On the Optimality and Sub-optimality of PCA in Spiked Random Matrix Models: supplementary proofs

Amelia Perry*†[1], Alexander S. Wein*‡[1], Afonso S. Bandeira§[1,3], and Ankur Moitra¶[1,2]

[1]Department of Mathematics, Massachusetts Institute of Technology
[2]Computer Science and Artificial Intelligence Lab, Massachusetts Institute of Technology
[3]Department of Mathematics and Center for Data Science, Courant Institute of Mathematical Sciences, New York University, NY, USA

September 29, 2016

A Proof of Theorem 3.7: spherically spiked Wigner

Theorem 3.7. *Consider the spherical prior* $\mathcal{X}_{\mathrm{sph}}$. *If* $\lambda < 1$ *then* $\mathrm{GWig}(\lambda, \mathcal{X}_{\mathrm{sph}})$ *is contiguous to* $\mathrm{GWig}(0)$.

Proof. By symmetry, we reduce the second moment above as

$$\mathop{\mathbb{E}}_{x,x'} \exp\left(\frac{n\lambda^2}{2}\langle x, x'\rangle^2\right) = \mathop{\mathbb{E}}_{x} \exp\left(\frac{n\lambda^2}{2}\langle x, e_1\rangle^2\right) = \mathop{\mathbb{E}}_{x_1} \exp\left(\frac{n\lambda^2}{2}x_1^2\right),$$

where e_1 denotes the first standard basis vector. Note that the first coordinate x_1 of a point uniformly drawn from the unit sphere in $\mathbb{R}^n$ is distributed proportionally to $(1-x_1^2)^{(n-3)/2}$, so that its square y is distributed proportionally to $(1-y)^{(n-3)/2}y^{-1/2}$. Hence y is distributed as $\mathrm{Beta}(\frac{1}{2}, \frac{n-1}{2})$. The second moment is thus the moment generating function of $\mathrm{Beta}(\frac{1}{2}, \frac{n-1}{2})$ evaluated at $n\lambda^2/2$, and as such, we have

$$\mathop{\mathbb{E}}_{Q_n}\left(\frac{\mathrm{d}P_n}{\mathrm{d}Q_n}\right)^2 = {}_1F_1\left(\frac{1}{2}; \frac{n}{2}; \frac{\lambda^2 n}{2}\right), \tag{7}$$

where ${}_1F_1$ denotes the confluent hypergeometric function.

Suppose $\lambda < 1$. Equation 13.8.4 from DLMF grants us that, as $n \to \infty$,

$$\begin{aligned}{}_1F_1\left(\frac{1}{2}; \frac{n}{2}; \frac{\lambda^2 n}{2}\right) = (1+o(1))\left(\frac{n}{2}\right)^{1/4} e^{\zeta^2 n/8}\Bigg(&\lambda^2\sqrt{\frac{\zeta}{1-\lambda^2}}U(0, \zeta\sqrt{n/2}) \\ &+ \left(-\lambda^2\sqrt{\frac{\zeta}{1-\lambda^2}} + \sqrt{\frac{\zeta}{1-\lambda^2}}\right)\frac{U(-1, \zeta\sqrt{n/2})}{\zeta\sqrt{n/2}}\Bigg),\end{aligned}$$

*The first two authors contributed equally.

†Email: ameliaperry@mit.edu. This work is supported in part by NSF CAREER Award CCF-1453261 and a grant from the MIT NEC Corporation.

‡Email: awein@mit.edu. This research was conducted with Government support under and awarded by DoD, Air Force Office of Scientific Research, National Defense Science and Engineering Graduate (NDSEG) Fellowship, 32 CFR 168a.

§Email: bandeira@cims.nyu.edu. A.S.B. was supported by NSF Grant DMS-1317308. Part of this work was done while A.S.B. was with the Department of Mathematics at the Massachusetts Institute of Technology.

¶Email: moitra@mit.edu. This work is supported in part by NSF CAREER Award CCF-1453261, a grant from the MIT NEC Corporation and a Google Faculty Research Award.

where $\zeta = \sqrt{2(\lambda^2 - 1 - 2\log\lambda)}$ and U is the parabolic cylinder function,

$$= (1+o(1)) \left(\frac{n}{2}\right)^{1/4} e^{\zeta^2 n/8} \left(\lambda^2 \sqrt{\frac{\zeta}{1-\lambda^2}} e^{-\zeta^2 n/8} (\zeta\sqrt{n/2})^{-1/2} \right.$$
$$\left. + \left(-\lambda^2 \sqrt{\frac{\zeta}{1-\lambda^2}} + \sqrt{\frac{\zeta}{1-\lambda^2}} \right) \frac{e^{-\zeta^2 n/8} (\zeta\sqrt{n/2})^{1/2}}{\zeta\sqrt{n/2}} \right),$$

by Equation 12.9.1 from DLMF,

$$= (1+o(1))(1-\lambda^2)^{-1/2},$$

which is bounded as $n \to \infty$, for all $\lambda < 1$. The result follows from Lemma 2.3. □

B Conditioning method for Gaussian Wigner model

In this section we give the full details of the conditioning method for the Gaussian Wigner model. We assume that the prior $\mathcal{X} = \mathrm{iid}(\pi)$ draws each entry of x independently from $\frac{1}{\sqrt{n}}\pi$ where π is a finitely-supported distribution on $\mathbb{R}$ with mean zero and variance one.

The argument that we will use is based on Banks et al. [2016a], in particular their Proposition 5. Suppose ω_n is a set of 'good' x values so that $x \in \omega_n$ with probability $1 - o(1)$. Let $P_n = \mathrm{GWig}_n(\lambda, \mathcal{X})$ and let $Q_n = \mathrm{GWig}_n(0)$. Let $\tilde{\mathcal{X}}$ be the prior that draws x from $\mathcal{X}$, outputs x if $x \in \omega_n$, and outputs the zero vector otherwise. Let $\tilde{P}_n = \mathrm{GWig}_n(\lambda, \tilde{\mathcal{X}})$. Our goal is to show $\tilde{P}_n \lhd Q_n$, from which it follows that $P_n \lhd Q_n$ (see Lemma 2.4). In our case, the bad events are when the empirical distribution of x differs significantly from π, i.e. x has atypical proportions of entries. If we let Ω_n be the event that x and x' are both in ω_n, our second moment becomes

$$\mathop{\mathbb{E}}_{Q_n} \left(\frac{\mathrm{d}\tilde{P}_n}{\mathrm{d}Q_n}\right)^2 = \mathop{\mathbb{E}}_{\tilde{x},\tilde{x}' \sim \tilde{\mathcal{X}}} \left[\exp\left(\frac{n\lambda^2}{2} \langle \tilde{x}, \tilde{x}' \rangle^2 \right) \right] = \mathop{\mathbb{E}}_{x,x' \sim \mathcal{X}} \left[\mathbb{1}_{\Omega_n} \exp\left(\frac{n\lambda^2}{2} \langle x, x' \rangle^2 \right) \right] + o(1).$$

Let $\Sigma \subseteq \mathbb{R}$ (a finite set) be the support of π, and let $s = |\Sigma|$. We will index Σ by $[s] = \{1, 2, \ldots, s\}$ and identify π with the vector of probabilities $\pi \in \mathbb{R}^s$. For $a, b \in \Sigma$, let N_{ab} denote the number of indices i for which $x_i = \frac{a}{\sqrt{n}}$ and $x'_i = \frac{b}{\sqrt{n}}$ (recall x_i is drawn from $\frac{1}{\sqrt{n}}\pi$). Note that N follows a multinomial distribution with n trials, s^2 outcomes, and with probabilities given by $\bar{\alpha} = \pi\pi^\top \in \mathbb{R}^{s \times s}$. We have

$$\frac{n\lambda^2}{2}\langle x, x' \rangle^2 = \frac{\lambda^2}{2n} \left(\sum_{a,b \in \Sigma} abN_{ab} \right)^2 = \frac{\lambda^2}{2n} \sum_{a,b,a',b'} aba'b'N_{ab}N_{a'b'} = \frac{1}{n} N^\top AN$$

where A is the $s^2 \times s^2$ matrix $A_{ab,a'b'} = \frac{\lambda^2}{2}aba'b'$, and the quadratic form $N^\top AN$ is computed by treating N as a vector of length s^2.

