On the TAP equation for the perceptron

Erwin Bolthausen, University of Zurich

Ilya70-Fest, Dec 22, 2017

The perceptron: Neural net. Gardner (partly with Derrida) 1987-88 had results on the memory capacity, based on non-rigorous replica computations. The simplest case: $H_k, 1 \leq k \leq M$, random half spaces in $\mathbb{R}^N$

$$H_k := \left\{ x \in \mathbb{R}^N : \sum_{i=1}^N x_i J_{ik} \geq 0 \right\}, \ J_{ik} \text{ indep. standard Gaussians}$$

For $\alpha > 0$, they obtained a formula for

$$f(\alpha) := \lim_{N \to \infty} \frac{1}{N} \log \left| \bigcap_{k=1}^{M=\alpha N} H_k \cap \Sigma_N \right|, \quad \Sigma_N := \{-1, 1\}^N .$$

Soft version: $u : \mathbb{R} \to \mathbb{R}$

$$Z_{N,u,\alpha} = \sum_{\sigma \in \Sigma_N} 2^{-N} \exp \left[\sum_{k=1}^{\alpha N} u \left(N^{-1/2} \sum_{i=1}^N \sigma_i J_{ik} \right) \right],$$

$$f_u(\alpha) := \lim_{N \to \infty} \frac{1}{N} \log Z_{N,u,\alpha}.$$

For the random half spaces: $u(x) = -\infty 1_{x<0}$.

Theorem (Talagrand, Shcherbina-Tirozzi). For $u : \mathbb{R} \to \mathbb{R} \cup \{-\infty\}$, bounded above, and α small enough

$$\frac{1}{N} \log Z_{N,u,\alpha} \to \mathrm{RS}(\alpha, u), \text{ a.s.}$$

with

$$\mathrm{RS}(\alpha, u) := -\frac{r}{2}(1-q) + E_Z \log\cosh\left(\sqrt{\alpha r} Z\right) + \alpha E_Z \log E_{Z'} u\left(\sqrt{q} Z + \sqrt{1-q} Z'\right),$$

where Z, Z' are standard Gaussians, and $r = r(\alpha, u)$ and $q = q(\alpha, u)$ solve

$$q = E \tanh^2\left(\sqrt{\alpha r} Z\right), \; r = E\psi_q^2(\sqrt{q} Z), \; \psi_q(x) := \frac{1}{\sqrt{1-q}} \frac{EZ \exp\left[u\left(x + \sqrt{1-q} Z\right)\right]}{E \exp\left[u\left(x + \sqrt{1-q} Z\right)\right]}$$

Aim: Give a proof based on the Thouless-Anderson-Palmer approach (originally proposed for SK).

Background: **Curie-Weiss**

$$\begin{aligned} \mathsf{GIBBS}_{\beta,h,N}(\sigma) &= \frac{2^{-N}}{Z_{N,\beta,h}} \exp\left[\frac{\beta}{2}\sum\nolimits_{i,j=1}^{N}\sigma_i\sigma_j + h\sum\nolimits_{i=1}^{N}\sigma_i\right] \\ &= \frac{2^{-N}}{Z_{N,\beta,h}} \exp\left[\frac{N\beta}{2}\bar{\sigma}^2 + hN\bar{\sigma}\right],\ \bar{\sigma} := \frac{1}{N}\sum\nolimits_{i=1}^{N}\sigma_i \end{aligned}$$

$$P^{\text{coin toss}}(\bar{\sigma} \sim x) = \exp\left[-NI(x)\right],$$

and then

$$\lim_{N\to\infty}\frac{1}{N}\log Z_N = \sup_x\left(\frac{\beta}{2}x^2 + hx - I(x)\right).$$

The sup is attained at an $x = m$ which solves

$$m = \tanh(\beta m + h).$$

For $h \neq 0$ and for $h = 0,\ \beta \leq 1$: Unique maximizer m and $\mathsf{GIBBS}(\bar{\sigma} \approx m) \approx 1$.

SK-Model with external field: Random Gibbs measure

$$\mathsf{GIBBS}(\sigma) := \frac{2^{-N}}{Z_{N,\beta,h}} \exp\left[\frac{\beta}{\sqrt{N}} \sum_{1\leq i<j\leq N} J_{ij}\sigma_i\sigma_j + h\sum_{i=1}^{N}\sigma_i\right],$$

J_{ij} i.i.d. standard Gaussians, defined on $(\Omega, \mathcal{F}, \mathbb{P})$, $\sigma \in \{-1,1\}^N$, $h \in \mathbb{R}$.

