STAT260 Mean Field Asymptotics in Statistical Learning	Lecture 6 - 02/08/2021

Lecture 6: Concentration phenomena in mean-field asymptotic

Lecturer: Song Mei *Scriber: Omer Ronen* *Proof reader: Tae Joo Ahn*

1 Concentration phenomena in mean-field asymptotic

In the **non-asymptotic** regime a typical bound is of the form

$$\mathbb{P}\Big(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2/d \geq C\sqrt{\frac{d\log d/\delta}{n}}\Big) \leq \delta. \tag{1}$$

In the **asymptotic** regime the picture is a bit different. Typically the risk $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2/d$ would not go to zero when we consider a regime in which d/n converges to some constant. Under such regime the risk lower bound in (1) would go to ∞ and the risk would not converge to zero, but rather concentrate around it's expectation, formally

1. $\lim_{n\to\infty} \mathbb{P}(|\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2/n - E\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2/n| \geq \epsilon) = 0$,
2. $\lim_{n\to\infty} \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2/n =$ some formula.

In this lecture we will focus on step 1. Step 2 will be covered in later lectures.

2 The Gaussian concentration inequality

Proposition 1 (Gaussian concentration inequality)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be an L-Lipschitz function:*

$$|f(x_1) - f(x_2)| \leq L\|x_1 - x_2\|_2,\ \forall x_1, x_2 \in \mathbb{R}^d,$$

and let $\boldsymbol{G} = (G_1, \ldots, G_d)$ be a vector of standard Gaussian random variables ($G_i \sim_{i.i.d} N(0, 1)$). Then we have

$$\mathbb{P}(|f(\boldsymbol{G}) - \mathbb{E}f(\boldsymbol{G})| \geq t) \leq 2\exp\{-t^2/(2L^2)\}.$$

This proposition can be derived from the Gaussian log Sobolev inequality coupled with the Herbst's argument (c.f. [Thm 3.25] in Ramon van Handel's notes).

3 $\mathbb{Z}_2$ synchronization problem

Let us define the problem, we have our true parameter $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)^\mathsf{T} \in \mathbb{R}^n$ with $\theta_i \sim_{i.i.d}$ Unif($\mathbb{Z}_2 = \{\pm 1\}$). We have the observation vector $\boldsymbol{Y} = \frac{\lambda}{n}\boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T} + \boldsymbol{W} \in \mathbb{R}^{n\times n}$ with $\boldsymbol{W} \sim$ GOE(n). That is $W_{ij} \sim_{i.i.d} N(0, n^{-1})$ for $1 \leq i \leq j \leq n$ and $W_{ii} \sim_{\text{i.i.d}}$ for $1 \leq i \leq n$ with the restriction that $\boldsymbol{W}$ is symmetric ($W_{ij} = W_{ji}$).

Remark 2. $\boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T}$ scaling. *An interesting thing to consider is the choice of scaling. We choose the scaling of n so that the operator norms of $\frac{\lambda}{n}\boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T}$ and $\boldsymbol{W}$ would be on the same scale (we consider to get a non-trivial behavior of the spectral estimator which involves the calculation of the leading eigenvector). Indeed $\|\boldsymbol{W}\|_{op} \overset{2}{\approx} C$ and $\|\boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T}\|_{op} \approx \|\boldsymbol{\theta}\|_2 = n$ hence the choice of scaling is n, such that the signal to noise ration is balances between the two norms.*

We observe $\boldsymbol{Y}$ and try to estimate $\boldsymbol{\theta}$, the spectral estimator is defined to be

$$\hat{\boldsymbol{\theta}}(\boldsymbol{Y}) = \hat{\boldsymbol{\theta}}_{\rm spec}(\boldsymbol{Y}) = \arg\max_{\boldsymbol{\sigma}\in\mathbb{S}^{n-1}(\sqrt{n})} \langle \boldsymbol{\sigma}, \boldsymbol{Y}\boldsymbol{\sigma}\rangle/n$$

Our first goal is to show $\max_{\boldsymbol{\sigma}\in\mathbb{S}^{n-1}(\sqrt{n})} \langle \boldsymbol{\sigma}, \boldsymbol{Y}\boldsymbol{\sigma}\rangle/n \approx$ it's expectation (w.h.p), which follows from the proposition below.

