Lecture 1: Introduction to the mean field asymptotics

Lecturer: Song Mei *Scriber: Kumar Krishna Agrawal* *Proof reader: Tae Joo Ahn*

In this course, we study the computational and statistical aspects of statistical models in the high dimensional asymptotic limit (the mean-field asymptotics). We will introduce heuristic tools in physics including the replica method and the cavity method. These tools can be made rigorous using approaches including the Gaussian comparison inequality, the leave-one-out analysis, and approximate message passing algorithms. Applications of these methods include the spiked matrix model, the LASSO problem, and the double-descent phenomenon.

1 Motivating example: The LASSO problem

We will get a flavor of the difference between the non-asymptotic theory and the asymptotic theory using the example of LASSO.

Let $\boldsymbol{x}_0 \in \mathbb{R}^d$, $\boldsymbol{A} \in \mathbb{R}^{n\times d}$, $\boldsymbol{w} \in \mathbb{R}^n$ and $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_0 + \boldsymbol{w} \in \mathbb{R}^n$. We consider the case $d \gg n$ but hope that $\boldsymbol{x}_0$ is sparse in some sense (e.g $\boldsymbol{x}_0$ is k-sparse if $\boldsymbol{x}_0$ has k *non-zero* elements). To recover $\boldsymbol{x}_0$ given $\boldsymbol{A}$ and $\boldsymbol{y}$, we solve the following LASSO problem

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \frac{\lambda}{n}\|\boldsymbol{x}\|_1. \tag{1}$$

Figure 1 illustrates loss landscape of linear regression with mean-squared error, where LASSO encourages solutions within some l_1 level set. Our objective is to quantify/bound the normalized mean squared error, $||\hat{x} - \boldsymbol{x}_0||_2^2/||\boldsymbol{x}_0||_2^2$. Note that different papers use different normalization of the LASSO problem. Here the normalization I used is such that the presentation is simpler. When you read a paper on LASSO, you should first look at their normalization and then interpret the results.

1.1 Non-asymptotic theory of LASSO

A line of papers studied the LASSO risk in the non-asymptotic regime. The following result is due to [NRWY12]. Theorem 2 is a fully deterministic statement: the result is satisfied by any deterministic $\boldsymbol{A}$, $\boldsymbol{x}_0$, $\boldsymbol{w}$, and $\boldsymbol{y}$.

Definition 1 (Restricted strong convexity)**.** *We say a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times d}$ satisfies the restricted strong convexity property, if there exists universal constants c_1 and c_2, such that for any $\boldsymbol{v} \in \mathbb{R}^d$, we have*

$$\frac{\|\boldsymbol{A}\boldsymbol{v}\|_2^2}{n} \geq c_1\|\boldsymbol{v}\|_2^2 - c_2\frac{\log d}{n}\|\boldsymbol{v}\|_1^2. \tag{2}$$

Why this property is called restricted strong convexity? If we define $f(\boldsymbol{x}) = (1/2n)\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2$, strong convexity property says that $\nabla^2 f(\boldsymbol{x}) \succeq c_1\mathbf{I}_d$, so that for any direction $\boldsymbol{v}$, we have

$$\frac{\|\boldsymbol{A}\boldsymbol{v}\|_2^2}{n} \geq c_1\|\boldsymbol{v}\|_2^2.$$

Restricted strong convexity simply says that f is strongly convex in the direction $\boldsymbol{v}$ when $\|\boldsymbol{v}\|_1$ is small.

For sensing matrix $\boldsymbol{A}$ that satisfies RSC property, we have the following control of the LASSO risk.

Theorem 2 ([NRWY12]). *For any $\boldsymbol{A} \in \mathbb{R}^{n\times d}$ satisfying the RSC property (2) with constant c_1 and c_2, there exists universal constant $c < \infty$ (depending only on c_1, c_2), such that as long as $\lambda \geq 2\|\boldsymbol{A}^{\mathsf{T}}\boldsymbol{w}\|_\infty$, for any $\boldsymbol{x}_0 \in \mathbb{R}^d$ and $S \subseteq [d]$ with $|S| \leq n/(c \log d)$, the LASSO estimator (1) satisfies*

$$\|\hat{\boldsymbol{x}} - \boldsymbol{x}_0\|_2^2 \leq c\frac{\lambda^2|S|}{n^2} + c\frac{\lambda}{n}\|\boldsymbol{x}_{0,S^c}\|_1 + c\frac{\log d}{n}\|\boldsymbol{x}_{0,S^c}\|_1^2.$$

Theorem 2 does not tell us whether there exists a matrix that satisfies the RSC property. The following proposition tells us that, for Gaussian random matrix $\boldsymbol{A}$, RSC property holds with high probability.

