STAT260 Mean Field Asymptotics in Statistical Learning Lecture 7 - 02/10/2021

Lecture 7: Concentration Inequalities and Field theoretic calculations

Lecturer: Song Mei *Scriber: Mengqi Lin* *Proof reader: Tae joo Ahn*

In this lecture, we will finish the discussion of concentration inequalities, which includes the general recipe of showing the concentration of ensemble average of an observable and another example of its application: Trace of resolvent of Wishart Matrix. And then talk about field theoretic calculations, which is the prerequisites of Replica Methods, and an application example of Large deviation of overlap matrix will be discussed as well.

1 Concentration Inequalities

1.1 General Recipe of showing the concentration of ensemble average of an observable

Suppose we have a perturbed Hamiltonian $H_\lambda(\sigma) = H_0(\sigma) + \lambda M(\sigma)$, where $H_0(\sigma)$ is a random Hamiltonian, e.g $H_0(\sigma) = \langle \sigma, \boldsymbol{W}\sigma\rangle$, and $M(\sigma)$ is the observable, e.g $M(\sigma) = \langle \sigma, \theta\rangle^2/n$, the ensemble average $\langle M\rangle_{\beta_*,\lambda_*}/n$ is the quantity of interest:

$$\langle M\rangle_{\beta,\lambda} \equiv \int_\Omega M(\boldsymbol{\sigma})\mathbb{P}_{\beta,\lambda}(d\sigma), \quad \mathbb{P}_{\beta,\lambda}(d\sigma) \propto \exp\{-\beta H_\lambda(\sigma)\}$$

Before calculating the limits of this ensemble average, we would like to firstly calculate the concentration of it. The general steps are as follows:

Step 1: Showing the concentration of Free energy

Define the normalized free energy

$$F_n(\beta,\lambda)/n = -\frac{1}{n\beta}\log\int_\Omega \exp\{-\beta H_\lambda(\sigma)\}$$

Let $U_n(\lambda) \equiv F_n(\beta_*,\lambda)/n$, we would first to show the concentration of $U_n(\lambda)$:

$$\lim_{n\to\infty}\mathbb{P}(|U_n(\lambda) - \mathbb{E}[U_n(\lambda)]| \ge \epsilon) = 0 \tag{1}$$

If the Hamiltonian involves Gaussian randomness, we can apply Gaussian concentration inequality to show this.

Step 2: Derivative of the limiting free energy

Assume we are able to calculate the limiting free energy :

$$\lim_{n\to\infty}\mathbb{E}[U_n(\lambda)] = u(\lambda), \quad \forall\lambda$$

and $u(\lambda)$ is differentiable at particular λ_*.

Define the discrete derivative of the limiting free energy:

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

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

we have

$$\lim_{\delta\to 0+} \Delta^+(\lambda_*, \delta) = \lim_{\delta\to 0+} \Delta^-(\lambda_*, \delta) = u'(\lambda_*) \tag{2}$$

Note this step is completely by assumption.

Step 3: Derivative of the prelimit free energy
Note $\Delta_n(\lambda) \equiv U_n'(\lambda) = \langle M\rangle_{\beta_*,\lambda}/n$. Define the discrete derivative of the prelimit free energy:

$$\Delta_n^+(\lambda, \delta) \equiv \frac{U_n(\lambda+\delta) - U_n(\lambda)}{\delta}$$

$$\Delta_n^-(\lambda, \delta) \equiv \frac{U_n(\lambda) - U_n(\lambda-\delta)}{\delta}$$

Note $U_n''(\lambda) = -cVar_{\beta_*,\lambda}(M) \le 0$ is concave, so we have:

$$\Delta_n^+(\lambda, \delta) \le \Delta_n(\lambda) \le \Delta_n^-(\lambda, \delta) \tag{3}$$

Step 4: Bridge the above 3 steps

$$\Delta_n^+(\lambda_*, \delta) \xrightarrow[n\to\infty]{(1)} \Delta^+(\lambda_*, \delta) \xrightarrow[\delta\to 0]{(2)} u'(\lambda_*)$$

$$\Delta_n^-(\lambda_*, \delta) \xrightarrow[n\to\infty]{(1)} \Delta^-(\lambda_*, \delta) \xrightarrow[\delta\to 0]{(2)} u'(\lambda_*)$$

$$\Rightarrow \Delta_n(\lambda_*) \xrightarrow[n\to\infty]{(3)} u'(\lambda_*)$$

where $\Delta_n(\lambda_*) = \langle M\rangle_{\beta_*,\lambda_*}$ which is what we want to show : the concentration of ensemble average of the observable.