Proposition 3. *Let $\mathcal{R}_n \subseteq \{\boldsymbol{\theta}\in\mathbb{R}^n : \|\boldsymbol{\theta}\|_2^2 n \le 1\}$ and let $\boldsymbol{Y} = \boldsymbol{A} + \boldsymbol{W} \in \mathbb{R}^{n\times n}$, where $\boldsymbol{A}$ is deterministic and $\boldsymbol{W} \sim GOE(n)$*
Then $\delta\rangle 0$, we have

$$\mathbb{P}_{\boldsymbol{W}}\Big(|\sup_{\boldsymbol{\sigma}\in\mathcal{R}_n}\langle \boldsymbol{\sigma}, \boldsymbol{Y}\boldsymbol{\sigma}\rangle/n - \mathbb{E}[\sup_{\boldsymbol{\sigma}\in\mathcal{R}_n}\langle \boldsymbol{\sigma}, \boldsymbol{Y}\boldsymbol{\sigma}\rangle/n]| \le \sqrt{\frac{4\log(2/\lambda)}{n}}\Big) \ge 1-\delta$$

We will now use the Gaussian inequality to prove the above mentioned proposition

Proof. Let $\boldsymbol{G} \in \mathbb{R}n\times n$ with $G_{ij} \sim_{i.i.d} N(0,1)$ for $1 \le i,j \le n$ (which is not symmetric).
Denote $\tilde{\boldsymbol{W}} = (\boldsymbol{G}+\boldsymbol{G}^\mathsf{T})/\sqrt{2n}$ then we have $\tilde{\boldsymbol{W}} \stackrel{d}{=} \boldsymbol{W}$.
Define $\tilde{\boldsymbol{Y}} = \boldsymbol{A} + \tilde{\boldsymbol{W}}$ and

$$\begin{aligned} f(G) &= \sup_{\boldsymbol{\sigma}\in\mathcal{R}_n}\langle \boldsymbol{\sigma}, \tilde{\boldsymbol{Y}}\boldsymbol{\sigma}\rangle/n \\ &= \sup_{\boldsymbol{\sigma}\in\mathcal{R}_n}\langle \boldsymbol{\sigma}, (G+G^\mathsf{T})/\sqrt{2n}\boldsymbol{\sigma}\rangle/n \end{aligned}$$

We would like to show that f is a L-Lipschitz function and then apply the inequality.
Let $\boldsymbol{G}_1, \boldsymbol{G}_2 \in \mathbb{R}^{n\times n}$ and let $\boldsymbol{\sigma}^\star = \arg\sup_{\boldsymbol{\sigma}\in\mathcal{R}_n}\langle \boldsymbol{\sigma}, \tilde{\boldsymbol{Y}}_1\boldsymbol{\sigma}\rangle/n$

$$\begin{aligned} f(G_1) - f(G_2) &= \sup_{\boldsymbol{\sigma}\in\mathcal{R}_n}\langle \boldsymbol{\sigma}, \tilde{\boldsymbol{Y}}_1\boldsymbol{\sigma}\rangle/n + \inf_{\boldsymbol{\sigma}\in\mathcal{R}_n} -\langle \boldsymbol{\sigma}, \tilde{\boldsymbol{Y}}_2\boldsymbol{\sigma}\rangle/n \\ &\le \langle \boldsymbol{\sigma}^\star, \tilde{\boldsymbol{Y}}_1\boldsymbol{\sigma}^\star\rangle/n + -\langle \boldsymbol{\sigma}^\star, \tilde{\boldsymbol{Y}}_2\boldsymbol{\sigma}^\star\rangle/n \\ &= \langle \boldsymbol{G}_1 - \boldsymbol{G}_2, \boldsymbol{\sigma}^\star(\boldsymbol{\sigma}^\star)^\mathsf{T}\rangle\sqrt{\frac{2}{n}}/n \\ &\le \|\boldsymbol{G}_1 - \boldsymbol{G}_2\|_{\rm op}\underbrace{\|\boldsymbol{\sigma}^\star\|_2^2/n}_{\le 1}\sqrt{\frac{2}{n}} \\ &\le \|\boldsymbol{G}_1 - \boldsymbol{G}_2\|_{\rm F}\sqrt{\frac{2}{n}} \end{aligned}$$

