Stat 217, Spring 2023 *Jiaze Qiu* (jiazeqiu@g.harvard.edu)

Section 9: Detection v.s. Recovery [a]

- Sections: Wed, 7:30-8:30pm (SC 705); OHs: Wed 8:30-9:30pm (SC 316.07).
- All the section materials (handouts & solutions) can be found either on Canvas or here.

[a] This handout is based on [9].

Definition 1. *Let distributions P_n, Q_n be defined on the measurable space $(\Omega_n, \mathcal{F}_n)$. We say that the sequence Q_n is contiguous to P_n, and write $Q_n \triangleleft P_n$, if for any sequence A_n of events,*

$$\lim_{n\to\infty} P_n\left(A_n\right) = 0 \quad \Longrightarrow \quad \lim_{n\to\infty} Q_n\left(A_n\right) = 0.$$

Lemma 1. *If $Q_n \triangleleft P_n$, then there is no hypothesis test of the alternative Q_n against the null P_n with* $\Pr[\textit{type I error}] + \Pr[\textit{type II error}] = o(1)$.

Note that $Q_n \triangleleft P_n$ and $P_n \triangleleft Q_n$ are not equivalent, but either of them implies non-distinguishability.

Our goal today is to show thresholds below which spiked and unspiked random matrix models are contiguous.

Lemma 2. *Let $\{P_n\}$ and $\{Q_n\}$ be two sequences of distributions on $(\Omega_n, \mathcal{F}_n)$. If the second moment*

$$\mathop{\mathbb{E}}_{P_n}\left[\left(\frac{\mathrm{d}Q_n}{\mathrm{d}P_n}\right)^2\right]$$

exists and remains bounded as $n \to \infty$, then $Q_n \triangleleft P_n$.

1. Prove this lemma.

> **Solution:** Let $\{A_n\}$ be a sequence of events. Using Cauchy-Schwarz,
>
> $$Q_n\left(A_n\right) = \int_{A_n} \frac{\mathrm{d}Q_n}{\mathrm{d}P_n}\, \mathrm{d}P_n \leq \sqrt{\int_{A_n}\left(\frac{\mathrm{d}Q_n}{\mathrm{d}P_n}\right)^2 \mathrm{d}P_n} \cdot \sqrt{\int_{A_n} \mathrm{d}P_n}.$$
>
> The first factor on the right-hand side is bounded; so if $P_n\left(A_n\right) \to 0$ then also $Q_n\left(A_n\right) \to 0$

> **Solution:** Moreover, given a value of the second moment, we are able to obtain bounds on the tradeoff between type I and type II error in hypothesis testing, which are valid nonasymptotically. Note that this implies, showing that two (sequences of) distributions are contiguous does not rule out the existence of a test that distinguishes between them with constant error probability (better than random guessing).

Lemma 3. *Consider a hypothesis test of a simple alternative Q against a simple null P. Let α be the probability of type I error, and β the probability of type II error. Regardless of the test, we must have*

$$\frac{(1-\beta)^2}{\alpha} + \frac{\beta^2}{(1-\alpha)} \le \mathop{\mathbb{E}}_{P} \left(\frac{\mathrm{d}Q}{\mathrm{d}P}\right)^2,$$

assuming the right-hand side is defined and finite. Furthermore, this bound is tight: for any $\alpha, \beta \in (0,1)$ there exist P, Q and a test for which equality holds.

