The adaptive interpolation method for the Wigner spike model

Jean Barbier
Communication Theory Laboratory, EPFL, Switzerland

I. Setting and result

A. *The Wigner spike model, or "planted" Sherrington-Kirkpatrick model*

$$\boxed{\underline{\underline{y}} = \frac{\underline{x}^*(\underline{x}^*)^\intercal}{\sqrt{N}} + \underline{\underline{z}} \quad \text{(forgetting the diagonal)} \quad \Leftrightarrow \quad y_{ij} = \frac{x_i^* x_j^*}{\sqrt{N}} + z_{ij} \quad \text{for} \quad 1 \le i < j \le N} \tag{1}$$

with $X_i^* \sim \mathbb{P}_0$ for $1 \le i \le N$, $Z_{ij} = Z_{ji} \sim \mathcal{N}(0,1)$ for $1 \le i < j \le N$ all independently.

Notations: Matrices are doubly underlined, vectors simply underlined, scalars are not. Fixed realizations of random variables $\underline{\underline{y}}$, $\underline{x}^*$, $\underline{\underline{z}}$, etc are small letters. The associated random variables $\underline{\underline{Y}}$, $\underline{X}^*$, $\underline{\underline{Z}}$ are capital letters.

Problem: Infer $\underline{x}^*$ from the knowledge of $\underline{\underline{y}}$. Model of extraction of low-rank information from noisy data matrix, such as PCA.

Signal-to-noise ratio (per observation):

$$\text{SNR} = \mathbb{E}\Big[\Big(\frac{X_i^* X_j^*}{\sqrt{N}}\Big)^2\Big] \Big\backslash \mathbb{E}[Z_{ij}^2] = \frac{\mathbb{E}_{\mathbb{P}_0}[(X^*)^2]^2}{N} = \frac{\rho^2}{N}, \qquad \rho \equiv \mathbb{E}_{\mathbb{P}_0}[(X^*)^2] : \text{signal power} \tag{2}$$

High-dimensional regime: (relevant in "Big-data" applications)

$$\frac{\#\,\text{observations} \cdot \text{SNR}}{\#\,\text{parameters to infer}} = O(1) \quad \to \quad \text{Wigner spike model:} \quad \frac{N(N-1)/2 \cdot \rho^2/N}{N} = \frac{\rho^2}{2} + O(1/N) = O(1) \tag{3}$$

B. *(Optimal) Bayesian setting*

Posterior: Bayes-optimal setting: We assume that all hyper-parameters are known (here $\mathbb{P}_0$ and that the noise variance is 1):

$$\begin{aligned}\mathbb{P}(\underline{X}^* = \underline{x} | \underline{\underline{Y}} = \underline{\underline{y}}) = \mathbb{P}(\underline{x}|\underline{\underline{y}}) &\propto \prod_{i=1}^N \mathbb{P}_0(x_i) \prod_{i<j} \exp\Big\{-\frac{1}{2}\Big(y_{ij} - \frac{x_i x_j}{\sqrt{N}}\Big)^2\Big\} \\ &= \frac{1}{\mathcal{Z}} \mathbb{P}_0(\underline{x}) \prod_{i<j} \exp\Big\{-\Big(\frac{x_i^2 x_j^2}{2N} - \frac{x_i x_j x_i^* x_j^*}{N} - \frac{x_i x_j z_{ij}}{\sqrt{N}}\Big)\Big\}\end{aligned} \tag{4}$$

$$\mathcal{Z}(\underline{x}^*, \underline{\underline{z}}) = \mathcal{Z} \equiv \int d\mathbb{P}_0(\underline{x}) \prod_{i<j} \exp\Big\{-\Big(\frac{x_i^2 x_j^2}{2N} - \frac{x_i x_j x_i^* x_j^*}{N} - \frac{x_i x_j z_{ij}}{\sqrt{N}}\Big)\Big\} \tag{5}$$

Phase transitions and free energy: In inference we often observe first-order (i.e. discontinuous) phase transitions: Let $\hat{\underline{x}}_{\rm opt}(\underline{\underline{y}}, \rho) = \hat{\underline{x}}_{\rm opt} \equiv \operatorname{argmin}_{\hat{\underline{x}}} \mathbb{E}_{\underline{X}|\underline{\underline{y}}}[\|\hat{\underline{x}} - \underline{X}\|^2]$ be the minimum mean-square error estimator.

