta\sqrt{n}\bar{\boldsymbol{v}}$ with $\bar{\boldsymbol{v}} \in \mathbb{S}^{d-1}$, and explicitly optimize the vector $\bar{\boldsymbol{v}}$ on the unit sphere. In step (9d), we make use of concentration inequalities for the norm of Gaussian random vectors, such that $\|\boldsymbol{g}\|_2, \|\boldsymbol{w}\|_2/\sigma \approx \sqrt{n}$ with high probability.

From Eq (9d), we note that the terms involving $\boldsymbol{u}$ are separable along d-dimensions except for the square-root term. Recall that the goal is to relate lasso risk with that of a one-dimensional optimization problem. In order to make it coorindate-wise separable, we make use of the variational representation $\sqrt{x} = \min_{\tau > 0} \left\{ \frac{x}{2\tau} + \frac{\tau}{2} \right\}$, and obtain the following equations:

$$\begin{aligned}
& \text{Eq (9d)} \\
&= \min_{\boldsymbol{u} \in \mathbb{R}^d} \max_{\beta \geq 0} \min_{\tau > 0} \left\{ \beta \cdot \left(\frac{1}{2\tau}\left(n\sigma^2 + \|\boldsymbol{u}\|_2^2\right) + \frac{n\tau}{2} + \langle \boldsymbol{h}, \boldsymbol{u} \rangle \right) - \frac{n}{2}\beta^2 + \lambda \|\boldsymbol{u} + \boldsymbol{x}_0\|_1 + \gamma \sum_{j=1}^d \psi(u_j + x_{0,j}, x_{0,j}) \right\} && \text{(9e)} \\
&= \max_{\beta \geq 0} \min_{\tau > 0} \sum_{j=1}^d \min_{u_j \in \mathbb{R}} \left\{ \left(\frac{\sigma^2}{\tau} + \tau\right)\frac{\beta\delta}{2} - \frac{\delta}{2}\beta^2 + \frac{\beta}{2\tau}u_j^2 + \beta h_j u_j + \lambda |u_j + x_{0,j}| + \gamma \psi(u_j + x_{0,j}, x_{0,j}) \right\}. && \text{(9f)}
\end{aligned}$$

In step (9e), we substitute with the variational representation of square root, and in the step (9f), we use the strong duality result for convex-concave minimax optimization, which, once again, holds true when ψ is convex in its first argument for positive γ (and concave for negative γ).

Using standard concentration inequalities and ε-net arguments, we have the uniform convergence result:

$$\left\{\frac{1}{d}\sum_{j=1}^{d}\min_{u_j}\left(\frac{\beta}{2\tau}u_j^2+\beta h_j u_j+\lambda|u_j+x_{0,j}|+\gamma\psi(u_j+x_{0,j},x_{0,j})\right)\right\}_{(\beta,\tau)}$$
$$\xrightarrow{p}\left\{\mathbb{E}_{(X_0,G)\sim\mathbb{P}_0\times\mathcal{N}(0,1)}\left[\min_{u}\left(\frac{\beta}{2\tau}u^2+\beta G+\lambda|u+X_0|+\gamma\psi(u+X_0,X_0)\right)\right]\right\}_{(\beta,\tau)}.$$

We therefore have the relation

$$\text{Eq (9f)}\approx d\times\max_{\beta\geq 0}\min_{\tau>0}\left\{(\frac{\sigma^2}{\tau}+\tau)\frac{\beta\delta}{2}-\frac{\delta}{2}\beta^2+\mathbb{E}\left[\min_{u}\left(\frac{\beta}{2\tau}u^2+\beta G+\lambda|u+X_0|+\gamma\psi(u+X_0,X_0)\right)\right]\right\}, \tag{9g}$$

for the random variables $(X_0,G)\sim\mathbb{P}_0\times\mathcal{N}(0,1)$. Solving for the optimality condition, we get the self-consistent equations (3a) and (3b) for $(\beta_\star,\tau_\star)$. Using the implicit differentiation theorem, we can then obtain the expression for the empirical average of ψ.

