Strong replica symmetry for log-concave Gibbs measures

Jean Barbier (ICTP - Italy)
Dmitry Panchenko (U. Toronto - Canada)
and Manuel Sáenz (ICTP - Italy)

Inference problems: algorithms and lower bounds
Goethe-Universität Frankfurt
September 2020

Introduction

Overview

- For disordered systems with concave Hamiltonians
- we prove the concentration of multioverlaps,
- a representation of the asymptotic distribution of the spins,
- and asymptotic strong Gibbs decorrelation of spins.

(!) $\longrightarrow$ this symbol marks important points during the talk

Setting and notations

Bounded spin Gibbs measure

Let J be real valued r.v.s and $\sigma \in [-1;1]^N$ random vector with density

$$G_N(\sigma|J) := \frac{1}{\mathcal{Z}_N(J)} e^{\mathcal{H}_N(\sigma|J)},$$

where a.s. $\mathcal{H}_N \in \mathcal{C}^2(\Sigma_N)$ is concave w.r.t. σ and $\mathcal{Z}_N(J)$ normalisation.

Define the log partition function as

$$f_N := \mathbb{E} F_N := \mathbb{E} \log \mathcal{Z}_N(J).$$

We make the usual assumption that $\operatorname{Var} F_N / N \xrightarrow{N \to +\infty} 0$.

Setting and notations

Replicas := independent samples from Gibbs measure with same disorder J.

σ_i^l := i-th spin of l-th replica.

K := set of sets of finite integers.

Definition

For $k := (k_1, \ldots, k_n) \in K$, the associated *multioverlap* (m.o.) is

$$R^{(k)} := N^{-1} \sum_{i=1}^{N} (\sigma_i^1)^{k_1} \cdots (\sigma_i^n)^{k_n}.$$

- $n = 1$ and $k = 1$: *magnetisation*.
- $n = 2$ and $k = (1, 1)$: *overlap*.

High dimensional inference

$\boldsymbol{x}_0 \in \Sigma^N$ parameters to estimate and $\boldsymbol{y}(\boldsymbol{x}_0) \in (\Sigma')^M$ observations. We have

$$P(\boldsymbol{x}|\boldsymbol{y}) = \frac{1}{\mathcal{Z}} P(\boldsymbol{x}) P(\boldsymbol{y}|\boldsymbol{x}).$$

High dimension: we consider $N, M \to +\infty$, $N/M = \Theta(1)$.

The *Hamiltonian* is

$$\mathcal{H}(\boldsymbol{x}|\boldsymbol{y}) := \log(P(\boldsymbol{x}) P(\boldsymbol{y}|\boldsymbol{x})).$$

(!) *Objective:* estimate log partition $\log \mathcal{Z}$ or mutual information $I(\boldsymbol{x}_0; \boldsymbol{y})$. Both are related by a constant.

Empirical risk minimisation

$\mathcal{L} : \Sigma_N \times \Sigma'_M \to \mathbb{R}_{\geq 0}$ is a *loss*. Define the estimator

$$\boldsymbol{x} := \mathrm{argmin}_{\boldsymbol{x}' \in \Sigma_N} \mathcal{L}(\boldsymbol{x}', \boldsymbol{y}).$$

Let $\beta > 0$ be an *inverse temperature* and

$$P(\boldsymbol{x}|\boldsymbol{y}) = \frac{1}{\mathcal{Z}} e^{-\beta \mathcal{L}(\boldsymbol{x}, \boldsymbol{y})}.$$

Then $\mathcal{H}(\boldsymbol{x}|\boldsymbol{y}) := -\mathcal{L}(\boldsymbol{x}, \boldsymbol{y})$. $\boldsymbol{x}$ is recovered at zero temperature.

(!) The estimator can be studied through $\log \mathcal{Z}$.