2. Prove the lemma above and discuss the difference between Lemma 2 and Lemma 3.

Solution: Let A be the event that the test selects the alternative Q, and let $\bar{A}$ be its complement.

$$\begin{aligned}
\mathop{\mathbb{E}}_{P} \left(\frac{\mathrm{d}Q}{\mathrm{d}P}\right)^2 &= \int \frac{\mathrm{d}Q}{\mathrm{d}P}\,\mathrm{d}Q = \int_A \frac{\mathrm{d}Q}{\mathrm{d}P}\,\mathrm{d}Q + \int_{\bar{A}} \frac{\mathrm{d}Q}{\mathrm{d}P}\,\mathrm{d}Q \\
&\ge \frac{\left(\int_A \mathrm{d}Q\right)^2}{\int_A (\mathrm{d}P/\mathrm{d}Q)\mathrm{d}Q} + \frac{\left(\int_{\bar{A}} \mathrm{d}Q\right)^2}{\int_{\bar{A}} (\mathrm{d}P/\mathrm{d}Q)\mathrm{d}Q} = \frac{(1-\beta)^2}{\alpha} + \frac{\beta^2}{(1-\alpha)},
\end{aligned}$$

where the inequality follows from Cauchy-Schwarz. The following example shows tightness: let $P = \mathrm{Bernoulli}(\alpha)$ and let $Q = \mathrm{Bernoulli}(1-\beta)$. On input 0 , the test chooses P, and on input 1 , it chooses Q.

Definition 2 (Gaussian Wigner Spiked Matrix Model)**.** *We observe $Y = \lambda xx + \frac{1}{\sqrt{n}}W$, where W is an $n \times n$ random symmetric matrix with entries drawn iid (up to symmetry) from a fixed distribution of mean 0 and variance 1.*

Question 1. *Can we "detect" whether there is a spike or not?*

3. Try to formalize the question above. Is there a difference between "detection" and "recovery"?

Solution:

We will adopt a Bayesian point of view from now on. Namely, we assume a priori $x \sim \mathcal{X}$, where $\mathcal{X} = \mathcal{X}_n$ is a sequence of distributions on $\mathbb{R}^n$, with the default example being $\mathcal{N}(0, I_n/n)$. It is understood that $\|x\| \approx 1$. We use $\mathrm{GWig}_n(\lambda, \mathcal{X})$ to denote the corresponding distribution of Y.

Lemma 4. *Let $\lambda \ge 0$. Let $Q_n = \mathrm{GWig}_n(\lambda, \mathcal{X})$ and $P_n = \mathrm{GWig}_n(0)$. Let x and x' be independently drawn from $\mathcal{X}_n$. Then*

$$\mathop{\mathbb{E}}_{P_n} \left(\frac{\mathrm{d}Q_n}{\mathrm{d}P_n}\right)^2 = \mathop{\mathbb{E}}_{x,x'} \exp\left(\frac{n\lambda^2}{2} \langle x, x' \rangle^2\right)$$

4. Prove Lemma 4.

Solution: Let $Q_n = \mathrm{GWig}_n(\lambda, \mathcal{X})$, i.e., the spiked distribution, and $P_n = \mathrm{GWig}_n(0)$, i.e., the unspiked distribution. First, we simplify the likelihood ratio:

$$\frac{\mathrm{d}Q_n}{\mathrm{d}P_n} = \frac{\mathbb{E}_{x \sim \mathcal{X}_n} \exp\left(-\frac{n}{4}\left\langle Y - \lambda xx^\top, Y - \lambda xx^\top \right\rangle\right)}{\exp\left(-\frac{n}{4}\langle Y, Y\rangle\right)}$$
$$= \mathop{\mathbb{E}}_{x \sim \mathcal{X}_n} \exp\left(\frac{\lambda n}{2}\left\langle Y, xx^\top \right\rangle - \frac{n\lambda^2}{4}\left\langle xx^\top, xx^\top \right\rangle\right).$$

Now passing to the second moment:

$$\mathop{\mathbb{E}}_{P_n}\left(\frac{\mathrm{d}Q_n}{\mathrm{d}P_n}\right)^2 = \mathop{\mathbb{E}}_{x,x' \sim \mathcal{X}_n} \mathop{\mathbb{E}}_{\sim P_n} \exp\left(\frac{\lambda n}{2}\left\langle Y, xx^\top + x'x'^\top \right\rangle - \frac{n\lambda^2}{4}\left(\left\langle xx^\top, xx^\top \right\rangle + \left\langle x'x'^\top, x'x'^\top \right\rangle\right)\right)$$

where x and x' are drawn independently from $\mathcal{X}_n$. We now simplify the Gaussian moment-generating function over the randomness of Y, and cancel terms, to arrive at the expression

$$= \mathop{\mathbb{E}}_{x,x'} \exp\left(\frac{n\lambda^2}{2}\left\langle x, x' \right\rangle^2\right),$$

which proves Lemma 4.