- *Information theoretic (i.e. optimal) threshold:* $\rho_{\rm IT}$ s.t. $\lim_{N\to\infty} \frac{1}{N} \hat{\underline{x}}_{\rm opt}^\intercal \underline{x}^*$ jumps from "low" to "high".
- *Algorithmic threshold:* $\rho_{\rm algo}$ s.t. $\lim_{N\to\infty} \frac{1}{N} \hat{\underline{x}}_{\rm algo}^\intercal \underline{x}^*$ jumps from "low" to "high".

The location of the phase transitions and the optimal achievable estimation error are contained in the *free energy*: $-\frac{1}{N}\ln \mathcal{Z}$, but intractable... Fortunately this object is self-averaging (i.e. concentrates on its mean) w.r.t. the problem realization as $N \to \infty$:

$$\boxed{\underline{\text{Averaged free energy:}}\ f_N \equiv -\frac{1}{N} \mathbb{E}_{\underline{X}^*, \underline{\underline{Z}}} \ln \mathcal{Z} \qquad \underline{\text{Mutual information:}}\ \frac{1}{N} I(\underline{\underline{Y}}; \underline{X}^*) = f_N + \frac{\rho^2}{4}} \tag{6}$$

Remark: A phase transition corresponds to a non-analyticity point of the free energy. The algorithmic threshold is not a phase transition from the thermodynamical point of view as the free energy is analytic at this point: It is a "dynamical phase transition", i.e. (it is conjectured that) only an exponential time algorithm may reach the equilibrium state, i.e. non-trivially estimate the planted signal $\underline{x}^*$, for $\rho \leq \rho_{\rm algo}$.

C. *The replica-symmetric formula*

***Theorem** 1.1 (Replica-symmetric variational formula: [1–4]):*

$$\boxed{\lim_{N\to\infty} f_N = \inf_{q\in[0,\rho]} \sup_{r\in[0,\rho]} f_{\rm RS}(q,r)} \tag{7}$$

$$\boxed{f_{\rm RS}(q,r) \equiv \frac{q\,r}{2} - \frac{q^2}{4} \underbrace{-\mathbb{E}_{X^*\sim\mathbb{P}_0, Z\sim\mathcal{N}(0,1)} \ln \int d\mathbb{P}_0(x) \exp\Big\{ - r\Big(\frac{x^2}{2} - xX^* - \frac{xZ}{\sqrt{r}}\Big)\Big\}}_{\tilde f(r):\ \text{averaged free energy of } Y=\sqrt{r}X^*+Z \text{ where } X^*\sim\mathbb{P}_0,\, Z\sim\mathcal{N}(0,1)}} \tag{8}$$

II. Proof by the adaptive interpolation method

Simple but quite powerful evolution of the Guerra-Toninelli interpolation method for spin glasses [5].

A. *Interpolating model*

Define, for the sake of the proof, the following (random) observation model:

$$\boxed{\begin{cases} Y_{ij}(t) = \frac{X_i^* X_j^*}{\sqrt{N}}\sqrt{1-t} + Z_{ij} & 1 \leq i < j \leq N \\ \tilde Y_i(t) = X_i^* \sqrt{\int_0^t r(s)ds} + \tilde Z_i & 1 \leq i \leq N \end{cases}} \tag{9}$$

$Z_{ij} = Z_{ji} \sim \mathcal{N}(0,1)$, $X_i \sim \mathbb{P}_0$ all independent and $t \in [0,1]$: interpolation parameter, $r : [0,1] \mapsto [0,\rho]$: interpolating function.

$$\begin{aligned} \mathbb{P}(\underline{x}|\underline{\underline{Y}}(t), \underline{\tilde Y}(t)) &\propto \prod_{i=1}^N \mathbb{P}_0(x_i) \prod_{i<j} \exp\Big\{-\frac{1}{2}\Big(Y_{ij}(t) - x_i x_j \sqrt{\frac{1-t}{N}}\Big)^2\Big\} \prod_{i=1}^N \exp\Big\{-\frac{1}{2}\Big(\tilde Y_i(t) - x_i\sqrt{\textstyle\int_0^t r(s)ds}\Big)^2\Big\} \\ &= \frac{1}{\mathcal{Z}(t)} \mathbb{P}_0(\underline{x}) \prod_{i<j} \exp\Big\{ -(1-t)\Big(\frac{x_i^2 x_j^2}{2N} - \frac{x_i x_j X_i^* X_j^*}{N} - \frac{x_i x_j Z_{ij}}{\sqrt{N}\sqrt{1-t}}\Big)\Big\} \\ &\quad \times \prod_{i=1}^N \exp\Big\{ - \textstyle\int_0^t r(s)ds \Big(\frac{x_i^2}{2} - x_i X_i^* - \frac{x_i \tilde Z_i}{\sqrt{\int_0^t r(s)ds}}\Big)\Big\} \end{aligned} \tag{10}$$