Remark 1. *If we directly apply Gaussian concentration inequality on $\langle M\rangle_{\beta_*,\lambda_*}$, we cannot get a tight concentration bound. That's why we need to introduce the free engergy function.*

1.2 Example: Concentration of Trace of resolvent of Wishart matrix(Stieltjes transform)

Let $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ with $X_{ij} \sim_{i.i.d.} N(0, 1/d)$. Let $S(\lambda) = \mathrm{tr}[(\boldsymbol{X}^\mathsf{T}\boldsymbol{X} + \lambda \boldsymbol{I}_d)^{-1}]/d,\ (\lambda > 0),\ n/d \to \gamma$

Proposition For any $\delta > 0$, we have:

$$\mathbb{P}\Big(|S(\lambda) - \mathbb{E}[S(\lambda)]| \le \sqrt{\frac{c\log(2/\delta)}{\lambda^3 d^2}}\Big) \ge 1 - \delta$$

Note: the bound rate is $O(1/d)$, which is different from the previous examples of rate $O(1/\sqrt{d})$.

Proof. Goal: To show S is 1/d-Lipschitz in Gaussian r.v.
Let $\boldsymbol{G} = \sqrt{d}\boldsymbol{X} \in \mathbb{R}^{n\times d}$ with $G_{ij} \sim_{i.i.d.} N(0, 1)$

$$\begin{aligned}\bar{S}(\boldsymbol{G}) &= \mathrm{tr}[(\lambda \boldsymbol{I}_d + \boldsymbol{G}^\mathsf{T}\boldsymbol{G}/d)^{-1}]/d \\ &= \mathrm{tr}[(\lambda d \boldsymbol{I}_d + \boldsymbol{G}^\mathsf{T}\boldsymbol{G})^{-1}]\end{aligned}$$

Use the fact that taking derivative and taking trace can be exchanged, $\frac{\partial \boldsymbol{A}(t)^{-1}}{\partial t} = -\boldsymbol{A}(t)^{-1}\frac{\partial \boldsymbol{A}(t)}{\partial t}\boldsymbol{A}(t)^{-1}$ and $\mathrm{tr}(\boldsymbol{AB}) = \mathrm{tr}(\boldsymbol{BA})$:

$$\begin{aligned}
\partial_{G_{ij}} \bar{S}(\boldsymbol{G}) = & -2\mathrm{tr}[(\lambda d \boldsymbol{I}_d + \boldsymbol{G}^\top \boldsymbol{G})^{-1} \boldsymbol{G}^\top \boldsymbol{E}_{ij} (\lambda d \boldsymbol{I}_d + \boldsymbol{G}^\top \boldsymbol{G})^{-1}] \\
= & -2\mathrm{tr}[(\lambda d \boldsymbol{I}_d + \boldsymbol{G}^\top \boldsymbol{G})^{-2} \boldsymbol{G}^\top \boldsymbol{E}_{ij}] \\
= & -2\langle \boldsymbol{G}(\lambda d \boldsymbol{I}_d + \boldsymbol{G}^\top \boldsymbol{G})^{-2}), \boldsymbol{E}_{ij} \rangle
\end{aligned}$$

Hence, in terms of $\boldsymbol{G}$:

$$\nabla_{\boldsymbol{G}} \bar{S}(\boldsymbol{G}) = -2\boldsymbol{G}(\lambda d \boldsymbol{I}_d + \boldsymbol{G}^\top \boldsymbol{G})^{-2})$$

SVD decompose matrix $\boldsymbol{G} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ and Calculating the Lipschitz constant:

$$\begin{aligned}
L^2 &= \sup_{\boldsymbol{G}} \|2\boldsymbol{G}(\lambda d \boldsymbol{I}_d + \boldsymbol{G}^\top \boldsymbol{G})^{-2})\|_F^2 && (4) \\
&= \sup_{\boldsymbol{G}} \|2\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top(\lambda d \boldsymbol{I}_d + \boldsymbol{V}\boldsymbol{\Sigma}^2 \boldsymbol{V}^\top)^{-2})\|_F^2 && (5) \\
&= \sup_{\boldsymbol{G}} \|2\boldsymbol{\Sigma}(\lambda d \boldsymbol{I}_d + \boldsymbol{\Sigma}^2)^{-2})\|_F^2 && (6) \\
&= 4 \sup_{\boldsymbol{G}} \sum_{i=1}^{n} \frac{\sigma_i(\boldsymbol{G})^2}{(\lambda d + \sigma_i(\boldsymbol{G})^2)^4} && (7) \\
&\leq 4 \sum_{i=1}^{n} \frac{c\lambda d}{(\lambda d + c\lambda d)^4} && (8) \\
&= c_2 \frac{n}{\lambda^3 d^3} && (9)
\end{aligned}$$

where (8) follows from the fact that (7) maximizes when $\sigma_i(\boldsymbol{G})^2 = c\lambda d$ for some constant c. □