We are now in a position to apply Proposition 5 from Banks et al. [2016a]. Define $Y = (N - n\bar{\alpha})/\sqrt{n}$. Let Ω_n be the event defined in Appendix A of Banks et al. [2016a], which enforces that the empirical distributions of x and x' are close to π (in a specific sense).

Note that $\bar{\alpha}$ (treated as a vector of length s^2) is in the kernel of A because π is mean-zero: the inner product between $\bar{\alpha}$ and the (a, b) row of A is

$$\sum_{a',b'} A_{ab,a'b'}\bar{\alpha}_{a'b'} = \frac{\lambda^2}{2} \sum_{a',b'} aba'b'\pi_{a'}\pi_{b'} = \frac{\lambda^2}{2} ab \left(\sum_{a'} a'\pi_{a'} \right) \left(\sum_{b'} b'\pi_{b'} \right) = 0.$$

Therefore we have $\frac{1}{n}N^\top AN = Y^\top AY$ and so we can write our second moment as $\mathbb{E}[\mathbb{1}_{\Omega_n} \exp(Y^\top AY)] + o(1)$.

Let $\Delta_{s^2}(\pi)$ denote the set of nonnegative vectors $\alpha \in \mathbb{R}^{s^2}$ with row- and column-sums prescribed by π, i.e. treating α as an $s \times s$ matrix, we have (for all i) that row i and column i of α each sum to π_i. Let $D(u,v)$ denote the KL divergence between two vectors: $D(u,v) = \sum_i u_i \log(u_i/v_i)$. For convenience, we restate Proposition 5 in Banks et al. [2016a].

Proposition B.1 (Banks et al. [2016a] Proposition 5)**.** *Let $\pi \in \mathbb{R}^s$ be any vector of probabilities. Let A be any $s^2 \times s^2$ matrix. Define N, Y, $\bar{\alpha}$, and Ω_n as above (depending on π). Let*

$$m = \sup_{\alpha \in \Delta_{s^2}(\pi)} \frac{(\alpha - \bar{\alpha})^\top A (\alpha - \bar{\alpha})}{D(\alpha, \bar{\alpha})}.$$

If $m < 1$ then $\lim_{n\to\infty} \mathbb{E}[\mathbb{1}_{\Omega_n} \exp(Y^\top A Y)] = \mathbb{E}[\exp(Z^\top A Z)] < \infty$, where $Z \sim \mathcal{N}(0, \operatorname{diag}(\bar{\alpha}) - \bar{\alpha}\bar{\alpha}^\top)$. If $m > 1$ then $\lim_{n\to\infty} \mathbb{E}[\mathbb{1}_{\Omega_n} \exp(Y^\top A Y)] = \infty$.

We apply Proposition B.1 to our specific choice of π and A:

Theorem 3.11 (conditioning method)**.** *Let $\mathcal{X}_n = \operatorname{iid}(\pi)$ where π has finite support $\Sigma \subseteq \mathbb{R}$ with $|\Sigma| = s$. Let $P_n = \operatorname{GWig}_n(\lambda, \mathcal{X})$, $\tilde{P}_n = \operatorname{GWig}_n(\lambda, \tilde{\mathcal{X}})$, and $Q_n = \operatorname{GWig}_n(0)$. Define the $s \times s$ matrix $\beta_{ab} = ab$ for $a, b \in \Sigma$. Let $D(u,v)$ denote the KL divergence between two vectors: $D(u,v) = \sum_i u_i \log(u_i/v_i)$. Identify π with the vector of probabilities $\pi \in \mathbb{R}^\Sigma$, and define $\bar{\alpha} = \pi\pi^\top$. Let $\Delta_{s^2}(\pi)$ denote the set of $s \times s$ matrices with row- and column-sums prescribed by π, i.e. row i and column i of α each sum to π_i. Let*

$$\lambda^*_{\mathcal{X}} = \left[\sup_{\alpha \in \Delta_{s^2}(\pi)} \frac{\langle \alpha, \beta \rangle^2}{2D(\alpha, \bar{\alpha})} \right]^{-1/2}. \tag{8}$$

If $\lambda < \lambda^*_{\mathcal{X}}$ then $\lim_{n\to\infty} \mathbb{E}_{Q_n} (\mathrm{d}\tilde{P}_n/\mathrm{d}Q_n)^2 = (1-\lambda^2)^{-1/2} < \infty$ and so $P_n \lhd Q_n$. Conversely, if $\lambda > \lambda^*_{\mathcal{X}}$ then $\lim_{n\to\infty} \mathbb{E}_{Q_n} \left(\mathrm{d}\tilde{P}_n/\mathrm{d}Q_n\right)^2 = \infty$.

Note that this is a tight characterization of when the second moment is bounded, but not necessarily of when contiguity holds.

Above we have computed the limit value of the second moment in the case $\lambda < \lambda^*_{\mathcal{X}}$ as follows. Defining Z as in Proposition B.1 we have $\langle Z, \beta \rangle \sim \mathcal{N}(0, \sigma^2)$ where

$$\begin{aligned} \sigma^2 = \beta^\top (\operatorname{diag}(\bar{\alpha}) - \bar{\alpha}\bar{\alpha}^\top)\beta &= \sum_{ab} \beta_{ab}^2 \bar{\alpha}_{ab} + \left(\sum_{ab} \beta_{ab} \bar{\alpha}_{ab} \right)^2 \\ &= \left(\sum_a a^2 \pi_a \right) \left(\sum_b b^2 \pi_b \right) + \left(\sum_a a\pi_a \sum_b b\pi_b \right)^2 = 1, \end{aligned}$$

since π is mean-zero and unit-variance, and so

$$\mathbb{E}[\exp(Z^\top A Z)] = \mathbb{E}\left[\exp\left(\frac{\lambda^2}{2} \langle Z, \beta \rangle^2 \right) \right] = \mathbb{E}\left[\exp\left(\frac{\lambda^2}{2} \chi_1^2 \right) \right] = (1-\lambda^2)^{-1/2}.$$

C Sparse Rademacher prior

In this section we give details for our results on the spiked Gaussian Wigner model with the i.i.d. sparse Rademacher prior: $\pi = \sqrt{1/\rho}\,\mathcal{R}(\rho)$ where $\mathcal{R}(\rho)$ is the sparse Rademacher distribution with sparsity $\rho \in [0,1]$:

$$\mathcal{R}(\rho) = \begin{cases} 0 & \text{w.p.} \quad 1-\rho \\ +1 & \text{w.p.} \quad \rho/2 \\ -1 & \text{w.p.} \quad \rho/2 \end{cases}.$$

First we try the sub-Gaussian method of Section 3.4. Note that $\pi\pi' = \frac{1}{\rho}\mathcal{R}(\rho^2)$. The variance proxy σ^2 for $\pi\pi'$ needs to satisfy

$$\exp\left(\frac{1}{2}\sigma^2 t^2\right) \geq \mathbb{E}\exp(t\pi\pi') = 1 - \rho^2 + \frac{\rho^2}{2}\cosh(t/\rho) \tag{9}$$

for all $t \in \mathbb{R}$ so the best (smallest) choice for σ^2 is

$$(\sigma^*)^2 = \sup_{t\in\mathbb{R}} \frac{2}{t^2}\log\left[1 - \rho^2 + \frac{\rho^2}{2}\cosh(t/\rho)\right].$$