Proposition 3. *For $\boldsymbol{A} \in \mathbb{R}^{n\times d}$ with $A_{ij} \sim \mathcal{N}(0,1)$, Eq. (2) is satisfied for some constant c_1 and c_2 with high probability as $n \to \infty$.*

In the following, we will make simpler assumptions to understand Theorem 2.

Corollary 4. *Let $\boldsymbol{A} \in \mathbb{R}^{n\times d}$ with $A_{ij} \sim \mathcal{N}(0, 1/\|\boldsymbol{x}_0\|_2^2)$. Let $\boldsymbol{x}_0 \in \mathbb{R}^d$ be k-sparse with the support of $\boldsymbol{x}_0$ given by S. Let $\boldsymbol{w}$ be σ^2-sub-Gaussian. Then for any $\delta > 0$, there exists constant $C(\delta)$ such that, as long as we take $n \geq C(\delta)k \log d$ and $\lambda \geq C(\delta) \cdot \sigma\sqrt{n \log d}$, then with probability at least $1 - \delta$, the LASSO estimator (1) satisfies*

$$\frac{\|\hat{\boldsymbol{x}} - \boldsymbol{x}_0\|_2^2}{\|\boldsymbol{x}_0\|_2^2} \leq \frac{C(\delta)\sigma^2 k \log d}{n}.$$

The corollary tells us that, to well-estimate a k-sparse ground truth vector, it is enough to have sample size $n \gg k \log d$.

Remark 5. *In the non-asymptotic setting, everything is explicit, i.e there are no limiting statements. Additionally, the assumptions on the distribution of $\boldsymbol{x}_0$ are quite weak.*

1.2 High dimensional asymptotics of LASSO

Note that the non-asymptotic theory of LASSO does not allow us to consider the proportional regime $n \propto k \propto d$. In many cases, however, this proportional regime is very interesting. It would be desirable to establish a theory to characterize the performance of LASSO in this regime.

Theorem 6 ([BM11]). *We consider the asymptotic limit when $n/d \to \delta \in (0, \infty)$ as $d \to \infty$. Let $\boldsymbol{A} \in \mathbb{R}^{n\times d}$ with $A_{ij} \sim \mathcal{N}(0, 1/n)$. Let $\boldsymbol{x}_0 \in \mathbb{R}^d$ with $\boldsymbol{x}_{0,i} \sim_{iid} \mathbb{P}_0$. Let $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\mathbf{I}_n)$. Let $\hat{\boldsymbol{x}}$ be the LASSO estimator (1). Then we have*

$$\lim_{d,n\to\infty} \frac{1}{d}\|\hat{\boldsymbol{x}} - \boldsymbol{x}_0\|_2^2 = \mathbb{E}_{(X_0,Z)\sim\mathbb{P}_0\times\mathcal{N}(0,1)}[(\eta(X_0 + \tau_\star Z; \theta_\star) - X_0)^2],$$

where $\eta(x) = \text{sign}(x) \cdot (|x| - 1)_+$ is the soft thresholding function and $\tau_\star = \tau_\star(\alpha_\star)$. Here we denote $\tau_\star(\alpha)$ to be a function such that, for fixed α, $\tau_\star(\alpha)$ is the largest solution of

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

and we denote $\alpha_\star$ by the unique non-negative solution of

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

Moreover, for any Lipschitz function ψ, we have almost surely

$$\lim_{d\to\infty} \frac{1}{d}\sum_{i=1}^{d} \psi(\hat{x}_i, x_{0,i}) = \mathbb{E}_{(X_0,Z)\sim\mathbb{P}_0\times\mathcal{N}(0,1)}[\psi(\eta(X_0 + \tau_\star Z; \alpha_\star\tau_\star), X_0)].$$