2 Field Theoretic Calculation

We can think of this as Heuristic Physics Calculation. This calculation is also the basic for replica method. Quantity of interest:

$$Z_n = \int_\Omega \exp\{-\beta H_\lambda(\sigma)\} d\sigma$$

$$\lim_{n\to\infty} \mathbb{E}\Big[\frac{1}{n}\log Z_n\Big] \tag{10}$$

2.1 Delta Function $\boldsymbol{\delta}$(x-a)

Mathematically:

- $\mu(\{a\}) = 1,\ \mu(\{\mathbb{R}/\{a\}\}) = 0$
- $\delta(x-a) = \lim_{\sigma\to 0} \phi_\sigma(x-a)$, where $\phi_\sigma(x-a)$ is the Gaussian density with mean a and variance σ^2.
  $\int f(x)\delta(x-a)dx = \lim_{\sigma\to 0} \int f(x)\phi_\sigma(x-a)dx$, for some test function f.

Physically:

- $$f(a) = \int_R f(x)\delta(x-a)dx \tag{11}$$

- Delta identity formula

$$a \in \mathbb{R},\ \delta(x-a) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{ip(x-a)}dp \tag{12}$$

$$\boldsymbol{a} \in \mathbb{R}^d,\ \delta(\boldsymbol{x}-\boldsymbol{a}) = \frac{1}{(2\pi)^d}\int_{-\infty}^{+\infty} e^{i\langle \boldsymbol{p},\boldsymbol{x}-\boldsymbol{a}\rangle}d\boldsymbol{p} \tag{13}$$

  intuition:
  We know from the Fourier transform:

$$\begin{aligned} f(x) &= \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{ipx}\Big(\int_{-\infty}^{+\infty} e^{-i\alpha p}f(\alpha)d\alpha\Big)dp \\ &= \int_{-\infty}^{+\infty} \underline{\Big(\frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{ip(x-\alpha)}dp\Big)} f(\alpha)d\alpha \end{aligned}$$

  And by (11), we know:

$$f(x) = \int_{-\infty}^{+\infty} f(\alpha)\underline{\delta(x-\alpha)}d\alpha$$

- $1 = \int_R \delta(x-a)dx$

2.2 Gaussian Identity Formula

We begin with some quantity in the LHS of the following formulas, and would try to derive it into an integration of some Gaussian randomness.

$$\exp\{\|\boldsymbol{x}-\boldsymbol{a}\|^2_{\boldsymbol{A}}/2\} = \det(\boldsymbol{A})^{-\frac{1}{2}}\int \frac{1}{(\sqrt{2\pi})^d}\exp\{\langle \boldsymbol{p},\boldsymbol{x}-\boldsymbol{a}\rangle - \frac{1}{2}\|\boldsymbol{p}\|^2_{\boldsymbol{A}^{-1}}\}d\boldsymbol{p} \tag{14}$$

$$\exp\{-\|\boldsymbol{x}-\boldsymbol{a}\|^2_{\boldsymbol{A}}/2\} = \det(\boldsymbol{A})^{-\frac{1}{2}}\int \frac{1}{(\sqrt{2\pi})^d}\exp\{i\langle \boldsymbol{p},\boldsymbol{x}-\boldsymbol{a}\rangle - \frac{1}{2}\|\boldsymbol{p}\|^2_{\boldsymbol{A}^{-1}}\}d\boldsymbol{p} \tag{15}$$

(14) is derived from Gaussian moment generating function , (15) is derived from the Gaussian characteristic formula. To remember the two formula, we can use the fact that:

$$\int \frac{1}{(2\pi)^{d/2}\det(\boldsymbol{A})^{1/2}}\exp\{-(\boldsymbol{x}-\boldsymbol{a}-\boldsymbol{A}\boldsymbol{p})^\top\boldsymbol{A}^{-1}(\boldsymbol{x}-\boldsymbol{a}-\boldsymbol{A}\boldsymbol{p})/2\}d\boldsymbol{p} = 1$$

2.3 Laplace Method

- If $f_n(\boldsymbol{\lambda}) \to f(\boldsymbol{\lambda})$ as $n\to\infty$, then $\int_{\mathbb{R}^k}\exp\{nf_n(\boldsymbol{\lambda})\}d\boldsymbol{\lambda} \doteq \sup_{\boldsymbol{\lambda}\in\mathbb{R}^k}\exp\{nf_n(\boldsymbol{\lambda})\}$.
  where $a_n \doteq b_n \iff \lim_{n\to\infty}\frac{1}{n}\log a_n = \lim_{n\to\infty}\frac{1}{n}\log b_n$
  And we look into this kind of equivalence because of our quantity of interest (10).