It is well known that our spiked Wigner model admits the following spectral behavior.

Theorem 1. *Let Y be drawn from $\mathrm{GWig}(\lambda, \mathcal{X})$ with any spike prior $\mathcal{X}$ supported on unit vectors ($\|x\| = 1$):*

- *If $\lambda \leq 1$, the top eigenvalue of Y converges almost surely to 2 as $n \to \infty$, and the top (unit-norm) eigenvector v has trivial correlation with the spike: $\langle v, x\rangle^2 \to 0$ almost surely.*
- *If $\lambda > 1$, the top eigenvalue converges almost surely to $\lambda + 1/\lambda > 2$, and v estimates the spike nontrivially: $\langle v, x\rangle^2 \to 1 - 1/\lambda^2$ almost surely.*

5. Prove that for $\lambda < 1$ "detection" is impossible, assuming $x_i \overset{iid}{\sim} \mathcal{N}(0, 1/n)$.

Solution: Please see [9][Prop. 3.8.]

6. Compare this result to Theorem 1. Do the thresholds for detection adn recovery match? What about more general noise distributions and more general priors on x?

Solution: First of all, roughly speaking, the spectral behavior of this model exhibits universality: regardless of the choice of the noise distributions, many properties of the spectrum behave the same as if the noise came from a standard Gaussian distribution. In particular, for $\lambda \leq 1$, the spectrum bulk has a semicircular distribution and the maximum eigenvalue converges almost surely to 2. For $\lambda > 1$, an isolated eigenvalue emerges from the bulk with value converging to $\lambda + 1/\lambda$, and (under suitable assumptions) the top eigenvector has squared correlation $1 - 1/\lambda^2$ with the truth. In stark contrast, from a statistical standpoint, universality breaks down entirely: the detection problem becomes easier when the noise is non-Gaussian. Equivalently, the detection threshold is actually lower than 1 in the non-Guassian case, or in other words, Gaussian noise is the hardest! See [9] for more details.

References

[1] Bayati, M., and Montanari, A. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory 57*, 2 (2011), 764–785.

[2] Bolthausen, E. An iterative construction of solutions of the tap equations for the sherrington–kirkpatrick model. *Communications in Mathematical Physics 325*, 1 (2014), 333–366.

[3] Chatterjee, S. A simple invariance theorem. *arXiv preprint math/0508213* (2005).

[4] Feng, O. Y., Venkataramanan, R., Rush, C., and Samworth, R. J. A unifying tutorial on approximate message passing. *Foundations and Trends in Machine Learning 15*, 4 (2022), 335–536.

[5] Fox, C. W., and Roberts, S. J. A tutorial on variational bayesian inference. *Artificial intelligence review 38* (2012), 85–95.

[6] Guerra, F., and Toninelli, F. L. The thermodynamic limit in mean field spin glass models. *Communications in Mathematical Physics 230* (2002), 71–79.

[7] Montanari, A., and Sen, S. A short tutorial on mean-field spin glass techniques for non-physicists. *arXiv preprint arXiv:2204.02909* (2022).

[8] Panchenko, D. *The sherrington-kirkpatrick model*. Springer Science & Business Media, 2013.

[9] Perry, A., Wein, A. S., Bandeira, A. S., and Moitra, A. Optimality and sub-optimality of pca for spiked random matrices and synchronization. *arXiv preprint arXiv:1609.05573* (2016).

[10] Talagrand, M. The generalized parisi formula. *Comptes Rendus Mathematique 337*, 2 (2003), 111–114.