TAP equations for the Gibbs expectation $m_i := \langle \sigma_i \rangle$

$$m_i \approx \tanh\left(h + \frac{\beta}{\sqrt{N}}\sum_j J_{ij}m_j \underbrace{-\beta^2(1-q)\,m_i}_{\text{Onsager correction}}\right),\ J_{ij} = J_{ji},\ J_{ii} = 0$$

with q the unique fixed point of

$$q = q(\beta,h) = E\tanh^2\left(h + \sqrt{q}\beta Z\right),\ Z \text{ standard Gaussian}$$

Mathematical proofs for high temperature: Talagrand, Chatterjee. Low temperature recently by Auffinger-Jagannath.

Heuristic derivation by belief propagation: *If the model would be defined on a tree graph* (instead of the complete graph):
Equation for the marginals ν_i

$$\nu_i(\sigma_i) \sim \sum_{(\sigma_j)_{j\in\partial i}} \exp\left[h\sigma_i + \beta\sigma_i \frac{1}{\sqrt{N}} \sum_{j\in\partial i} J_{ij}\sigma_j\right] \prod_{j\in\partial i} \nu_j^{\text{cut } i}\left(\sigma_j\right),$$

where $\sim$ means equality up to normalization, and $\nu_j^{\text{cut } i}$ the j-th marginal *cutting* the connection with i. Therefore

$$m_i = \sum_{\sigma_i=\pm1} \sigma_i \nu_i(\sigma_i) = \frac{\sum_{(\sigma_j)_{j\in\partial i}} \sinh\left(h + \beta N^{-1/2} \sum_{j\in\partial i} J_{ij}\sigma_j\right) \prod_{j\in\partial i} \nu_j^{\text{cut } i}\left(\sigma_j\right)}{\sum_{(\sigma_j)_{j\in\partial i}} \cosh\left(h + \beta N^{-1/2} \sum_{j\in\partial i} J_{ij}\sigma_j\right) \prod_{j\in\partial i} \nu_j^{\text{cut } i}\left(\sigma_j\right)}$$

By a CLT, if $|\partial i|$ is large, for $\sum_{j\in\partial i} J_{ij}\sigma_j$ under $\prod_{j\in\partial i} \nu_j^{\text{cut } i}\left(\sigma_j\right)$, this is

$$m_i \approx \tanh\left(h + \beta N^{-1/2} \sum_{j:j\neq i} J_{ij} m_j^{\text{cut } i}\right), \; m_j^{\text{cut } i} := \left\langle \sigma_j \right\rangle_{\nu_j^{\text{cut } i}}$$

The argument is that the formula is true at high temperature (as $N \to \infty$), and also, by more complicated arguments in MPV, at low temperature if m_i is the mean inside a "pure state". Expanding the difference of $m_j - m_j^{\text{cut } i}$ leads to the **Onsager-correction** $-\beta^2 (1-q)\, m_i$, and the final form

$$m_i \approx \tanh\left(h + \frac{\beta}{\sqrt{N}} \sum_j J_{ij} m_j - \beta^2 (1-q)\, m_i\right).$$

In contrast to standard mean-field models, it was considered to be difficult to construct directly solutions even in high temperature: Bray and Moore 1979-1982, Nemoto-Takayama 1985, but the iterations behaved badly (see also the discussion in Mézard-Parisi-Virasoro).

Iterative construction: (B. CMP 2014): Define $m_i^{[k]},\ 1 \leq i \leq N,\ k \geq 0$:

$$m_i^{[0]} := 0,\ m_i^{[1]} := \sqrt{q},$$

$$m_i^{[k+1]} := \tanh\left(h + \beta N^{-1/2} \sum_j J_{ij} m_j^{[k]} - \beta^2 (1-q)\, m_i^{[k-1]}\right),\ k \geq 1.$$

Theorem

$$\lim_{k,l\to\infty} \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \left(m_i^{[k]} - m_i^{[l]}\right)^2 = 0,\ \text{a.s.}$$

holds if and only if the **de Almeida–Thouless condition** holds:

$$\beta^2 E \cosh^{-4}\left(h + \beta\sqrt{q} Z\right) \leq 1,\ Z \text{ standard Gauss.}$$