- Example: $f_n(\lambda) = \frac{1}{n}\log\int_{\mathbb{R}^n}\exp\{\sum_{i=1}^n h(\sigma_i;\lambda)\}\prod_{i=1}^n d\sigma_i$

  $$\int_{\mathbb{R}}\exp\{nf_n(\lambda)\}d\lambda \doteq \int_{\mathbb{R}}\int_{\mathbb{R}^n}\exp\{\sum_{i=1}^n h(\sigma_i;\lambda)\}\prod_{i=1}^n d\sigma_i \doteq \sup_{\lambda\in\mathbb{R}}\int_{\mathbb{R}^n}\exp\{\sum_{i=1}^n h(\sigma_i;\lambda)\}\prod_{i=1}^n d\sigma_i$$

2.4 Saddle point approximation (Method of steepest descent)

- If $f_n(i\lambda) \to f(\lambda) \text{as} n \to \infty$, then $\int_{\mathbb{R}} \exp\{nf_n(i\lambda)\}d\lambda \doteq \operatorname*{ext}_{\lambda\in\mathbb{C}} \exp\{nf_n(\lambda)\}$.
  where $\operatorname*{ext}_{\lambda\in\mathbb{C}} \exp\{nf_n(\lambda)\} \equiv \{f(\lambda_*) : f'(\lambda_*) = 0\}$

- Example: $f_n(i\lambda) = \frac{1}{n}\log\int_{\mathbb{R}^n} \exp\{\sum_{i=1}^n h(\sigma_i; i\lambda)\}\prod_{i=1}^n d\sigma_i$
  $\int_{\mathbb{R}} \exp\{nf_n(i\lambda)\}d\lambda \doteq \int_{\mathbb{R}}\int_{\mathbb{R}^n} \exp\{\sum_{i=1}^n h(\sigma_i; i\lambda)\}\prod_{i=1}^n d\sigma_i \doteq \sup_{\lambda\in\mathbb{C}}\int_{\mathbb{R}^n} \exp\{\sum_{i=1}^n h(\sigma_i; \lambda)\}\prod_{i=1}^n d\sigma_i$

2.5 Example: Large deviation of overlap matrix

Let $\boldsymbol{\sigma}_1, \dots, \boldsymbol{\sigma}_k \in \mathbb{R}^n$, k fixed, $n \to \infty$, $(\boldsymbol{\sigma}_i)_{i\in[k]} \sim_{i.i.d.} \text{Uinf}(\mathbb{S}^{n-1}(\sqrt{n}))$. $\boldsymbol{\sigma} = [\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2 \dots, \boldsymbol{\sigma}_k] \in \mathbb{R}^{n\times k}$.$\bar{\boldsymbol{Q}}(\boldsymbol{\sigma}) = \boldsymbol{\sigma}^\mathsf{T}\boldsymbol{\sigma}/n \in \mathbb{R}^{k\times k}$, so $\bar{\boldsymbol{Q}}(\boldsymbol{\sigma})$ is symmetric and $\bar{\boldsymbol{Q}}(\boldsymbol{\sigma})_{ii} = 1$. Let $\boldsymbol{Q} \in \mathbb{R}^{k\times k}$ be a symmetric matrix with $\boldsymbol{Q}(\boldsymbol{\sigma})_{ii} = 1$. Interest: $\mathbb{P}(\boldsymbol{Q} \approx \bar{\boldsymbol{Q}})$. i.e

$$\lim_{\epsilon\to 0}\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}\Big(\bar{Q}(\sigma)_{ij} \in [Q_{ij} - \epsilon, Q_{ij} + \epsilon], \forall i, j\Big) \tag{16}$$

goes to some nontrivial constant.