Remark 7. *The asymptotic error for high-dimensional LASSO estimator is equivalent to*

$$\mathbb{E}_{\hat{X},X_0}[(\hat{X}-X_0)^2],$$

where $(\hat{X}, X_0)$ *following the distribution of*

$$(X_0, Z) \sim \mathbb{P}_0 \times \mathcal{N}(0,1), \qquad Y = X_0 + \tau_\star Z,$$
$$\hat{X} = \arg\min_v \left\{(Y-v)^2 + \tau_\star \alpha_\star |v|\right\} = \eta(Y, \tau_\star \alpha_\star).$$

This can be interpreted as an one dimensional LASSO problem.

We can plot the limiting risk versus the regularization parameter λ, which is given in Figure 2. This curve gives the precise U-shaped curve for the Bias and Variance tradeoff of LASSO estimator. Note that this U-shaped curve cannot be completely captured by the non-asymptotic theory, since the non-asymptotic theory doesn't give lower and upper bounds that match up to $1 + o(1)$. The sharp characterization of the risk is an advantage of the high dimensional asymptotic theory.

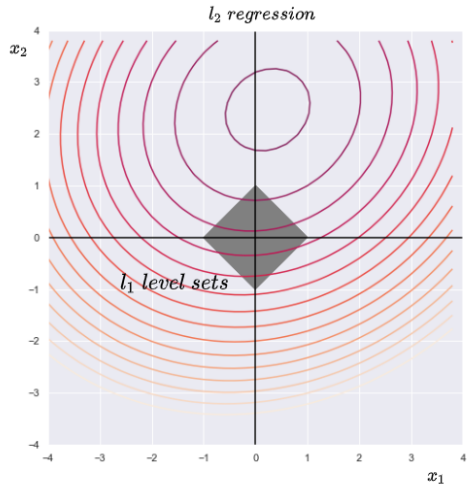

Figure 1: LASSO regularizer encourages sparsity.

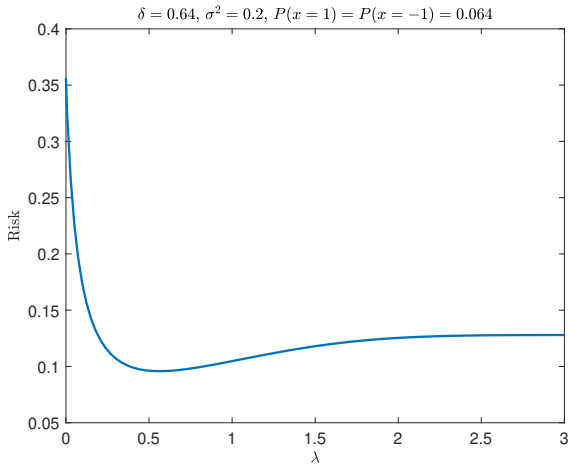

Figure 2: The risk of the LASSO estimator

1.3 Comparison of non-asymptotic theory and high dimensional asympotics

Here we present a table that compares the non-asymptotic theory versus the asymptotic theory.

	Non-asymptotics theory	High dimensional asymptotics
Typical regime	(Relatively) Strong signal-to-noise ratio ($n \gg k \log d$)	Constant signal-to-noise ratio ($n \propto d \propto k$)
Advantages	Less model assumptions. Result holds for any finite parameter size.	Precise asymptotic formula: upper and lower bounds match sharply.
Limitations	A gap of upper and lower bounds up to constant or logarithmic factors.	More detailed model assumptions. (Sometimes) hard to control how large should the parameter be so that the asymptotic regime kick in.
When useful?	Characterize the behavior of a model or an algorithm with general assumptions.	Identify the exact location of phase transition.
Examples	Statistical learning theory: bounding excessive risk by uniform convergence. Analyzing the non-convex landscape of empirical risk minimization.	The phase transition phenomenon for compressed sensing. Understanding the double-descent phenomenon. The optimal loss function in machine learning

2 Mean-field theory and statistical physics

2.1 The mean field theory

The following definition of the mean field theory is adapted from wikipedia.