Recall that Theorem 3.9 (sub-Gaussian method) gives contiguity for all $\lambda < 1/\sigma^*$. We now resolve a conjecture stated in Banks et al. [2016b]. For sufficiently large ρ, we have $\sigma^* = 1$, implying that PCA is tight:

Theorem 3.12. *When $\rho \geq 1/\sqrt{3} \approx 0.577$, we have $\sigma^* = 1$, yielding contiguity for all $\lambda < 1$. On the other hand, if $\rho < 1/\sqrt{3}$, then $\sigma^* > 1$.*

Proof. We equivalently consider the following reformulation of (9):

$$\frac{1}{2}\sigma^2 t^2 \overset{?}{\geq} \log\left(1 - \rho^2 + \rho^2\cosh(t/\rho)\right) \triangleq k_\rho(t). \tag{10}$$

Both sides of the inequality are even functions, agreeing in value at $t = 0$. When $\sigma^2 < 1$, the inequality fails, by comparing their second-order behavior about $t = 0$. When $\sigma^2 = 1$ but $\rho < 1/\sqrt{3}$, the inequality fails, as the two sides have matching behavior up to third order, but $k_\rho^{(4)}(0) = 3 - \rho^{-2} < 0$.

It remains to show that the inequality (10) does hold for $\rho > 1/\sqrt{3}$ and $\sigma^2 = 1$. As the left and right sides agree to first order at $t = 0$, and are both even functions, it suffices to show that for all $t \geq 0$,

$$1 \overset{?}{\geq} k_\rho''(t) = \frac{\rho^2 + (1-\rho^2)\cosh(t/\rho)}{(1 - \rho^2 + \rho^2\cosh(t/\rho))^2}.$$

Completing the square for cosh, we have the equivalent inequality:

$$0 \overset{?}{\leq} 1 - 3\rho^2 + \rho^4 + \Bigg(\underbrace{\rho^2\cosh(t/\rho) + \frac{(2\rho^2-1)(1-\rho^2)}{2\rho^2}}_{(*)}\Bigg)^2 - \frac{(2\rho^2-1)^2(1-\rho^2)^2}{4\rho^4}.$$

Note that cosh is bounded below by 1; thus for $\rho > 1/\sqrt{3}$, the underbraced term $(*)$ is nonnegative, and hence minimized in absolute value when $t = 0$. It then suffices to show the above inequality in the case $t = 0$, so that $\cosh(t/\rho) = 1$; but here the inequality is an equality, by simple algebra. □

Using the conditioning method of Section 3.6, we will now improve the range of ρ for which PCA is optimal, although our argument here relies on numerical optimization.

Example 3.13. *Let $\mathcal{X}$ be the sparse Rademacher prior* $\mathrm{iid}(\sqrt{1/\rho}\,\mathcal{R}(\rho))$. *There exists a critical value $\rho^* \approx$ 0.184 (numerically computed) such that if $\rho \geq \rho^*$ and $\lambda < 1$ then* $\mathrm{GWig}(\lambda, \mathcal{X})$ *is contiguous to* $\mathrm{GWig}(0, \mathcal{X})$. *When $\rho < \rho^*$ we are only able to show contiguity when $\lambda < \lambda_\rho^*$ for some $\lambda_\rho^* < 1$.*

Details. Consider the optimization problem of Theorem 3.11 (conditioning method). We will first use symmetry to argue that the optimal α must take a simple form. Abbreviate the support of π as $\{0, +, -\}$. For a given α matrix, define its complement by swapping $+$ and $-$, e.g. swap α_{0+} with α_{0-} and swap α_{-+} with α_{+-}. Note that if we average α with its complement, the numerator $\langle\alpha, \beta\rangle^2$ remains unchanged, the denominator $D(\alpha, \overline{\alpha})$ can only decrease, and the row- and column-sum constraints remain satisfied; this means the new solution is at least as good as the original α. Therefore we only need to consider α values satisfying $\alpha_{++} = \alpha_{--}$ and $\alpha_{+-} = \alpha_{-+}$. Note that the remaining entries of α are uniquely determined by the row- and column-sum constraints, and so we have reduced the problem to only two variables. It is now easy to solve the optimization problem numerically, say by grid search. □

D Proof of Proposition 4.2: Gaussian noise is hardest

Proposition 4.2. *Let $\mathcal{P}$ be a continuous distribution with a C^1 density function $p(w)$ with $p(w) > 0$ everywhere. Suppose* $\mathrm{Var}[\mathcal{P}] = 1$*. Then $F_{\mathcal{P}} \geq 1$ with equality if and only if $\mathcal{P}$ is a standard Gaussian.*

Proof. Since $F_{\mathcal{P}}$ is translation-invariant, assume $\mathbb{E}[\mathcal{P}] = 0$ without loss of generality. We have

$$\begin{aligned}
0 &\leq \int_{-\infty}^{\infty} \frac{1}{p(w)} \left(p'(w) + wp(w)\right)^2 \,\mathrm{d}w \\
&= \int_{-\infty}^{\infty} \left[\frac{p'(w)^2}{p(w)} + 2wp'(w) + w^2 p(w)\right] \mathrm{d}w \\
&= F_{\mathcal{P}} + \int_{-\infty}^{\infty} 2wp'(w)\,\mathrm{d}w + 1
\end{aligned}$$

since $\mathbb{E}[\mathcal{P}] = 0$ and $\mathrm{Var}[\mathcal{P}] = 1$. (The integral in the first line is finite, provided that $F_{\mathcal{P}}$ and $\mathrm{Var}[\mathcal{P}]$ are finite.) Using integration by parts,

$$\int_{-\infty}^{\infty} 2wp'(w)\,\mathrm{d}w = 2wp(w)\Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} 2p(w)\,\mathrm{d}w = -2$$

since $wp(w) \to 0$ as $w \to \pm\infty$ or else $p(w)$ would not be integrable. (Here we have used the fact that the limits $\lim_{w\to\pm\infty} wp(w)$ must exists, since the left-hand side is defined.) We now have $F_{\mathcal{P}} \geq 1$. Equality holds only if $p'(w) = -wp(w)$ for all w. We solve this differential equation as $p(w) = C\exp\left(-\frac{w^2}{2}\right)$, which is a standard Gaussian. □

E Proof of non-Gaussian Wigner lower bounds

In this section we prove Theorem 4.4, and verify its hypotheses for spherical and i.i.d. priors.

Theorem 4.4. *Under Assumption 4.3,* $\mathrm{Wig}(\lambda, \mathcal{P}, \mathcal{X})$ *is contiguous to* $\mathrm{Wig}(0, \mathcal{P})$ *for all $\lambda < \lambda^*_{\mathcal{X}}/\sqrt{F_{\mathcal{P}}}$.*

Proof. We begin by defining a modification $\tilde{\mathcal{X}}$ of the prior $\mathcal{X}$, by returning the spike 0 whenever one of the tail events described in Assumption 4.3 occur—namely, when some entry x_i exceeds $n^{-1/3}$ in magnitude, or when $\|x\|_q > \alpha_q$ for some $q \in \{2, 4, 6, 8\}$. By hypothesis, with probability $1 - o(1)$, no such tail event occurs; hence if $\mathrm{Wig}(\lambda, \mathcal{P}, \tilde{\mathcal{X}})$ is contiguous to $\mathrm{Wig}(0, \mathcal{P})$ then so is $\mathrm{Wig}(\lambda, \mathcal{P}, \mathcal{X})$. Let $P_n = \mathrm{Wig}_n(\lambda, \mathcal{P}, \mathcal{X})$, $\tilde{P}_n = \mathrm{Wig}_n(\lambda, \mathcal{P}, \tilde{\mathcal{X}})$, and $Q_n = \mathrm{Wig}_n(0, \mathcal{P})$.