<u>*Interpolating averaged free energy:*</u>

$$\boxed{f_N(t) \equiv -\frac{1}{N}\mathbb{E}_{\underline{X}^*,\underline{\underline{Z}},\underline{\tilde Z}} \ln \mathcal{Z}(t) \quad \to \quad \begin{cases} f_N(t=0) = f_N \\ f_N(t=1) = \tilde f(\int_0^1 r(t)dt) \end{cases}} \tag{11}$$

B. *Adaptive interpolation*

$$f_N(t=0) = f_N(t=1) - \int_0^1 f_N'(t)dt \quad \to \quad f_N = \tilde f\Big(\int_0^1 r(t)dt\Big) - \int_0^1 f_N'(t)dt \tag{12}$$

$$f_N'(t) = \mathbb{E}\langle g(\underline{X}, \underline{X}^*)\rangle_t \quad \text{for some function } g\text{, with} \quad \langle g(\underline{X}, \underline{X}^*)\rangle_t \equiv \int g(\underline{x}, \underline{X}^*)\, \mathbb{P}(\underline{x}|\underline{\underline{Y}}(t), \underline{\tilde Y}(t))\, d\underline{x} \tag{13}$$

$\mathbb{E}$ is the expectation w.r.t. the quenched variables $\underline{X}^*$, $\underline{\underline{Y}}(t)$, $\underline{\tilde Y}(t)$ generated from (9), or equivalently w.r.t. $\underline{X}^*$, $\underline{\underline{Z}}$, $\underline{\tilde Z}$.

<u>*Nishimori identity:*</u> This is where the Bayes optimality is crucial:

$$X^* \to Y \to X : \mathbb{E}_{X^*}\mathbb{E}_{Y|X^*}\mathbb{E}_{X|Y} g(X,X^*) = \mathbb{E}_Y \underbrace{\mathbb{E}_{X^*|Y}}_{\text{same}} \underbrace{\mathbb{E}_{X|Y}}_{\text{same}} g(X,X^*) = \mathbb{E}_Y \mathbb{E}_{X'|Y}\mathbb{E}_{X|Y} g(X,X')$$

$$\Leftrightarrow \quad \boxed{\mathbb{E}\langle g(X,X^*)\rangle = \mathbb{E}\langle g(X,X')\rangle} \tag{14}$$

X, X' two i.i.d. "replicas" drawn from $\mathbb{P}(\cdot|Y)$. Thus replicas are independent *given* Y.

$$\Rightarrow f'_N(t) = \frac{1}{4}\mathbb{E}\langle Q^2\rangle_t - \frac{1}{2}\mathbb{E}\langle Q\rangle_t\, r(t) \quad \text{with the } \textit{overlap} \quad Q \equiv \frac{1}{N}\underline{X}^\intercal \underline{X}^* \quad \text{where} \quad \underline{X} \sim \mathbb{P}(\cdot|\underline{\underline{Y}}(t), \underline{\tilde{Y}}(t)) \tag{15}$$

Fundamental sum rule:

$$f_N = \tilde{f}\Big(\int_0^1 r(t)dt\Big) - \frac{1}{4}\int_0^1 \Big\{\mathbb{E}\langle Q^2\rangle_t - 2\,\mathbb{E}\langle Q\rangle_t\, r(t)\Big\}dt \tag{16}$$

$$\Rightarrow \boxed{f_N = \tilde{f}\Big(\int_0^1 r(t)dt\Big) - \frac{1}{4}\int_0^1 \Big\{\big(\mathbb{E}\langle Q\rangle_t\big)^2 - 2\,\mathbb{E}\langle Q\rangle_t\, r(t)\Big\}dt + o_N(1)} \tag{17}$$

Self-averaging/concentration of overlap: This is called "replica-symmetric behavior" in physics:

$$\boxed{Q = \mathbb{E}\langle Q\rangle_t + o_N(1), \qquad \lim_{N\to\infty} o_N(1) = 0} \tag{18}$$

Two type of fluctuations must be controlled (see [3] for a generic proof for inference). This requires a slight perturbation of the model "a la Ghirlanda-Guerra" but *that maintains the Nishimori/Bayes-optimality property*, i.e. it must come from an inference problem with known hyper-parameters; an additional "side-channel": $\hat{\underline{Y}} = \sqrt{\epsilon_N}\,\underline{X}^* + \hat{\underline{Z}}$, $\hat{Z}_i \sim \mathcal{N}(0,1)$, $\epsilon_N \to 0$.