Problems of interest

Regularised least squares linear regression:

$$\boldsymbol{x} := \operatorname{argmin}_{\boldsymbol{x}' \in \Sigma_N} \left|\left|\boldsymbol{A}\boldsymbol{x}' - \boldsymbol{y}\right|\right|_2^2 + f(\boldsymbol{x}'),$$

with $\boldsymbol{A} \to$ known matrix / $f \to$ convex function.

The associated Hamiltonian

$$\mathcal{H}(\boldsymbol{x}|\boldsymbol{y}) := -\left|\left|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\right|\right|_2^2 - f(\boldsymbol{x})$$

is then concave.

(!) This includes ridge ($f = ||\cdot||_1$) and LASSO ($f = ||\cdot||_2^2$).

Problems of interest

Generalised linear models:

$$\boldsymbol{y} = \phi(\boldsymbol{A}\boldsymbol{x}_0) + \boldsymbol{z},$$

with $\phi(\cdot)$ some function and $\boldsymbol{z}$ a normal vector of covariance $\Delta\mathbb{I}$.

If assumed model is linear model (mismatch)

$$\mathcal{H}(\boldsymbol{x}|\boldsymbol{y}) := -\log P_0(\boldsymbol{x}) - \Delta^{-1}\left|\left|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\right|\right|_2^2,$$

where $P_0 \to$ assumed prior. If $\log P_0(\cdot)$ is convex, $\mathcal{H}$ is concave.

(!) Here there are no Bayes-optimal identities.

Prior work

[1] connection between Bayes-optimal inference and M-estimators.

[2] many sparse inference problems.

[3] informational theoretical limit of a binary sparse model.

(!) The proof requires multioverlap concentration.

---

[1] *Advani & Ganguli (2016). Adv. in Neur. Inf. Proc. Sys.*
[2] *Coja-Oghlan et al. (2018). Advances in Mathematics, 333, 694-795.*
[3] *Barbier et al. (2019) arXiv preprint arXiv:1806.05121.*

Prior work

[4] two-layer neural network with a first randomly weighted layer.

[5] Empirical Risk Minimisation applied to GLM.

[6] regularised Empirical Risk Minimisation for GLM data.

[4] *Mei & Montanari (2019). arXiv preprint arXiv:1908.05355.*
[5] *Aubin et al. (2020). arXiv preprint arXiv:2006.06560.*
[6] *Taheri et al. (2020). arXiv preprint arXiv:2006.08917.*

Prior work

[7] and [8] regularised least squares for general feature matrices more general than Gaussian.

[9] m.o. concentration for Bayes-optimal inference.

(!) Many approaches based on Gordon min-max, interpolation and cavity methods.

---

[7] *Gerbelot et al. (2020). arXiv preprint arXiv:2006.06581.*
[8] *Gerbelot et al. In Conference on Learning Theory (pp. 1682-1713).*
[9] *Barbier & Panchenko (2020). arXiv preprint arXiv:2005.03115.*

Main results

Perturbed model

(!) Perturbations give "good properties" that ensure concentrations.

Gaussian perturbation

We add a *ridge regularisation* term

$$\mathcal{H}_N^{\text{gauss}}(\sigma) := -\frac{\epsilon_N}{2} \|\sigma\|^2 ,$$

with $\epsilon_N \to 0$ and $N\epsilon_N \to +\infty$.

(!) This forces m.o. to concentrate w.r.t. Gibbs measure.

Perturbed model

For $I \in \mathcal{I}$, consider polynomials $P_I : [-1;1] \to [0;1]$ s.t.

$$P_I(x) := \sum_{p=0}^{m-1} a_p (x+1)^p.$$

The P_I are convex on $[-1;1]$.

The definition of $\mathcal{I}$ makes $\sum_{I \in \mathcal{I}} P_I$ uniformly summable and the coefficients accumulate at 0.

Also, $P_I(x) \in [0,1]$.

Perturbed model

$\pi := (\pi_I)_{I\in\mathcal{I}}$ i.i.d. Poisson of mean s_N.