We proceed from the second moment:

$$\begin{aligned}
\mathop{\mathbb{E}}_{Q_n}\left(\frac{\mathrm{d}\tilde{P}_n}{\mathrm{d}Q_n}\right)^2 &= \mathop{\mathbb{E}}_{Y\sim Q_n}\left[\mathop{\mathbb{E}}_{x,x'\sim\tilde{\mathcal{X}}} \prod_{i<j} \frac{p(\sqrt{n}Y_{ij} - \lambda\sqrt{n}x_i x_j)}{p(\sqrt{n}Y_{ij})} \frac{p(\sqrt{n}Y_{ij} - \lambda\sqrt{n}x'_i x'_j)}{p(\sqrt{n}Y_{ij})}\right] \\
&= \mathop{\mathbb{E}}_{x,x'\sim\tilde{\mathcal{X}}}\left[\prod_{i<j} \mathop{\mathbb{E}}_{\sqrt{n}Y_{ij}\sim\mathcal{P}} \frac{p(\sqrt{n}Y_{ij} - \lambda\sqrt{n}x_i x_j)}{p(\sqrt{n}Y_{ij})} \frac{p(\sqrt{n}Y_{ij} - \lambda\sqrt{n}x'_i x'_j)}{p(\sqrt{n}Y_{ij})}\right] \\
&= \mathop{\mathbb{E}}_{x,x'\sim\tilde{\mathcal{X}}}\left[\exp\left(\sum_{i<j} \tau(\lambda\sqrt{n}x_i x_j, \lambda\sqrt{n}x'_i x'_j)\right)\right].
\end{aligned}$$

We will expand τ using Taylor's theorem, using the C^4 assumption:

$$\tau(a,b) = \sum_{0\leq k+\ell\leq 3} \frac{\partial^{k+\ell}\tau}{\partial a^k \partial b^\ell}(0,0)\; a^k b^\ell + \sum_{k+\ell=4} \left(\frac{\partial^4\tau}{\partial a^k \partial b^\ell}(0,0) + h_{k,\ell}(a,b)\right) a^k b^\ell$$

for some remainder function $h_{k,\ell}(a,b)$ tending to 0 as $(a,b) \to (0,0)$. Given the bounds assumed on the entries of x and x', these remainder terms $h_{k,\ell}(\lambda\sqrt{n}x_i x_j, \lambda\sqrt{n}x'_i x'_j)$ are $o(1)$ as $n \to \infty$. Note that $\tau(a,0) = 0 = \tau(0,b)$, so that the non-mixed partials of τ vanish. Further, by the hypothesis of noise symmetry, we have $\tau(-a,-b) = \tau(a,b)$, so that all partials of odd total degree vanish; in particular the mixed third partials vanish. We note also that $\frac{\partial^2 \tau}{\partial a \partial b}(0,0) = F_{\mathcal{P}}$, the Fisher information defined above. Thus,

$$\begin{aligned}
\mathop{\mathbb{E}}_{Q_n} \left(\frac{\mathrm{d}\tilde{P}_n}{\mathrm{d}Q_n} \right)^2 &= \mathop{\mathbb{E}}_{x,x' \sim \tilde{\mathcal{X}}} \left[\exp\left(F_{\mathcal{P}} \lambda^2 n \sum_{i<j} x_i x_j x'_i x'_j + \sum_{k+\ell=4} \left(\frac{\partial^4 \tau}{\partial a^k \partial b^\ell}(0,0) + o(1) \right) \lambda^4 n^2 \sum_{i<j} x_i^k x_j^k (x'_i)^\ell (x'_j)^\ell \right) \right] \\
&\le \mathop{\mathbb{E}}_{x,x' \sim \tilde{\mathcal{X}}} \left[\exp\left(\frac{F_{\mathcal{P}} \lambda^2 n}{2} \langle x, x' \rangle^2 \right) \prod_{k+\ell=4} \exp\left(\left(\frac{\partial^4 \tau}{\partial a^k \partial b^\ell}(0,0) + o(1) \right) \frac{\lambda^4 n^2}{2} \langle x^k, (x')^\ell \rangle^2 \right) \right],
\end{aligned}$$

where x^k denotes entrywise kth power. For all $\varepsilon > 0$, we can apply the weighted AM–GM inequality:

$$\begin{aligned}
&\le \mathop{\mathbb{E}}_{x,x' \sim \tilde{\mathcal{X}}} \left[(1-\varepsilon) \exp\left(\frac{F_{\mathcal{P}} \lambda^2 n}{2} \langle x, x' \rangle^2 \right)^{(1-\varepsilon)^{-1}} + \sum_{k+\ell=4} \frac{\varepsilon}{5} \exp\left(\left(\frac{\partial^4 \tau}{\partial a^k \partial b^\ell}(0,0) + o(1) \right) \frac{\lambda^4 n^2}{2} \langle x^k, (x')^\ell \rangle^2 \right)^{5/\varepsilon} \right] \\
&= \mathop{\mathbb{E}}_{x,x' \sim \tilde{\mathcal{X}}} \left[(1-\varepsilon) \exp\left(\frac{(1-\varepsilon)^{-1} F_{\mathcal{P}} \lambda^2 n}{2} \langle x, x' \rangle^2 \right) \right] \\
&\quad + \sum_{k+\ell=4} \frac{\varepsilon}{5} \mathop{\mathbb{E}}_{x,x' \sim \tilde{\mathcal{X}}} \left[\exp\left(\left(\frac{\partial^4 \tau}{\partial a^k \partial b^\ell}(0,0) + o(1) \right) \frac{5\lambda^4 n^2}{2\varepsilon} \langle x^k, (x')^\ell \rangle^2 \right) \right], \qquad (11)
\end{aligned}$$

so it suffices to bound each of these expectations.

By hypothesis, $\lambda < \lambda^*_{\mathcal{X}} / \sqrt{F_{\mathcal{P}}}$, implying that we can choose $\varepsilon > 0$ such that $(1-\varepsilon)^{-1} F_{\mathcal{P}} \lambda^2 < (\lambda^*_{\mathcal{X}})^2$. But $\tilde{\mathcal{X}}$ is dominated as a measure by the sum of $\mathcal{X}$ and an $o(1)$ mass at 0; it follows that $\lambda_{\mathcal{X}} \le \lambda_{\tilde{\mathcal{X}}}$, and the first expectation in (11) is bounded.

We bound each of the other expectations using Cauchy–Schwarz:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x,x' \sim \tilde{\mathcal{X}}} \left[\exp\left(\left(\frac{\partial^4 \tau}{\partial a^k \partial b^\ell}(0,0) + o(1) \right) \frac{5\lambda^4 n^2}{2\varepsilon} \langle x^k, (x')^\ell \rangle^2 \right) \right] \\
&\le \mathop{\mathbb{E}}_{x,x' \sim \tilde{\mathcal{X}}} \left[\exp\left(\left(\frac{\partial^4 \tau}{\partial a^k \partial b^\ell}(0,0) + o(1) \right) \frac{5\lambda^4 n^2}{2\varepsilon} \|x^k\|_2^2 \, \|(x')^\ell\|_2^2 \right) \right] \\
&= \mathop{\mathbb{E}}_{x,x' \sim \tilde{\mathcal{X}}} \left[\exp\left(\left(\frac{\partial^4 \tau}{\partial a^k \partial b^\ell}(0,0) + o(1) \right) \frac{5\lambda^4 n^2}{2\varepsilon} \|x\|_{2k}^{2k} \, \|x'\|_{2\ell}^{2\ell} \right) \right] \\
&\le \mathop{\mathbb{E}}_{x,x' \sim \tilde{\mathcal{X}}} \left[\exp\left(\left(\frac{\partial^4 \tau}{\partial a^k \partial b^\ell}(0,0) + o(1) \right) \frac{5\lambda^4 n^2}{2\varepsilon} \alpha_{2k}^{2k} n^{1-k} \alpha_{2\ell}^{2\ell} n^{1-\ell} \right) \right],
\end{aligned}$$

due to the norm restrictions on prior $\tilde{\mathcal{X}}$,

$$= \exp\left(\left(\frac{\partial^4 \tau}{\partial a^k \partial b^\ell}(0,0) + o(1) \right) \frac{5\lambda^4}{2\varepsilon} \alpha_{2k}^{2k} \alpha_{2\ell}^{2\ell} \right),$$

which remains bounded as $n \to \infty$.