- "Thermal" fluctuations $\mathbb{E}\langle (Q - \langle Q\rangle_t)^2\rangle_t \overset{N\to\infty}{\to} 0$: Follows from the concavity + continuity in ϵ_N of f_N.
- "Quenched" fluctuations $\mathbb{E}[(\langle Q\rangle_t - \mathbb{E}\langle Q\rangle_t)^2] \overset{N\to\infty}{\to} 0$: Follows from the concavity in ϵ_N of f_N + Nishimori identity that allows to relate the overlap quenched fluctuations to the fluctuations of the free energy $\mathbb{E}[(\frac{1}{N}\ln\mathcal{Z} - \frac{1}{N}\mathbb{E}\ln\mathcal{Z})^2] \overset{N\to\infty}{\to} 0$. Very similar in spirit to the derivation of the Ghirlanda-Guerra identities in spin glasses [6].

Optimal interpolation path: We want the RS formula to appear: *Choose* in (17) $r(t)$ such that, for some fixed $q \in [0,\rho]$,

$$\big(\mathbb{E}\langle Q\rangle_t\big)^2 - 2\,\mathbb{E}\langle Q\rangle_t\, r(t) = q^2 - 2\,q\,r(t) \quad \Leftrightarrow \quad \boxed{r(t) = \frac{q + \mathbb{E}\langle Q\rangle_t}{2} \in [0,\rho]} \tag{19}$$

We recognize a (parametric in q) 1st order differential equation written in integral form (i.e. over $\int_0^t r(s)ds$):

$r(t) = g\big(\int_0^t r(s)ds, t; q\big)$ with $(\int_0^t r(s)ds)_{t=0} = 0$ and $g\big(\int_0^t r(s)ds, t; q\big) \equiv \dfrac{q + \mathbb{E}\langle Q\rangle_t}{2}$ ($\mathbb{E}\langle Q\rangle_t$ depends on $\int_0^t r(s)ds$, t)

By the Cauchy-Lipschitz theorem it possesses a unique solution $\mathcal{C}^0$ in t and q:

$$r^{(q)} : [0,1] \mapsto [0,\rho] \tag{20}$$

(Non-variational) single-letter formula: We obtain for *any* $q \in [0,\rho]$

$$f_N = \tilde{f}\Big(\int_0^1 r^{(q)}(t)dt\Big) - \frac{q^2}{4} + \frac{q}{2}\int_0^1 r^{(q)}(t)dt + o_N(1) \tag{21}$$

$$\Rightarrow \boxed{f_N = f_{\rm RS}\Big(q, \underbrace{\int_0^1 r^{(q)}(t)dt}_{R(q)\in[0,\rho]}\Big) + o_N(1)} \tag{22}$$

C. Matching bounds

The two bounds are obtained starting from (22).

Upper bound: For *any* $q \in [0,\rho]$

$$f_N \le \sup_{r\in[0,\rho]} f_{\rm RS}(q,r) + o_N(1) \quad \Rightarrow \quad \boxed{\limsup_{N\to\infty} f_N \le \inf_{q\in[0,\rho]}\sup_{r\in[0,\rho]} f_{\rm RS}(q,r)} \tag{23}$$

Lower bound: Assume $\exists$ a map $\mathcal{Q}$ s.t. $\mathcal{Q}\circ R:[0,\rho]\mapsto[0,\rho]$ is $\mathcal{C}^0$: It thus admits a fixed point $q^*=\mathcal{Q}(R(q^*))=\mathcal{Q}(r^*)$, where $r^*\equiv R(q^*)$. Using q^* in (22):

$$\begin{aligned}
f_N &= f_{\rm RS}(q^*,r^*)+o_N(1)\\
&= f_{\rm RS}(\mathcal{Q}(r^*),r^*)+o_N(1)\\
&\overset{(A)}{=} \sup_{r\in[0,\rho]} f_{\rm RS}(\mathcal{Q}(r^*),r)+o_N(1)\\
&\geq \inf_{q\in[0,\rho]}\sup_{r\in[0,\rho]} f_{\rm RS}(q,r)+o_N(1)\\
\Rightarrow\ & \boxed{\liminf_{N\to\infty} f_N \geq \inf_{q\in[0,\rho]}\sup_{r\in[0,\rho]} f_{\rm RS}(q,r)}
\end{aligned} \tag{24}$$