$U := (U_j^I)_{j\in[N],I\in\mathcal{I}}$ i.i.d. uniform in $[N]$.

$\lambda := (\lambda_I)_{I\in\mathcal{I}}$ i.i.d. uniform in $[1/2;1]$.

s_N s.t. $s_N \to +\infty$ and $\frac{s_N}{N} \to 0$.

Poisson perturbation

Add a second perturbation defined by

$$\mathcal{H}_N^{\text{poiss}}(\sigma|\pi, U, \lambda) := -\sum_{I\in\mathcal{I}} \lambda_I \sum_{j=1}^{\pi_I} P_I(\sigma_{U_j^I}).$$

(!) $\mathcal{H}_N^{\text{poiss}}$ is a.s. concave w.r.t. σ.

(!) Poisson perturbation forces full m.o. concentration.

Hypothesis

We assume the following hypothesis:

- $[H1]$: a.s. $\mathcal{H}_N(\sigma|J) \in \mathcal{C}^2(\Sigma_N)$ and concave w.r.t. σ,
- $[H2]$: $\mathcal{H}_N(\sigma_1, \ldots, \sigma_N|J) \overset{d}{=} \mathcal{H}_N(\sigma_{P(1)}, \ldots, \sigma_{P(N)}|J)$.

$\mathbb{E}(\cdot) :=$ expectation w.r.t. J.

$\mathbb{E}_\lambda(\cdot) :=$ expectation w.r.t. λ.

Main results

Proposition (Gibbs m.o. concentration)

Assume $[H1]$. *For all* k,

$$\mathbb{E}_\lambda \mathbb{E} \left\langle \left(R^{(k)} - \left\langle R^{(k)} \right\rangle \right)^2 \right\rangle \leq \frac{\sum_{i=1}^n k_i^2}{N \epsilon_N}$$

Theorem (m.o. concentration)

Assume $[H1] - [H2]$. *For all* k,

$$\lim_{N \to +\infty} \mathbb{E}_\lambda \mathbb{E} \left\langle \left(R^{(k)} - \mathbb{E} \left\langle R^{(k)} \right\rangle \right)^2 \right\rangle = 0\,.$$

Important consequences

Corollary (Asymptotic spin distribution)

Under $[H1]-[H2]$*, for every* $(N_j)_{j\geq 1}$ *s.t.* $\forall i, l \geq 1$*,* σ_i^l *converge in dist., there exists a probability measure* $\nu \in \mathcal{B}([-1;1])$ *and (for* $i \geq 1$*)* $\mu_i \sim \nu$ *i.i.d., so that* $(\sigma_i^l)_{l\geq 1}$ *converge jointly in dist. to samples from* $\mu_i(\cdot)$*.*

Corollary (Strong asymptotic spin independence)

Under $[H1]-[H2]$*, for all distinct* $i_1, \ldots, i_k$*, and* $h_1, \ldots, h_k$ *continuous functions,*

$$\mathbb{E}(\langle h_1(\sigma_{i_1}) \ \cdots \ h_k(\sigma_{i_k})\rangle - \langle h_1(\sigma_{i_1})\rangle \ \cdots \ \langle h_k(\sigma_{i_k})\rangle)^2 \xrightarrow{N\to+\infty} 0$$

In [10], a softer decorrelation is derived from overlap concentration.

---

[10] *Talagrand. (2010). Mean field models for spin glasses: Vol. I.*

Discussion of the results

These results are a step forward in two directions:

(!) Extending adaptive interpolation methods to the non Bayes-optimal regime of inference and ML setting/ERM.

(!) Studying the relationship between the approaches based on interpolation and Gordon's min-max theorem.

Strategy of the proofs

Technical background: log-concave densities

Definition

Log-concave density $f(\cdot)$ if $f = e^{\phi}$, for some $\phi : \mathbb{R}^N \to \mathbb{R}$ concave.