$$\lim_{d\to+\infty}\mathbb{E}\left[\frac{1}{d}\sum_{i=1}^{d}\psi(\widehat{x}_i,x_{0,i})\right]\approx f'(0)\approx\mathbb{E}\left[\eta\big(X_0+\tau_\star G;\frac{\tau_\star\lambda}{\beta_\star}\big)\right]. \tag{10}$$

This concludes the derivation of the asymptotic empirical distribution for the Lasso estimator in Theorem 1

3 How to make it rigorous?

In this section, we discuss the technical components that make the derivation steps in the previous section rigorous. We will also present an alternative strategy that directly prove the Wasserstein convergence result, without going through the perturbed free energy trick.

We first discuss the "≈" steps appearing in the long derivation in the prevoius section.

- To apply CGMT and obtain step (9b), besides the compactness and convexity conditions, one would need the value of the minimax problems to concentrates well around its expectation. Conditionally on the random variables $\boldsymbol{x}_0$ and $\boldsymbol{w}$, for bounded $\boldsymbol{u}$ and $\boldsymbol{v}$, the objective in the minimax problem is a Lipschitz function of (rescaled) Gaussian random matrix $\boldsymbol{A}$. We can then invoke the Borel-TIS lemma to obtain the concentration result.
- In step (9d), we note that for $\boldsymbol{g}\sim\mathcal{N}(0,I_d)$ and $\boldsymbol{w}\sim\mathcal{N}(0,\sigma^2 I_d)$ independent, for given $\boldsymbol{u}$, we have that

  $$\|\boldsymbol{u}\|_2\boldsymbol{g}+\sqrt{n}\cdot\boldsymbol{w}\sim\mathcal{N}\left(0,(n\sigma^2+\|u\|_2^2)I_d\right).$$

  The result then follows from the standard result for concentration of χ^2 random variables.
- The uniform concentration result holds true for (β,τ) in any compact set. However, the optimization is over an unbounded domain. In order to obtain Eq (9g), we also need to apply a truncation argument: showing the uniform concentration in a large compact domain and also show that the minimax equilibrium is achieved in this domain with high probability.

- In order to go from concentration of the function value to that of the derivative in Eq (10), we can apply the same trick as used in Lecture 6. In particular, define the discrete derivatives:

$$\Delta_n^+(\gamma, \varepsilon) = \frac{f(\gamma+\varepsilon) - f(\gamma)}{\varepsilon}, \quad \text{and} \quad \Delta_n^-(\gamma, \varepsilon) = \frac{f(\gamma) - f(\gamma-\varepsilon)}{\varepsilon}.$$

  Since f is a concave function, we have $\Delta_n^+ \leq f' \leq \Delta_n^-$. We can then show the concentration inequalities for Δ_n^+ and Δ_n^- for any finite ε, and then apply monotone convergence theorem.

- The results can be shown for test functions ψ that are convex in its first coordinate. On the other hand, note that any univariate C^2 smooth function is the difference of two convex and smooth ones. The convergence result can therefore be extended to the class of functions with uniformly bounded second derivatives. To further extend it to the class of pseudo-Lipschitz functions, one can start with a function ψ and mollify it with a Gaussian kernel to get a smooth function $\widetilde{\psi} := \psi \star \phi_\varepsilon$, where ϕ_ε is the Gaussian density with variance ε^2. Showing the convergence with test function $\widetilde{\psi}$ and estimating the approximation error leads to a $\mathcal{W}_2$ convergence result (albeit with worse rate of convergence).