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

In our example, the LASSO problem is a high dimensional random model, while the one dimensional model in remark 7 is the simpler model that approximates the original one.

2.2 Method from statistical physics

The focus of this course is to analyze statistical models through the high dimensional asymptotic viewpoint. In many cases, we are interested in deriving the asymptotic formula instead of proving the formula rigorously, and statistical physics tools can be used to predict these formula. These predicted formula can be simply verified through experiments. While some predictions have been made rigorous in some way, typically proving these formula is much more complicated than deriving them.

Figure 3 motivates some of the connections between statitical physics and statistical learning. In this course, we will introduce the "replica method" introduced by physicists early in 1970s. We will show how it can be used to predict the behaviors of statistical models and algorithms in the asymptotic limit. Simple models will be used as examples in class: the spiked GOE matrix and the LASSO problem. We will revisit these models several times. We will first show how the replica method can be used to predict the behavior of these models. Then we will show how these predictions can be proved using rigorous tools. These rigorous tools include the Gaussian comparison theorem, the Stieltjes transforms, and approximate message passing (AMP) algorithms.

3 Level of rigorous of this course

In this course, we will sometimes adopt a physics level of rigorous, and sometimes adopt a mathematics level of rigorous. We will not involve in the measure theoretic issues. That being said, we will assume every function is measurable and most of the time integrable. Sometimes we will assume differentiability, assume exchange of limits, and assume exchange of limits and differentiation. We will clarify these heuristic when we did so.

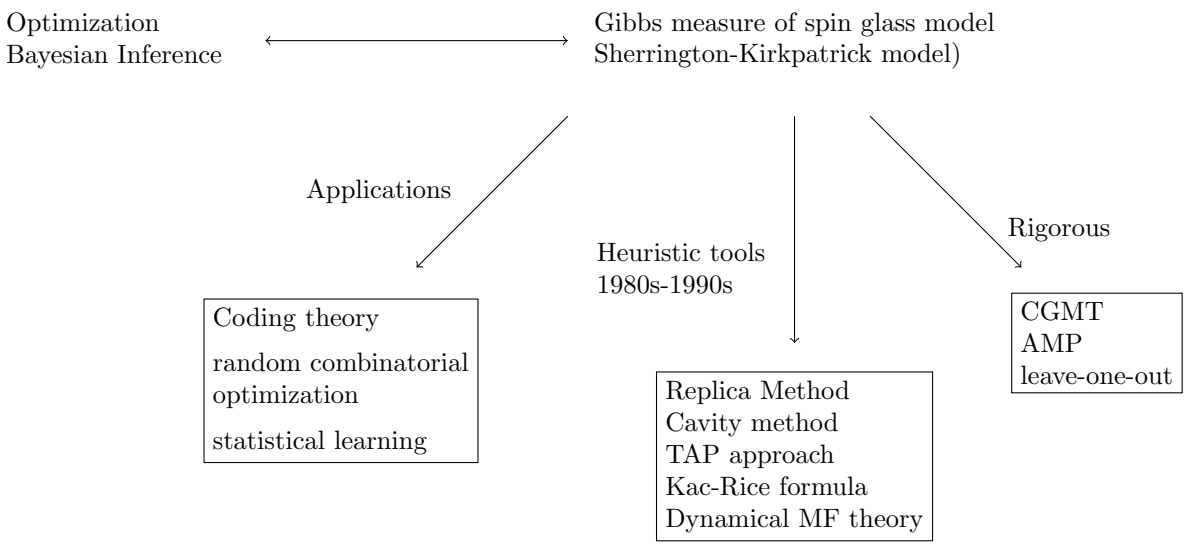

Figure 3: Tools developed in statistical physics with applications to statistical learning.

The reason why we don't adopt the fully rigorous approach is that, it can take a long time to explain every details in checking these exchange of limits assumptions, which may make the audience lose the intuition and the main idea.

References

[BM11] Mohsen Bayati and Andrea Montanari, *The lasso risk for gaussian matrices*, IEEE Transactions on Information Theory **58** (2011), no. 4, 1997–2017.

[NRWY12] Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu, *A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers*, Statistical science **27** (2012), no. 4, 538–557.