With the overall second moment $\mathbb{E}_{Q_n} \left(\frac{\mathrm{d}\tilde{P}_n}{\mathrm{d}Q_n} \right)^2$ bounded as $n \to \infty$, the result follows from Lemma 2.4. □

Proposition 4.5. *Consider the spherical prior $\mathcal{X}_{\text{sph}}$. Then conditions (i) and (ii) in Assumption 4.3 are satisfied.*

Proof. For the spherical prior we have $\lambda^*_{\mathcal{X}_{\text{sph}}} = 1$, as computed in Theorem 3.7. Note that one can sample $x \sim \mathcal{X}_{\text{sph}}$ by first sampling $y \sim \mathcal{N}(0,1)^n$ and then taking $x = y/\|y\|_2$. By Chebyshev, $\left| \|y\|_2^2 - n \right| < n^{3/4}$ with probability $1 - o(1)$.

(i) Supposing that $\|y\|_2^2 > n - n^{3/4}$, which occurs with probability $1 - o(1)$, we have

$$\Pr[|x_u| \geq n^{-1/3}] \leq \Pr[|y_u| \geq n^{1/6}\sqrt{1 - n^{-1/4}}] \leq e^{-n^{1/3}(1-n^{-1/4})/2} = o(1/n),$$

so that with probability $1 - o(1)$, we have for all u, $|x_u| < n^{-1/3}$.

(ii) We have $\|x\|_2 = 1$. For $q \in \{4, 6, 8\}$, $\|y\|_q^q$ has expectation $n(q-1)!!$ and variance

$$n[(2q-1)!! - ((q-1)!!)^2].$$

Supposing that $\|y\|_2^2 > n - n^{3/4} > n/2$, which occurs with probability $1 - o(1)$, we have for any α_q that

$$\begin{aligned} \Pr[\|x_q\| > \alpha_q n^{\frac{1}{q}-\frac{1}{2}}] &= \Pr[\|x\|_q^q > \alpha_q^q n^{1-\frac{q}{2}}] \\ &= \Pr[\|y\|_q^q > \alpha_q^q n^{1-\frac{q}{2}} \|y\|_2^q] \\ &\leq \Pr[\|y\|_q^q > \alpha_q^q 2^{-q/2} n] \\ &\leq \frac{n((2q-1)!! - ((q-1)!!)^2)}{n^2(2^{-q}\alpha_q^{2q} - (q-1)!!)^2}, \end{aligned}$$

by Chebyshev. This probability is $o(1)$ so long as we take $\alpha_q^{2q} > 2^q(q-1)!!$. □

Proposition 4.6. *Consider an i.i.d. prior $\mathcal{X} = \mathrm{iid}(\pi)$ where π is zero-mean and unit-variance with $\mathbb{E}[\pi^{16}] < \infty$. Then conditions (i) and (ii) in Assumption 4.3 are satisfied.*

An immediate implication of this is that conditions (i) and (ii) are also satisfied for a 'conditioned' prior which draws x from $\mathrm{iid}(\pi)$ but then outputs zero if a 'bad' event occured.

Proof. We have $x_i = \frac{1}{\sqrt{n}}\pi_i$ where π_i are independent copies of π. To prove (i),

$$\Pr[|x_i| \geq n^{-1/3}] = \Pr[|\pi_i| \geq n^{1/6}] = \Pr[\pi_i^8 \geq n^{4/3}] \leq \frac{\mathbb{E}[\pi^8]}{n^{4/3}} = \mathcal{O}(n^{-4/3})$$

using Markov's inequality and $\mathbb{E}[\pi^8] \leq 1 + \mathbb{E}[\pi^{16}] < \infty$. The proof follows by a union bound over all n coordinates.

To prove (ii), for $q \in \{2, 4, 6, 8\}$,

$$\begin{aligned} \Pr[\|x\|_q > \alpha_q n^{\frac{1}{q}-\frac{1}{2}}] &= \Pr[\|x\|_q^q > \alpha_q^q n^{1-\frac{q}{2}}] = \Pr\left[\sum_i x_i^q > \alpha_q^q n^{1-\frac{q}{2}}\right] \\ &= \Pr\left[\sum_i \pi_i^q > \alpha_q^q n\right] = \Pr\left[\sum_i \pi_i^q - n\mathbb{E}[\pi^q] > (\alpha_q^q - \mathbb{E}[\pi^q])n\right]. \end{aligned}$$

Choose α_q so that $C \equiv \alpha_q^q - \mathbb{E}[\pi^q] > 0$, and apply Chebyshev's inequality:

$$\leq \frac{\mathrm{Var}[\sum_i \pi_i^q]}{C^2 n^2} = \frac{n\mathrm{Var}[\pi^q]}{C^2 n^2} = \mathcal{O}(1/n).$$

Here we needed $\mathbb{E}[\pi^{2q}] < \infty$ so that $\mathrm{Var}[\pi^q] < \infty$. □

F Proof of pre-transformed PCA

In this section we prove our upper bound for the non-Gaussian Wigner model via pre-transformed PCA. We make the following assumptions on the spike prior $\mathcal{X}$ and the entrywise noise distribution $\mathcal{P}$.

Assumption 4.7. *Assumption on $\mathcal{X}$:*

(i) With probability $1 - o(1)$, all entries of x are small: $|x_i| \le n^{-1/2+\alpha}$ for some fixed $\alpha < \frac{1}{32}$.

Assumptions on $\mathcal{P}$:

(ii) $\mathcal{P}$ is a continuous distribution with a density function $p(w)$ that is three times differentiable.

(iii) $p(w) > 0$ everywhere.

(iv) Letting $f(w) = -p'(w)/p(w)$, we have that f and its first two derivatives are polynomially-bounded: there exists $C > 0$ and an even integer $m \ge 2$ such that $|f^{(\ell)}(w)| \le C + w^m$ for all $0 \le \ell \le 2$.

(v) With m as in (iv), $\mathcal{P}$ has finite moments up to $5m$: $\mathbb{E}|\mathcal{P}|^k < \infty$ for all $1 \le k \le 5m$.

An important consequence of assumptions (iv) and (v) is the following.

Lemma F.1. *$\mathbb{E}|f^{(\ell)}(\mathcal{P})|^q < \infty$ for all $0 \le \ell \le 2$ and $1 \le q \le 5$.*

Proof. Using $|a + b|^q \le |2a|^q + |2b|^q = 2^q(|a|^q + |b|^q)$ we have

$$\mathbb{E}|f^{(\ell)}(\mathcal{P})|^q \le \mathbb{E}|C + \mathcal{P}^m|^q \le 2^q(C^q + \mathbb{E}|\mathcal{P}|^{mq}) < \infty. \qquad \square$$

The main theorem of this section is the following.