Another approach is to use the original free energy but with a carefully-designed domain to apply CGMT. In particular, we let:

$$\Phi(\boldsymbol{u}) := \frac{1}{2d}\|\boldsymbol{A}\boldsymbol{u} + \boldsymbol{w}\|_2^2 + \frac{\lambda}{d}\|\boldsymbol{u} + \boldsymbol{x}_0\|_1, \tag{11a}$$

$$\Gamma(\boldsymbol{u}) := \beta_\star \cdot \left(\frac{1}{2\tau_\star}\big(\delta\sigma^2 + \frac{\|\boldsymbol{u}\|_2^2}{d}\big) + \frac{\tau_\star \delta}{2} + \frac{\langle \boldsymbol{h}, \boldsymbol{u}\rangle}{d}\right) - \frac{\delta}{2}\beta_\star^2 + \frac{\lambda}{d}\|\boldsymbol{u} + \boldsymbol{x}_0\|_1 \tag{11b}$$

Denote by $\Phi_\star$ the asymptotic risk of the lasso problem. Applying Theorem 2 and following the derivation until step (9e), we have:

- For any compact set S, we have:

$$\mathbb{P}\left(\min_{u \in S} \Phi(u) \leq \Phi_\star + \varepsilon\right) \lesssim \mathbb{P}\left(\min_{u \in S} \Gamma(u) \leq \Phi_\star + \varepsilon\right). \tag{12a}$$

- For compact set T, we have:

$$\mathbb{P}\left(|\min_{u \in T} \Phi(u) - \Phi_\star| \geq \varepsilon\right) \lesssim \mathbb{P}\left(|\min_{u \in T} \Gamma(u) - \Phi_\star| \geq \varepsilon\right). \tag{12b}$$

Take T to be a large compact set such that the minimum is contained in T with high probability. Following concentration property of $\min_{\boldsymbol{u}} \Gamma(\boldsymbol{u})$, we conclude that $\min_{\boldsymbol{u}} \Phi(\boldsymbol{u})$ is concentrating around $\Phi_\star$.

For any $\omega > 0$, we define the set:

$$S(\omega) := \mathbb{B}(0, R\sqrt{d}) \setminus \left\{\boldsymbol{u} : \mathcal{W}_2\big(\widehat{\mu}_{(\boldsymbol{x}_0+\boldsymbol{u}, \boldsymbol{x}_0)}, \bar{\mu}_\lambda\big) \leq \omega\right\},$$

where R is a large constant such that the minimum is contained in $\mathbb{B}(0, R\sqrt{d})$ with high probability.

Applying the bound (12a) with set $S = S(\omega)$, we note that $\mathbb{P}\left(\min_{u \in S(\omega)} \Gamma(u) \leq \Phi_\star + \varepsilon\right) \to 0$ implies the limit $\mathbb{P}\left(\min_{u \in S(\omega)} \Phi(u) \leq \Phi_\star + \varepsilon\right) \to 0$. By standard concentration inequalities and local strong convexity of Γ, one can show that the set approximate minima of Γ is attained outside the set $S(\omega)$, with high probability, so that we have:

$$\mathbb{P}\left(\min_{u \in S(\omega)} \Phi(u) \leq \Phi_\star + \varepsilon\right) \to 0.$$

This implies the limiting statement $\mathbb{P}\left(\min_{u \in S(\omega)} \Phi(u) \leq \Phi_\star + \varepsilon\right) \to 0$, which means the minimum is contained within set $S(\omega)$ with high probability.

Above derivation exemplifies the following general proof strategy: in order to show that $\arg\min_{\boldsymbol{u}} \Phi(u) \in E$ where E is a good set (a Wasserstein neighborhood in this case), it suffices to show $\min_{\boldsymbol{u} \in \mathbb{R}^d \setminus E} \Phi(u) \geq \min_{\boldsymbol{u} \in \mathbb{R}^d} \Phi(u) + \varepsilon$. The probability of latter event can typically lower bounded by the probability of an analogous event in a simpler problem using CGMT, leading to the desired non-asymptotic result.

References