Which means that f is $\sqrt{\frac{2}{n}}$−Lipschitz

Remark 4. Differential f proof sketch *To get an intuition for this proof we can consider the case where f is differentiable. We use the implicit differentiation theorem to calculate the gradient of f and evaluate it at $\boldsymbol{\sigma}^\star = \arg\sup_{\boldsymbol{\sigma}\in\mathcal{R}_n}$, indeed*

$$\|\nabla_{\boldsymbol{G}} f(\boldsymbol{G})\|_F = \sqrt{\frac{2}{n}}\|\frac{\boldsymbol{\sigma}^\star(\boldsymbol{\sigma}^\star)^\mathsf{T}}{n}\|_F = \sqrt{\frac{2}{n}}\|\boldsymbol{\sigma}^\star\|_2^2/n = \frac{2}{n}$$

Which means f is $\frac{2}{n}$−Lipschitz

We now use proposition 1 with $L = \frac{2}{n}$ and get

$$\mathbb{P}(|f(\boldsymbol{G}) - \mathbb{E}f(\boldsymbol{G})| \ge t) \le 2\exp\{-(nt^2/4)\}$$

we solve

$$2\exp\{-nt^2/4\} = \delta$$

and conclude our proof with $t = \sqrt{\frac{4\log(2/\delta)}{n}}$ $\square$

The next step is to show the concentration of $\langle \hat{\boldsymbol{\theta}}, \boldsymbol{\theta}\rangle^2/2$ with $\hat{\boldsymbol{\theta}} = \arg\sup_{\boldsymbol{\sigma}\in\mathbb{S}^{n-1}(\sqrt{n})} \langle \boldsymbol{\sigma}, \boldsymbol{Y}\boldsymbol{\sigma}\rangle/n$. We first state a result to be used:

Remark 5. BBP phase transition *We define* $U_n(\lambda) = \sup_{\boldsymbol{\sigma}\in\mathbb{S}^{n-1}(\sqrt{n})}\langle \boldsymbol{\sigma}, \boldsymbol{Y}\boldsymbol{\sigma}\rangle/n$.
We have that

$$\lim_{n\to\infty} \mathbb{E}U_n(\lambda) = \begin{cases} 2 & ,\lambda \le 1 \\ \lambda + \frac{1}{\lambda} & ,\lambda \rangle 1 \end{cases}$$

Using implicit differentiation,

$$\begin{aligned} \frac{d}{d\lambda}U_n(\lambda) &= \frac{d}{d\lambda}\sup_{\boldsymbol{\sigma}\in\mathbb{S}}[\frac{\lambda}{n^2}\langle \boldsymbol{\sigma}, \boldsymbol{\theta}\rangle^2 + \langle \boldsymbol{\sigma}, \boldsymbol{W}\boldsymbol{\sigma}\rangle/2] \\ &= \frac{1}{n^2}\langle \hat{\boldsymbol{\theta}}, \boldsymbol{\theta}\rangle^2 \end{aligned}$$

We now give a heuristic proof, for some intuition, we have shown (U is the same function as f)

$$\lim_{n\to\infty} \mathbb{P}(|U_n(\lambda) - \mathbb{E}U_n(\lambda)| \ge \epsilon) = 0$$

and we calculate ($\approx$) the expectation

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n}\langle \boldsymbol{\theta}, \hat{\boldsymbol{\theta}}\rangle^2 &\approx \partial\lambda \lim_{n\to\infty} \mathbb{E}[U_n(\lambda)] \\ &= \partial\lambda \begin{cases} 2 & ,\lambda \le 1 \\ \lambda + \frac{1}{\lambda} & ,\lambda\rangle 1 \end{cases} \\ &= \begin{cases} 0 & ,\lambda \le 1 \\ 1 - \frac{1}{\lambda^2} & ,\lambda\rangle 1 \end{cases} \end{aligned}$$