Theorem 4.8. *Let $\lambda \ge 0$ and let $\mathcal{X}, \mathcal{P}$ satisfy Assumption 4.7. Let $\widehat{Y} = \sqrt{n}Y$ where Y is drawn from $\mathrm{Wig}(\lambda, \mathcal{X}, \mathcal{P})$. Let $f(\widehat{Y})$ denote entrywise application of the function $f(w) = -p'(w)/p(w)$ to $\widehat{Y}$, except the diagonal entries remain zero. Let*

$$F_{\mathcal{P}} = \mathbb{E}[f(\mathcal{P})^2] = \int_{-\infty}^{\infty} \frac{p'(w)^2}{p(w)} dw.$$

- *If $\lambda \le 1/\sqrt{F_{\mathcal{P}}}$ then $\frac{1}{\sqrt{n}}\lambda_{\max}(f(\widehat{Y})) \to 2\sqrt{F_{\mathcal{P}}}$ as $n \to \infty$.*
- *If $\lambda > 1/\sqrt{F_{\mathcal{P}}}$ then $\frac{1}{\sqrt{n}}\lambda_{\max}(f(\widehat{Y})) \to \lambda F_{\mathcal{P}} + \frac{1}{\lambda} > 2\sqrt{F_{\mathcal{P}}}$ as $n \to \infty$ and furthermore the top (unit-norm) eigenvector v of $f(\widehat{Y})$ correlates with the spike:*

$$\langle v, x\rangle^2 \ge \frac{(\lambda - 1/\sqrt{F_{\mathcal{P}}})^2}{\lambda^2} - o(1) \quad \textit{with probability } 1 - o(1).$$

Convergence is in probability. Here $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue of a matrix.

Note that Lemma F.1 implies that the expectation defining $F_{\mathcal{P}}$ is finite.

Proof. First we justify a local linear approximation of $f(\widehat{Y}_{ij})$. For $i \ne j$, define the error term $\mathcal{E}_{ij}$ by

$$f(\widehat{Y}_{ij}) = f(W_{ij}) + \lambda\sqrt{n}x_i x_j f'(W_{ij}) + \mathcal{E}_{ij}.$$

(Define $\mathcal{E}_{ii} = 0$.) We will show that the operator norm of $\mathcal{E}$ is small: $\|\mathcal{E}\| = o(\sqrt{n})$ with probability $1 - o(1)$. Apply the mean-value form of the Taylor approximation remainder: $\mathcal{E}_{ij} = \frac{1}{2}f''(W_{ij} + e_{ij})\lambda^2 n x_i^2 x_j^2$ for some $|e_{ij}| \le |\lambda\sqrt{n}x_i x_j|$. Bound the operator norm by the Frobenius norm:

$$\|\mathcal{E}\|^2 \le \|\mathcal{E}\|_F^2 = \frac{\lambda^4 n^2}{4} \sum_{i \ne j} x_i^4 x_j^4 f''(W_{ij} + e_{ij})^2 \le \frac{\lambda^4}{4} n^{8\alpha - 2} \sum_{i \ne j} f''(W_{ij} + e_{ij})^2.$$

Using the polynomial bound on f'' and the fact $|a+b|^k \le 2^k(|a|^k + |b|^k)$, we have

$$\begin{aligned} f''(W_{ij}+e_{ij})^2 &\le (C + (W_{ij}+e_{ij})^m)^2 \le 4C^2 + 4(W_{ij}+e_{ij})^{2m} \\ &\le 4C^2 + 4 \cdot 2^{2m}(W_{ij}^{2m} + e_{ij}^{2m}) \\ &\le 4C^2 + 2^{2m+2}(W_{ij}^{2m} + \lambda^{2m} n^{(4\alpha-1)m}) \\ &= 4C^2 + 2^{2m+2} W_{ij}^{2m} + o(1). \end{aligned}$$

Using finite moments of $W_{ij} \sim \mathcal{P}$, it follows that $\mathbb{E}\left[\sum_{i \ne j} f''(W_{ij}+e_{ij})^2\right] = \mathcal{O}(n^2)$, and so $\mathbb{E}\|\mathcal{E}\|^2 = \mathcal{O}(n^{8\alpha})$. Since $\alpha < \frac{1}{32}$, Markov's inequality now gives the desired result: with probability $1-o(1)$, $\|\mathcal{E}\|^2 = o(n^{1/4})$ and so $\|\mathcal{E}\| = o(\sqrt{n})$.

Our goal will be to show that $f(\widehat{Y})$ is, up to small error terms, another spiked Wigner matrix. Toward this goal we define another error term: for $i \ne j$, let $\Delta_{ij} = \lambda\sqrt{n} x_i x_j \left(f'(W_{ij}) - \mathbb{E}[f'(W_{ij})]\right)$, so that

$$f(\widehat{Y}_{ij}) = f(W_{ij}) + \lambda\sqrt{n} x_i x_j \mathbb{E}[f'(W_{ij})] + \mathcal{E}_{ij} + \Delta_{ij}. \tag{12}$$

(Define $\Delta_{ii} = 0$.) We will show that the operator norm of Δ is small: $\|\Delta\| = o(\sqrt{n})$ with probability $1-o(1)$. Let $A_{ij} = f'(W_{ij}) - \mathbb{E}[f'(W_{ij})]$ so that $\Delta_{ij} = \lambda\sqrt{n} x_i x_j A_{ij}$. (Define $A_{ii} = 0$.) We have $\|\Delta\| \le \lambda n^{-1/2+2\alpha}\|A\|$ because for any unit vector y,

$$\begin{aligned} y^\top \Delta y &= \sum_{i,j} \lambda\sqrt{n} x_i x_j A_{ij} y_i y_j \le \sum_{i,j} \lambda\sqrt{n} z_i A_{ij} z_j \qquad \text{where } z_i = x_i y_i \\ &\le \lambda\sqrt{n}\,\|A\| \cdot \|z\|^2 \le \lambda n^{-1/2+2\alpha}\|A\| \cdot \|y\| = \lambda n^{-1/2+2\alpha}\|A\|. \end{aligned}$$

Note that A is a Wigner matrix (i.e. a symmetric matrix with off-diagonal entries i.i.d.) and so $\|A\| = \mathcal{O}(\sqrt{n})$ with probability $1-o(1)$. This follows from Pizzo et al. [2013] Theorem 1.1, provided we can check that each entry of A has finite fifth moment. But this follows from Lemma F.1:

$$\mathbb{E}|A_{ij}|^5 \le 2^5\left(\mathbb{E}|f'(W_{ij})|^5 + |\mathbb{E}[f'(W_{ij})]|^5\right) < \infty.$$

Now we have $\|\Delta\| = \mathcal{O}(n^{2\alpha}) = o(\sqrt{n})$ with probability $1-o(1)$ as desired.

From (12) we now have that, up to small error terms, $f(\widehat{Y})$ is another spiked Wigner matrix:

$$f(\widehat{Y}) = f(W) + \lambda\sqrt{n}\,\mathbb{E}[f'(\mathcal{P})]\,xx^\top + \mathcal{E} + \Delta - \delta$$

where (to take care of the diagonal) we define $f(W)_{ii} = 0$, $\delta_{ij} = 0$, and $\delta_{ii} = \lambda\sqrt{n}\,\mathbb{E}[f'(\mathcal{P})]x_i^2$. Note that the final error term δ is also small: $\|\delta\| \le \|\delta\|_F = \mathcal{O}(n^{2\alpha}) = o(\sqrt{n})$. We now have

$$\frac{1}{\sqrt{n}}\lambda_{\max}(f(\widehat{Y})) = \lambda_{\max}\left(\frac{1}{\sqrt{n}} f(W) + \lambda\,\mathbb{E}[f'(\mathcal{P})]\,xx^\top\right) + o(1)$$

and so the theorem follows from known results on the spectrum of spiked Wigner matrices, namely Theorem 1.1 from Pizzo et al. [2013]. We need to check the following details. First note that the Wigner matrix $f(W)$ has off-diagonal i.i.d. entries that are centered:

$$\mathbb{E}[f(W_{ij})] = \int_{-\infty}^{\infty} \frac{-p'(w)}{p(w)} p(w)dw = p(-\infty) - p(\infty) = 0.$$

Each off-diagonal entry of $f(W)$ has variance $\mathbb{E}[f(W_{ij})^2] = F_{\mathcal{P}}$. The rank-1 deformation $\lambda\,\mathbb{E}[f'(\mathcal{P})]\,xx^\top$ has top eigenvalue $\lambda\,\mathbb{E}[f'(\mathcal{P})] \cdot \|x\|^2$. Recall that $\|x\|^2 \to 1$ in probability. Also,

$$f'(w) = \frac{d}{dw}\frac{-p'(w)}{p(w)} = -\frac{p''(w)p(w) - p'(w)^2}{p(w)^2}$$

and so

$$\mathbb{E}[f'(\mathcal{P})] = \int_{-\infty}^{\infty} \left[-p''(w) + \frac{p'(w)^2}{p(w)} \right] dw = \int_{-\infty}^{\infty} \frac{p'(w)^2}{p(w)} dw = F_{\mathcal{P}}.$$

Therefore the top eigenvalue of the rank-1 deformation converges in probability to $\lambda F_{\mathcal{P}}$. By Lemma F.1, the entries of $f(W)$ have finite fifth moment.