It remains to show what $\mathcal{Q}$ is, and that equality (A) stands:

$$\mathcal{Q}: r\in[0,\rho]\mapsto 2\tilde f(r)\in[0,\rho] \quad \text{concave (thus } \mathcal{C}^0)$$

$$\frac{d}{dr}f_{\rm RS}(\mathcal{Q}(r^*),r)=\frac{1}{2}\big(\mathcal{Q}(r^*)-\mathcal{Q}(r)\big) \quad\Rightarrow\quad \text{MAX attained, as } \mathcal{Q}(r) \text{ concave, at } r^*=r\in[0,\rho] \text{ and thus } (A) \text{ stands}$$

$\square$

Remarks and extensions:

- The method only requires what is believed to be the strict minimum for replica-symmetric formulas to be valid: Concentration of the overlap.
- It does not require any sign for some "remainder" as usually the case in the canonical interpolation method, as the remainder is directly canceled.
- Method developed in [3] with application to the Wigner spike model, random linear estimation and symmetric tensor estimation. Then applied in [7] to general tensor estimation, in [8] to generalized linear models, in [9] to random linear estimation with structured matrices, in [10, 11] to models of multi-layer neural networks and in [12] to inference for sparse models (in this case the censored block model, i.e. a simpler version of the stochastic block model, or a particular low density generator matrix code).

Two important open questions:

- How to move away from the Bayes optimal setting (i.e. from the Nishimori line) for inference/planted problems?
- How to extend the method to problems with replica symmetry breaking, i.e. no concentration of the overlap? E.g. combinatorial optimization and spin glasses.

References

[1] J. Barbier, M. Dia, N. Macris, F. Krzakala, T. Lesieur, and L. Zdeborová, "Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula," in Advances in Neural Information Processing Systems (NIPS) 29, 2016, pp. 424–432.

[2] M. Lelarge and L. Miolane, "Fundamental limits of symmetric low-rank matrix estimation," Probability Theory and Related Fields, Apr 2018. [Online]. Available: https://doi.org/10.1007/s00440-018-0845-x

[3] J. Barbier and N. Macris, "The adaptive interpolation method: A simple scheme to prove replica formulas in bayesian inference," CoRR, vol. abs/1705.02780, 2017. [Online]. Available: http://arxiv.org/abs/1705.02780

[4] A. E. Alaoui and F. Krzakala, "Estimation in the spiked wigner model: A short proof of the replica formula," CoRR, vol. abs/1801.01593, 2018.

[5] F. Guerra and F. L. Toninelli, "The thermodynamic limit in mean field spin glass models," Communications in Mathematical Physics, vol. 230, no. 1, pp. 71–79, 2002.

[6] S. Ghirlanda and F. Guerra, "General properties of overlap probability distributions in disordered spin systems. towards parisi ultrametricity," Journal of Physics A: Mathematical and General, vol. 31, no. 46, p. 9149, 1998.

[7] J. Barbier, N. Macris, and L. Miolane, "The Layered Structure of Tensor Estimation and its Mutual Information," in 47th Annual Allerton Conference on Communication, Control, and Computing (Allerton), Sep. 2017.

[8] J. Barbier, F. Krzakala, N. Macris, L. Miolane, and L. Zdeborová, "Phase transitions, optimal errors and optimality of message-passing in generalized linear models," arXiv preprint arXiv:1708.03395, 2017.

[9] J. Barbier, N. Macris, A. Maillard, and F. Krzakala, "The Mutual Information in Random Linear Estimation Beyond i.i.d. Matrices," in IEEE International Symposium on Information Theory (ISIT), 2018.

[10] M. Gabrié, A. Manoel, C. Luneau, J. Barbier, N. Macris, F. Krzakala, and L. Zdeborová, "Entropy and mutual information in models of deep neural networks," arXiv preprint arXiv:1805.09785, 2018.

[11] B. Aubin, A. Maillard, J. Barbier, F. Krzakala, N. Macris, and L. Zdeborová, "The committee machine: Computational to statistical gaps in learning a two-layers neural network," arXiv preprint arXiv:1806.05451, 2018.

[12] J. Barbier, C. L. Chan, and N. Macris, "Adaptive path interpolation for sparse systems: Application to a simple censored block model," arXiv preprint arXiv:1806.05121, 2018.