Basis of the proof: **Structure theorem** for the iterations

$$m_i^{[k+1]} \simeq \tanh\left(h + \eta_i^{[k]} + \beta \sum_{r=1}^{k-1} \gamma_r \xi_i^{[r]}\right), \; N \text{ large},$$
$$\eta_i^{[k]} := \frac{\beta}{\sqrt{N}} \sum_{j=1}^{N} J_{ij}^{[k]} m_j^{[k]}$$

where the rv $\xi^{[r]}$, the random matrices $J^{[k]}$, and $\gamma_r \in \mathbb{R}^+$ are recursively defined, $\xi_i^{[1]} = N^{-1/2} \sum_j J_{ij}$.

- The ξ's don't change with the iteration.

- The $\xi_i^{[1]}, \ldots, \xi_i^{[k-1]}, \eta_i^{[k]}$ are independent (asymptotically as $N \to \infty$),

- $\eta^{[k]}$ and $\eta^{[k+1]}$ are independent.

Key: Iterative construction of the $J^{[k]}$. First steps:
$m_i^{[2]} = \tanh\left(h + \beta\sqrt{q}\xi_i^{[1]}\right)$, where $\xi_i^{[1]} = N^{-1/2}\sum_j J_{ij}$. In

$$m_i^{[3]} = \tanh\left(h + \frac{\beta}{\sqrt{N}}\sum_j J_{ij}m_j^{[2]} - \beta\sqrt{q}\,(1-q)\right)$$

we make J independent of $m^{[2]}$, i.e. independent of $\xi^{[1]}$ by $J_{ij}^{[2]} = J_{ij}$−lin comb of ξ′s.

$$m_i^{[3]} \approx \tanh\left(h + \frac{\beta}{\sqrt{N}}\sum_j J_{ij}^{[2]}m_j^{[2]} + \gamma_1\xi_i^{[1]}\right)$$

For $m^{[4]}$ one does $J \to J^{[2]} \to J^{[3]}$. After the first: $N^{-1/2}\sum_j J_{ij}^{[2]}m_j^{[3]}$. $J^{[2]} \to J^{[3]}$ is done *conditionally* on $\xi^{[1]}$, so that $J^{[3]}$ becomes conditionally independent of $m^{[3]}$. This lead to

$$m_i^{[4]} \approx \tanh\left(h + \frac{\beta}{\sqrt{N}}\sum_j J_{ij}^{[3]}m_j^{[3]} + \gamma_1\xi_i^{[1]} + \gamma_2\xi_i^{[2]}\right).$$

Miraculously $J \to \cdots \to J^{[k]}$ cancels Onsager, provided it comes with a shift two. The size of $\eta_i^{[k]} := N^{-1/2} \sum_j J_{ij}^{[k]} m_j^{[k]}$ can be computed, and it disappears ($N \to \infty$ first, then $k \to \infty$) iff the **AT-condition** holds.

In the low temperature region, $\eta^{[k]}$ stabilizes as $k \to \infty$ in distribution, but behaves chaotic with $k \to k+1$.

Free energy:

$$f(\beta, h) = \lim_{N \to \infty} \frac{1}{N} \log Z_{N,\beta,h} = \lim_{N \to \infty} \frac{1}{N} \mathbb{E} \log Z_{N,\beta,h}.$$

For $h = 0$ and $\beta \leq 1$ (Aizenman-Lebowitz-Ruelle, Fröhlich-Zegarlinski):

$$f(\beta, 0) = f_{\text{ann}}(\beta, 0) := \lim_{N \to \infty} \frac{1}{N} \log \mathbb{E} Z_{N,\beta,h} = \frac{\beta^2}{4},$$

and can be proved by a second moment method. For $h \neq 0$, and all $\beta > 0$

$$f(\beta, h) \neq f_{\mathsf{ann}}(\beta, h).$$

However, take (m_i) from TAP, $m_i = \tanh(h_i)$, put $p(\sigma) := \prod_i p_i(\sigma_i)$, $p_i(\sigma_i) = \frac{1}{2}\exp[h_i\sigma_i] / \cosh(h_i)$,

$$Z_N = \sum_\sigma 2^{-N} \exp[\cdot] = \prod_i \cosh(h_i) \underbrace{\sum_\sigma p(\sigma) \exp\left[\cdot - \sum_i h_i\sigma_i\right]}_{\hat{Z}_N}.$$