The desired convergence of the top eigenvalue now follows. It remains to show that when $\lambda > 1/\sqrt{F_{\mathcal{P}}}$, the top eigenvalue of $f(\widehat{Y})$ correlates with the planted vector x. Let v be the top eigenvector of $f(\widehat{Y})$ with $\|v\| = 1$. From above we have

$$v^\top \left(\frac{1}{\sqrt{n}} f(\widehat{Y}) \right) v = v^\top \left(\frac{1}{\sqrt{n}} f(W) \right) v + \lambda F_{\mathcal{P}} \langle v, x \rangle^2 + o(1).$$

We know $\frac{1}{\sqrt{n}} f(\widehat{Y})$ has top eigenvalue $\lambda F_{\mathcal{P}} + 1/\lambda + o(1)$ and $\frac{1}{\sqrt{n}} f(W)$ has top eigenvalue $2\sqrt{F_{\mathcal{P}}} + o(1)$, which yields

$$\langle v, x \rangle^2 \geq \frac{1}{\lambda F_{\mathcal{P}}} (\lambda F_{\mathcal{P}} + 1/\lambda - 2\sqrt{F_{\mathcal{P}}}) - o(1) = \frac{(\lambda - 1/\sqrt{F_{\mathcal{P}}})^2}{\lambda^2} - o(1). \qquad \square$$

G Proof of Proposition 5.2: calculation of Wishart second moment

Proposition 5.2. *Let $\mathcal{X}$ be a spike prior. In distribution P_n, let a hidden spike x be drawn from $\mathcal{X}$, and let N independent samples y_i, $1 \leq i \leq N$, be revealed from the normal distribution $\mathcal{N}(0, I_{n \times n} + \beta x x^\top)$. In distribution Q_n, let N independent samples y_i, $1 \leq i \leq N$, be revealed from $\mathcal{N}(0, I_{n \times n})$. Then we have*

$$\mathbb{E}_{Q_n} \left[\left(\frac{\mathrm{d}P_n}{\mathrm{d}Q_n} \right)^2 \right] = \mathbb{E}_{x, x' \sim \mathcal{X}} \left[\left(1 - \beta^2 \langle x, x' \rangle^2 \right)^{-N/2} \right].$$

Proof. We first compute:

$$\begin{aligned} \frac{\mathrm{d}P_n}{\mathrm{d}Q_n}(y_1, \ldots, y_N) &= \mathbb{E}_{x' \sim \mathcal{X}} \left[\prod_{i=1}^{n} \frac{\exp(-\frac{1}{2} y_i^\top (I + \beta x'(x')^\top)^{-1} y_i)}{\sqrt{\det(I + \beta x'(x')^\top)} \exp(-\frac{1}{2} y_i^\top y_i)} \right] \\ &= \mathbb{E}_{x'} \left[\det(I + \beta x'(x')^\top)^{-N/2} \prod_{i=1}^{N} \exp\left(-\frac{1}{2} y_i^\top ((I + \beta x'(x')^\top)^{-1} - I) y_i \right) \right]. \end{aligned}$$

Note that $(I + \beta x'(x')^\top)^{-1}$ has eigenvalue $(1 + \beta |x'|^2)^{-1}$ on x' and eigenvalue 1 on the orthogonal complement of x'. Thus $(I + \beta x'(x')^\top)^{-1} - I = \frac{-\beta}{1 + \beta |x'|^2} x'(x')^\top$, and we have:

$$\begin{aligned} &= \mathbb{E}_{x'} \left[(1 + \beta |x'|^2)^{-N/2} \prod_{i=1}^{N} \exp\left(\frac{1}{2} \frac{\beta}{1 + \beta |x'|^2} y_i^\top x'(x')^\top y_i \right) \right] \\ &= \mathbb{E}_{x'} \left[(1 + \beta |x'|^2)^{-N/2} \prod_{i=1}^{N} \exp\left(\frac{1}{2} \frac{\beta}{1 + \beta |x'|^2} \langle y_i, x' \rangle^2 \right) \right]. \end{aligned}$$

Passing to the second moment, we compute:

$$\begin{aligned} \mathbb{E}_{Q_n} \left[\left(\frac{\mathrm{d}P_n}{\mathrm{d}Q_n} \right)^2 \right] &= \mathbb{E}_{P_n} \left[\frac{\mathrm{d}P_n}{\mathrm{d}Q_n} \right] \\ &= \mathbb{E}_{x, x'} \left[(1 + \beta |x'|^2)^{-N/2} \prod_{i=1}^{N} \mathbb{E}_{y_i \sim \mathcal{N}(0, I + \beta x x^\top)} \exp\left(\frac{1}{2} \frac{\beta}{1 + \beta |x'|^2} \langle y_i, x' \rangle^2 \right) \right]. \end{aligned}$$

Over the randomness of y_i, we have $\langle y_i, x'\rangle \sim \mathcal{N}(0, |x'|^2 + \beta\langle x, x'\rangle^2)$, so that the inner expectation is a moment generating function of a χ_1^2 random variable:

$$= \mathbb{E}_{x,x'}\left[(1+\beta|x'|^2)^{-N/2}\prod_{i=1}^{N}\left(1 - \frac{\beta}{1+\beta|x'|^2}(|x'|^2 + \beta\langle x, x'\rangle^2)\right)^{-1/2}\right]$$
$$= \mathbb{E}_{x,x'}\left[\left(1-\beta^2\langle x, x'\rangle^2\right)^{-N/2}\right]. \qquad \square$$

H Proofs of unboundedness of the Wishart second moment

In this section, we give proofs that the Wishart second moment diverges as $n \to \infty$ when certain lower bound conditions fail. We begin with a proof of the second part of the following theorem:

Theorem 5.3. *Let the spike prior $\mathcal{X}$ have rate function $f_{\mathcal{X}}$ which is finite on $(0,1)$.*

(i) Suppose that $\beta^2 < 1$, that $\beta^2/\gamma < (\lambda_{\mathcal{X}}^*)^2$, and that $f_{\mathcal{X}}(t) > \frac{-1}{2\gamma}\log(1-\beta^2 t)$ for all $t \in (0,1)$. Then $\mathrm{Wish}(\gamma, \beta, \mathcal{X}) \lhd \mathrm{Wish}(\gamma)$.

(ii) If $\beta^2 > 1$, or if $\beta^2/\gamma > (\lambda_{\mathcal{X}}^*)^2$, or if $f_{\mathcal{X}}(t) < \frac{-1}{2\gamma}\log(1-\beta^2 t)$ for some $t \in (0,1)$, then the second moment (2) is unbounded.