We know show the deviation a bit more carefully, we define the "discrete differential of size δ"

$$\begin{aligned} \Delta_n^+(\lambda, \delta) &= \frac{U_n(\lambda+\delta) - U_n(\lambda)}{\delta} \\ \Delta_n^-(\lambda, \delta) &= \frac{U_n(\lambda) - U_n(\lambda-\delta)}{\delta} \end{aligned}$$

Indeed

$$\begin{aligned} U_n(\lambda+\delta) &= \sup_{\boldsymbol{\sigma}\in\mathbb{S}}\langle \boldsymbol{\sigma}, \boldsymbol{W}\boldsymbol{\sigma}\rangle/n + \frac{\lambda+\delta}{n^2}\langle \boldsymbol{\sigma}, \boldsymbol{\theta}\rangle^2 \\ &\ge \underbrace{\frac{\langle \boldsymbol{\sigma}^\star, \boldsymbol{W}\boldsymbol{\sigma}^\star\rangle}{n} + \frac{\lambda}{n^2}\langle \boldsymbol{\sigma}^\star, \boldsymbol{\theta}\rangle}_{U_n(\lambda)} + \frac{\delta}{n^2}\langle \boldsymbol{\sigma}^\star, \boldsymbol{\theta}\rangle^2 \\ &= U_n(\lambda) + \frac{\delta}{n^2}\langle \hat{\boldsymbol{\theta}}, \boldsymbol{\theta}\rangle^2 \end{aligned}$$

We conclude (by symmetry) the following inequality

$$\Delta_n^-(\lambda, \delta) \le \frac{\langle \boldsymbol{\theta}, \hat{\boldsymbol{\theta}}\rangle^2}{n^2} \le \Delta_n^+(\lambda, \delta) \tag{2}$$

Using BBP phase transition result we get

$$\Delta^+(\lambda,\delta) = \frac{u(\lambda+\delta)-u(\lambda)}{\delta}$$
$$\Delta^-(\lambda,\delta) = \frac{u(\lambda)-u(\lambda-\delta)}{\delta}$$

for

$$u(\lambda) = \begin{cases} 2 & \lambda \le 1 \\ \lambda + \frac{1}{\lambda} & \lambda > 1 \end{cases}$$

Next we apply the Gaussian concentration inequality

$$\lim_{n\to\infty} \mathbb{P}(|\Delta_n^{\pm} - \Delta^{\pm}| \ge \epsilon) = 0 \tag{3}$$

Combining (3) with (2) we get

$$\lim_{n\to\infty} \mathbb{P}(\Delta^-(\lambda,\delta) - \epsilon \le \frac{\langle \boldsymbol{\theta}, \hat{\boldsymbol{\theta}} \rangle^2}{n^2} \le \Delta^+(\lambda,\delta) + \epsilon) = 1 \tag{4}$$

By the definition of the derivative

$$\lim_{\delta\to 0} \Delta^{\pm}(\delta,\lambda) = \begin{cases} 0 & \lambda \le 1 \\ 1 - \frac{1}{\lambda^2} & \lambda > 1 \end{cases} \tag{5}$$

And finally combining (5) and (4) we get

$$\lim_{n\to\infty} \mathbb{P}(|\langle \hat{\boldsymbol{\theta}}, \boldsymbol{\theta} \rangle^2 / n^2 - \Delta(\lambda)| \ge \epsilon) = 0$$

□

Remark 6. General recipe to show the concentration of any $M(\boldsymbol{\sigma})$

1. *Define the appropriate perturbed Hamiltonian* $H_h(\boldsymbol{\sigma}) = \langle \boldsymbol{\sigma}, \boldsymbol{Y}\boldsymbol{\sigma} \rangle / n + hM(\boldsymbol{\sigma})$
2. *Show the concentration of* $\sup_{\boldsymbol{\sigma}} H_h(\boldsymbol{\sigma})$ *at some* h *near* 0 *with* $H_h(\boldsymbol{\sigma})$ *differentiable with respect to* h *at* $h = 0$
3. *Obtain the result by taking derivative with respect to* h