The first part is easy (quenched). For $\hat{Z}_N$ do a *conditional* quenched=annealed argument, i.e. analyze $\mathbb{E}\left(\hat{Z}_N \middle| \mathcal{F}\right)$, $\mathcal{F} := \sigma\left(\xi^{[1]}, \xi^{[2]}, \ldots\right)$. For $\mathbb{E}(\exp[\cdot] | \mathcal{F})$, one has to do the shift $J \to J^{[k]}$, k large. By a second moment method, one gets

$$f(\beta, h) = \mathsf{RS}(\beta, h) := E \log\cosh(h + \beta\sqrt{q}Z) + \frac{\beta^2(1-q)}{4}.$$

Unfortunately, the conditional second moment method does not work up to the AT-line.

Back to the **perceptron** which has the (non-Gaussian!) Hamiltonian

$$\sum_{k=1}^{M} u(S_k), \; S_k := N^{-1/2} \sum_{i=1}^{N} \sigma_i J_{ik}.$$

The key point for TAP equations (Mézard 1988, 2017 for the Hopfield model): Use a bipartite structure $(\sigma_i)_{i\leq N} \leftrightarrows (S_k)_{k\leq \alpha N}$. With $m_i := \langle \sigma_i \rangle, \; \rho_k := \langle u'(S_k) \rangle$.

$$\begin{aligned} \langle \sigma_i \rangle &=: m_i = \tanh\left(N^{-1/2} \textstyle\sum_{k=1}^{M=\alpha N} J_{ik}\rho_k - \alpha E \psi_q'(\sqrt{q}Z)\, m_i \right) \\ \rho_k &= \psi_q \left(N^{-1/2} \textstyle\sum_{i=1}^{N} m_i J_{ik} - (1-q)\,\rho_k \right), \end{aligned}$$

with

$$\psi_q(x) := \frac{1}{\sqrt{1-q}} \frac{EZ \exp\left[u\left(x+\sqrt{1-q}Z\right)\right]}{E \exp\left[u\left(x+\sqrt{1-q}Z\right)\right]}.$$

Remark: ψ_q for $q < 1$ is smooth without u being smooth!

For the iteration, take $m_i^{[0]} := 0,\ m_i^{[1]} := \sqrt{q},\ \rho_k^{[0]} = 0,\ \rho_k^{[1]} := \sqrt{r}$, and

$$\begin{aligned} m_i^{[n+1]} &= \tanh\left(N^{-1/2}\sum_{k=1}^{\alpha N} J_{ik}\rho_k^{[n]} - m_i^{[n-1]}\alpha E\psi_q'(\sqrt{q}Z)\right), \\ \rho_k^{[n+1]} &= \psi_q\left(N^{-1/2}\sum_{i=1}^{N} m_i^{[n]} J_{ik} - \rho_k^{[n-1]}(1-q)\right). \end{aligned}$$

The iterations lead to a similar structure theorem as in the SK case:

$$\begin{aligned} m_i^{[n+1]} &= \tanh\left(N^{-1/2}\sum_{k=1}^{\alpha N} J_{ik}^{[n]}\rho_k^{[n]} + \gamma_1\xi_i^{[1]} + \cdots + \gamma_{n-1}\xi_i^{[n-1]}\right), \\ \rho_k^{[n+1]} &= \psi_q\left(N^{-1/2}\sum_{i=1}^{N} J_{ik}^{[n]} m_i^{[n]} + \beta_1\eta_k^{[1]} + \cdots + \beta_{n-1}\eta_k^{[n-1]}\right). \end{aligned}$$

The iterates converge if and only if

$$\alpha E\frac{1}{\cosh^4(\sqrt{r}Z)}E\left[\psi_q'(\sqrt{q}Z)\right]^2 \leq 1$$

provided that the fixed point equations for (r, q) have a unique solution, which is easy for small α (Talagrand).

The "transformation of measure argument" with a conditional second moment argument leads to the Gardner formula (work in progress).

Summary:

- The iterative scheme for TAP type equations can be widely applied. This is also investigated by Mézard (2017) for the Hopfield model, and multi-layer perceptrons.

- It seems to identify precisely the high-temperature region for many models.

- For the free energy, it is less satisfactory, as a conditional second moment method does not work in the full high-temperature region.

- Main open problem: Extend the method to low temperature. SK is probably not the ideal model to try first.

Happy birthday, Ilya!