Physicist prospective:

$$\mathbb{P}\Big(\boldsymbol{Q} \approx \bar{\boldsymbol{Q}}\Big) \doteq \frac{\int_{\mathbb{R}^{n\times k}} \delta(\bar{\boldsymbol{Q}}(\boldsymbol{\sigma}) - \boldsymbol{Q})\prod_{i=1}^k d\boldsymbol{\sigma}_i}{\int_{\mathbb{R}^{n\times k}} \prod_{i=1}^n \delta(\|\boldsymbol{\sigma}_i\|_2^2/n - 1)\prod_{i=1}^k d\boldsymbol{\sigma}_i} \doteq \frac{S_n(\boldsymbol{Q})}{T_n} \tag{17}$$

[Actually $T_n \doteq S_n(\boldsymbol{I})$]
We now calculate $S_n(\boldsymbol{Q})$: Using (12):

$$\begin{aligned} S_n(\boldsymbol{Q}) = \int_{\mathbb{R}^{n\times k}}\Big(\prod_{i=1}^k d\boldsymbol{\sigma}_i\Big)\frac{1}{(2\pi)^{k(k+1)/2}}\int_{\mathbb{R}^{k(k+1)/2}} &\prod_{1\le i<j\le k} \exp\{-i\lambda_{ij}\langle\boldsymbol{\sigma}_i, \boldsymbol{\sigma}_j\rangle + i\lambda_{ij}nQ_{ij}\}\prod_{1\le i<j\le k} d\lambda_{ij} \\ &\prod_{1\le i\le k} \exp\{-i\lambda_{ii}\|\boldsymbol{\sigma}_i\|_2^2/2 + i\lambda_{ii}nQ_{ii}/2\}\prod_{1\le i\le k} d\lambda_{ii} \end{aligned} \tag{18}$$

Using Saddle point approximation :

$$S_n(\boldsymbol{Q}) \doteq \inf_{\Lambda}\int_{\mathbb{R}^{n\times k}}\Big(\prod_{i=1}^k d\boldsymbol{\sigma}_i\Big)\exp\{-\sum_{ij=1}^k \lambda_{ij}\langle\boldsymbol{\sigma}_i, \boldsymbol{\sigma}_j\rangle/2 + n\sum_{ij=1}^k \lambda_{ij}Q_{ij}/2\} \tag{19}$$

$$= \inf_{\Lambda}\int_{\mathbb{R}^{n\times k}}\Big(\prod_{i=1}^k\prod_{\alpha=1}^n d\boldsymbol{\sigma}_i^\alpha\Big)\exp\{-\sum_{ij=1}^k \lambda_{ij}\sum_{\alpha=1}^n \boldsymbol{\sigma}_i^\alpha\boldsymbol{\sigma}_j^\alpha/2\}\exp\{n\sum_{ij=1}^k \lambda_{ij}Q_{ij}/2\} \tag{20}$$

$$= \inf_{\Lambda}\Big(\underline{\int_{\mathbb{R}^k}\Big(\prod_{i=1}^k d\sigma_i\Big)\exp\{-\sum_{ij=1}^k \lambda_{ij}\sum_{\alpha=1}^n \sigma_i\sigma_j/2\}}\Big)^n\exp\{n\sum_{ij=1}^k \lambda_{ij}Q_{ij}/2\} \tag{21}$$

$$\tag{22}$$

Using Gaussian Identity Formula (14):

$$S_n(\boldsymbol{Q}) \doteq \inf_{\Lambda}\Big(\det(\Lambda)^{-\frac{1}{2}}(\sqrt{2\pi})^k\Big)^n\exp\{n\langle\Lambda, \boldsymbol{Q}\rangle/2\} \tag{23}$$

$$= \inf_{\Lambda}\exp\{n[\langle\Lambda, \boldsymbol{Q}\rangle/2 - \frac{1}{2}\log\det(\Lambda) + \frac{k}{2}\log(2\pi)]\} \tag{24}$$

$$\tag{25}$$

Therefore,

$$\frac{1}{n}\log S_n(\boldsymbol{Q}) = \inf_{\Lambda}[\langle \Lambda, \boldsymbol{Q}\rangle/2 - \frac{1}{2}\log\det(\Lambda) + \frac{k}{2}\log(2\pi)] \tag{26}$$

$$= \frac{1}{2}\log\det(\boldsymbol{Q}) + \frac{k}{2}\log(2\pi) \tag{27}$$

Therefore,

$$\frac{1}{n}\log\mathbb{P}\Big(\bar{\boldsymbol{Q}} \approx \boldsymbol{Q}\Big) \doteq \frac{1}{n}\log S_n(\boldsymbol{Q}) - \frac{1}{n}\log S_n(\boldsymbol{I}) \tag{28}$$

$$= \frac{1}{2}\log\det(\boldsymbol{Q}) + \frac{k}{2}\log(2\pi) - \frac{k}{2}\log(2\pi) \tag{29}$$

$$= \frac{1}{2}\log\det(\boldsymbol{Q}) \tag{30}$$