Proof of (ii). Suppose first that $\beta^2/\gamma > (\lambda_{\mathcal{X}}^*)^2$. We bound the second moment (2) as follows:

$$\mathbb{E}_{x,x'\sim\mathcal{X}}\left[(1-\beta^2\langle x, x'\rangle^2)^{-n/2\gamma}\right] = \mathbb{E}_{x,x'}\left[\exp\left(\frac{-n}{2\gamma}\log(1-\beta^2\langle x, x'\rangle^2)\right)\right]$$
$$\geq \mathbb{E}_{x,x'}\left[\exp\left(\frac{n\beta^2}{2\gamma}\langle x, x'\rangle^2\right)\right],$$

which is unbounded as $n \to \infty$ as $\beta^2/\gamma > (\lambda_{\mathcal{X}}^*)^2$.

Next, suppose that $f_{\mathcal{X}}(t) < \frac{-1}{2\gamma}\log(1-\beta^2 t)$ for some $t \in (0,1]$. Then there exists $\varepsilon > 0$ so that, for all sufficiently large n,

$$\frac{-1}{n}\log \Pr_{x,x'\sim\mathcal{X}}[\langle x, x\rangle^2 \geq t] \leq -\varepsilon - \frac{-1}{2\gamma}\log(1-\beta^2 t). \tag{13}$$

we bound the second moment as follows:

$$\mathbb{E}_{x,x'\sim\mathcal{X}}\left[(1-\beta^2\langle x, x'\rangle^2)^{-n/2\gamma}\right]$$
$$\geq \Pr[\langle x, x'\rangle^2 \geq t](1-\beta^2 t)^{-n/2\gamma}$$
$$= \exp\left(n\left(\frac{1}{n}\log\Pr[\langle x, x'\rangle^2 \geq t] - \frac{1}{2\gamma}\log(1-\beta^2 t)\right)\right) \geq \exp(n\varepsilon)$$

which is unbounded as $n \to \infty$.

Finally, suppose that $\beta^2 > 1$. Note that $\frac{-1}{2\gamma}\log(1-\beta^2 t)$ tends to infinity as $t \to \beta^{-2}$ from below, whereas $f_{\mathcal{X}}(t)$ can only become infinite for $t \geq 1$. Hence we must have $f_{\mathcal{X}}(t) < \frac{-1}{2\gamma}\log(1-\beta^2 t)$ for some $t < \beta^{-2}$, so that the second moment is unbounded by the argument above. $\square$

We now prove an unboundedness result matching the bound of Proposition 5.8:

Proposition 5.9. *Let $\mathcal{X} = \mathrm{iid}(\{\pm 1\})$. For $\gamma > \frac{1}{3}$, there exists $\beta^2 < \gamma$ for which the second moment (2) diverges. Further, whenever $\beta^2 > 1 - e^{-(2\log 2)\gamma}$, the second moment diverges.*

Proof. For the first assertion, note that if we take $\beta^2 = \gamma$, then from the series expansion (6), $f_{\mathcal{X}}(t) + \frac{1}{2\gamma}\log(1-\beta^2 t)$ has vanishing t^0 and t^1 coefficients and negative t^2 coefficient for $\gamma \geq \frac{1}{3}$. It follows that there exists some $t > 0$ for which this quantity is negative. By continuity, this statement remains true if we fix γ and decrease β a sufficiently small amount. The assertion now follows from Theorem 5.3(ii).

The condition on the second assertion is precisely that $f_{\mathcal{X}}(t) + \frac{1}{2\gamma}\log(1-\beta^2 t)$ is negative at $t = 1$, as $f_{\mathcal{X}}(1) = \log 2$. Hence this assertion follows also from Theorem 5.3(ii). $\square$

I Proof of Theorem 5.10: MLE for Wishart with finite prior

Note the following well-known large deviations behavior for χ^2 distributions, which follows from Cramér's theorem:

Lemma I.1. *For all $z < 1$ and $c > 0$,*

$$\lim_{p\to\infty} \frac{1}{p} \Pr\left[\chi_p^2 < zp\right] = \frac{1}{2}(1 - z + \log z).$$

We now prove the following theorem:

Theorem 5.10. *Let $\beta < 0$. Let $\mathcal{X}_n$ be any prior supported on at most c^n points, for some fixed c. Then there is a computationally inefficient procedure that distinguishes between the spiked Wishart model* $\mathrm{Wish}(\gamma, \beta, \mathcal{X}))$ *and the unspiked model* $\mathrm{Wish}(\gamma)$*, with $o(1)$ probability of error, whenever*

$$(-\beta) + \log(1 - (-\beta)) < -2\gamma \log c,$$

Proof. Given a matrix Y, consider the test statistic $T = \min_{x\in\operatorname{supp}\mathcal{X}_n} \frac{1}{n} x^\top Y x$. Under $Y \sim \mathrm{Wish}(\gamma, \beta, \mathcal{X})$ with true spike x^*, we have that $\frac{1}{n}(x^*)^\top Y x^* \sim (1+\beta)\chi^2_{n/\gamma}$, which converges in probability to $(1+\beta)/\gamma$. Hence, for all $\varepsilon > 0$, we have that $T < (1 + \beta + \varepsilon)/\gamma$ with probability $1 - o(1)$ under the spiked model $\mathrm{Wish}(\gamma, \beta, \mathcal{X})$.

Under the unspiked model, we have

$$\begin{aligned}
\Pr[T \le (1+\beta+\varepsilon)/\gamma] &\le \sum_{x\in\operatorname{supp}\mathcal{X}} \Pr[x^\top Y x > (1+\beta-\varepsilon)n/\gamma] \\
&\le c^n \Pr\left[\chi^2_{n/2\gamma} \le (1+\beta+\varepsilon)n/\gamma\right] \\
&= \exp\left(n\left(\log c + \frac{1}{n}\Pr\left[\chi^2_{n/2\gamma} \le (1+\beta+\varepsilon)n/\gamma\right]\right)\right).
\end{aligned}$$

This is $o(1)$ so long as

$$\begin{aligned}
0 &> \log c + \lim_{n\to\infty} \frac{1}{n} \Pr\left[\chi^2_{n/2\gamma} \le (1+\beta+\varepsilon)n/\gamma\right] \\
&= \log c + \frac{1 - (1+\beta+\varepsilon) + \log(1+\beta+\varepsilon)}{2\gamma} \quad \text{by Lemma I.1;} \\
-2\gamma \log c &> -\beta - \varepsilon + \log(1+\beta+\varepsilon).
\end{aligned}$$

We can choose such $\varepsilon > 0$ precisely under the hypothesis of this theorem.

Hence, by thresholding the statistic T at $(1+\beta+\varepsilon)/\gamma$, we obtain a hypothesis test that distinguishes $Y \sim \mathrm{Wish}(\gamma, \beta, \mathcal{X})$ from $Y \sim \mathrm{Wish}(\gamma)$, with probability $o(1)$ of error of either type. ☐

References

J. Banks, C. Moore, J. Neeman, and P. Netrapalli. Information-theoretic thresholds for community detection in sparse networks. In *29th Annual Conference on Learning Theory*, pages 383–416, june 2016a.

J. Banks, C. Moore, R. Vershynin, and J. Xu. Information-theoretic bounds and phase transitions in clustering, sparse PCA, and submatrix localization. *arXiv:1607.05222*, july 2016b.

DLMF. NIST Digital Library of Mathematical Functions. http://dlmf.nist.gov/, Release 1.0.11 of 2016-06-08. URL `http://dlmf.nist.gov/`. Online companion to Olver et al. [2010].

F. W. J. Olver, D. W. Lozier, R. F. Boisvert, and C. W. Clark, editors. *NIST Handbook of Mathematical Functions*. Cambridge University Press, New York, NY, 2010. Print companion to DLMF.

A. Pizzo, D. Renfrew, and A. Soshnikov. On finite rank deformations of Wigner matrices. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 49, pages 64–94. Institut Henri Poincaré, 2013.