4 Concentration of Lasso training loss

Let us consider the Lasso problem - we have a random matrix $\boldsymbol{A} \in \mathbb{R}^{n\times d}, A_{ij} \sim N(0,1)$. Our estimand is $x_0 \in \mathbb{R}^d$ with $\|x_0\|_2^2/d \le M$. The observation vector is

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_0 + \boldsymbol{\varepsilon}$$

with $\epsilon \sim N(\mathbf{0}_n, \tau^2 \mathbf{I}_n)$.
We wish to consider two lasso estimators

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

$\hat{\boldsymbol{x}_2}$ is the familiar Lasso problem while $\hat{\boldsymbol{x}_1}$ is called "square root" lasso. We first show the concentration of $\hat{\boldsymbol{x}_1}$

Proposition 7. *Let $\Omega \subseteq \mathbb{R}^d$ be a compact region and consider*

$$\sup_{x\in\Omega}\{\|\boldsymbol{x}\|_2^2/d\} \le D$$

this is a compact set of radius $\sqrt{D}$ times a constant. We define

$$^{1} f_\Omega(\boldsymbol{A},\epsilon) = \min_{x\in\Omega}\frac{1}{\sqrt{n}}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2 + \frac{\lambda}{d}\|\boldsymbol{x}\|_1$$

Then $\exists K<\infty$ s.t $\forall\delta>0$ then

$$\mathbb{P}\Big(|f_\Omega(\boldsymbol{A},\epsilon)-\mathbb{E}f_\Omega(\boldsymbol{A},\epsilon)|\big) \ge \underbrace{K\big(\sqrt{\frac{d}{n}(M+D)}+\tau\big)}_{constant}\sqrt{\frac{\log(2/\delta)}{n}}\Big) \le \delta$$

Proof. We define the "standardized" version of $\boldsymbol{A}$ and ϵ, $\overline{\boldsymbol{A}} = \sqrt{n}\boldsymbol{A}, \bar{\epsilon} = \epsilon/\tau$. Let

$$F(\overline{\boldsymbol{A}},\bar{\epsilon}) = f_\Omega(\overline{\boldsymbol{A}}/\sqrt{n},\bar{\epsilon}\cdot\tau) = f_\Omega(\boldsymbol{A},\epsilon)$$

we want to show that F is Lipschitz and apply the Gaussian concentration inequality. By definition

$$\begin{aligned}
F(\overline{\boldsymbol{A}},\bar{\epsilon}) &= \min_{\boldsymbol{x}\in\Omega}\frac{1}{\sqrt{n}}\|\boldsymbol{A}(\boldsymbol{x}_0-\boldsymbol{x})+\epsilon\|_2 + \frac{\lambda}{n}\|\boldsymbol{x}\|_1 \\
&= \min_{\boldsymbol{x}\in\Omega}\sup_{\boldsymbol{\nu};\|\boldsymbol{\nu}\|_2\le 1}\underbrace{\frac{1}{n}\langle\boldsymbol{\nu},\overline{\boldsymbol{A}}(\boldsymbol{x}_0-\boldsymbol{x})\rangle + \frac{\tau}{\sqrt{n}}\langle\boldsymbol{\nu},\bar{\epsilon}\rangle + \frac{\lambda}{n}\|\boldsymbol{x}\|_1}_{L(\overline{\boldsymbol{A}},\bar{\epsilon},\boldsymbol{x},\boldsymbol{\nu})}
\end{aligned}$$

Now let us consider the difference $F(\overline{\boldsymbol{A}}_1,\bar{\epsilon}_1)-F(\overline{\boldsymbol{A}}_2,\bar{\epsilon}_2)$ and denote $\boldsymbol{\nu}_j^\star,\boldsymbol{x}_j^\star$ the argmax's for the optimization problem F solves with respect to $\overline{\boldsymbol{A}}_j,\bar{\epsilon}_j$.

$$\begin{aligned}
F(\overline{\boldsymbol{A}}_1,\bar{\epsilon}_1)-F(\overline{\boldsymbol{A}}_2,\bar{\epsilon}_2) &= \min_{\boldsymbol{x}\in\Omega}\sup_{\boldsymbol{\nu};\|\boldsymbol{\nu}\|_2\le1} L(\overline{\boldsymbol{A}}_1,\bar{\epsilon}_1,\boldsymbol{x},\boldsymbol{\nu}) + \max_{\boldsymbol{x}\in\Omega}\inf_{\boldsymbol{\nu};\|\boldsymbol{\nu}\|_2\le1} -L(\overline{\boldsymbol{A}}_2,\bar{\epsilon}_2,\boldsymbol{x},\boldsymbol{\nu}) \\
&\le \sup_{\boldsymbol{\nu};\|\boldsymbol{\nu}\|_2\le1} L(\overline{\boldsymbol{A}}_1,\bar{\epsilon}_1,\boldsymbol{x}_2^\star,\boldsymbol{\nu}) + \inf_{\boldsymbol{\nu};\|\boldsymbol{\nu}\|_2\le1} -L(\overline{\boldsymbol{A}}_2,\bar{\epsilon}_2,\boldsymbol{x}_2^\star,\boldsymbol{\nu}) \\
&\le L(\overline{\boldsymbol{A}}_1,\bar{\epsilon}_1,\boldsymbol{x}_2^\star,\boldsymbol{\nu}_1^\star) - L(\overline{\boldsymbol{A}}_2,\bar{\epsilon}_2,\boldsymbol{x}_2^\star,\boldsymbol{\nu}_1^\star) \\
&= \frac{1}{n}\langle\boldsymbol{\nu}_1^\star,(\overline{\boldsymbol{A}}_2-\overline{\boldsymbol{A}}_2)(\boldsymbol{x}_0-\boldsymbol{x}_2^\star))\rangle + \frac{\tau}{\sqrt{n}}\langle\boldsymbol{\nu}_1^\star,\bar{\epsilon}_1-\bar{\epsilon}_2\rangle \\
&\overset{\text{C.S}}{\le} \frac{1}{n}\|\overline{\boldsymbol{A}}_1-\overline{\boldsymbol{A}}_2\|_{\text{op}}\underbrace{\|\boldsymbol{\nu}_1^\star\|_2}_{\le1}\underbrace{\|\boldsymbol{x}_0-\boldsymbol{x}_2^\star\|_2}_{\le\sqrt{dM}\propto\sqrt{nM}} + \frac{\tau}{\sqrt{n}}\|\bar{\epsilon}_1-\bar{\epsilon}_2\|\underbrace{\|\boldsymbol{\nu}_1^\star\|_2}_{\le1} \\
&\le C(\|\overline{\boldsymbol{A}}_1-\overline{\boldsymbol{A}}_2\|_{\text{F}} + \|\bar{\epsilon}_1-\bar{\epsilon}_2\|_2)/\sqrt{n} \\
&\le C'\sqrt{(\|\overline{\boldsymbol{A}}_1-\overline{\boldsymbol{A}}_2\|_{\text{F}}^2 + \|\bar{\epsilon}_1-\bar{\epsilon}_2\|_2^2)/n}
\end{aligned}$$

Now let $G(\overline{\boldsymbol{A}},\bar{\epsilon})\in\mathbb{R}^{nd\times n}$ we have that F is $\frac{C'}{\sqrt{n}}$−Lipschitz in G ($\|G\|_2^2 = \|\overline{\boldsymbol{A}}_1-\overline{\boldsymbol{A}}_2\|_{\text{F}}^2 + \|\bar{\epsilon}_1-\bar{\epsilon}_2\|_2^2$).
We obtain the conclusion by applying the Gaussian inequality. □

[1] f_Ω is the square root lasso problem, with the minimization taken over a constrained set Ω. In future lectures we will show that w.h.p the square root lasso solution lies within Ω .