Brascamp-Lieb's variance inequality implies the following corollary.

Corollary

If Hessian of ϕ upper bounded by $-\epsilon\mathbb{I}$ ($\epsilon > 0$), then for $f \in \mathcal{C}^1$

$$\operatorname{Var} f(X) \leq \frac{1}{\epsilon}\mathbb{E}\,||\nabla f(X)||^2\,.$$

Technical background: Aldous-Hoover representation

Theorem (Aldous-Hoover)

$(X_{ij})_{i,j\geq 1}$ *invariant by permutations iff* $X_{ij} \stackrel{d}{=} f(u, v_i, w_j, x_{ij})$; *where* $f : [0;1]^4 \to \mathbb{R}$ *and* u, v_i, w_j, x_{ij} *i.i.d. Unif*$[0;1]$.

By tightness and this, there is $\sigma : [0;1]^4 \to [0;1] \to$ s.t.,

$$\sigma_i^l \xrightarrow{d} \sigma(u, v_i, w_l, x_{il}),$$

along some subsequence and with (u, v_i, w_l, x_{il}) as above.

(!) These variables parametrise correlations.

Technical background: Aldous-Hoover representation

Meaning of uniform variables in A-H representation:

- $u \to$ only depends on disorder, correlates every spin of every replica.
- $v_i \to$ correlates same spin in different replicas.
- $w_l \to$ correlates all the spins of the same replica.
- $x_{il} \to$ randomness particular of every single spin.

Obs: Gibbs mean $\langle \cdot \rangle$ goes to $\int_0^1 \cdots \int_0^1 (\cdot) dw\, dx$.

Technical background: limits of m.o.'s

Lemma

In this limit, we have that for every $n \geq 1$ *and* $k \in K_n$,

$$R^{(k)} \xrightarrow{d} R_{\infty}^{(k)}(u, w_1, \ldots, w_n) := \int_0^1 \prod_{l \leq n} \bar{\sigma}^{(k_l)}(u, v, w_l)\, dv,$$

with $\bar{\sigma}^{(k_l)}(u, v, w_l) := \int_0^1 \sigma^{k_l}(u, v, w_l, x) dx$.

By Gibbs concentration of m.o.'s (Main Results) we get:

Corollary

Let $k \geq 1$. *We have that a.s.* $\bar{\sigma}^{(k)}(u, v, w) = \bar{\sigma}^{(k)}(u, v)$.

Strategy of the proofs

Lemma (Energy concentration)

Assume $[H1]-[H2]$. *Let* $E_I(\sigma) := \sum_{j=0}^{\pi_I} P_I(\sigma_{U_j^I})$. *For* $I \in \mathcal{I}$,

$$\mathbb{E}_\lambda \mathbb{E} \left\langle \left| E_I(\sigma) - \mathbb{E}\left\langle E_I(\sigma) \right\rangle \right| \right\rangle \leq (5 v_N^{1/4} + \sqrt{2}) s_N^{1/2},$$

We get Franz-de Sanctis [11] type ineq (kind of spin glass' Ghirlanda-Guerra ids.). Define $\theta_{I,j}^l := P_I(\sigma_i^l)$.

Theorem

Given $[H1]-[H2]$, *for all* $n \geq 1$, $I \in \mathcal{I}$, *and* $f_n : \Sigma_N^n \to [-1;1]$

$$\mathbb{E}_\lambda \left| \mathbb{E} \frac{\left\langle f_n \theta_{I,1} e^{-\lambda_I \sum_{l=1}^n \theta_{I,1}^l} \right\rangle}{\left\langle e^{-\lambda_I \theta_{I,1}} \right\rangle^n} - \mathbb{E}\left\langle f_n \right\rangle \mathbb{E} \frac{\left\langle \theta_{I,1} e^{-\lambda_I \theta_{I,1}} \right\rangle}{\left\langle e^{-\lambda_I \theta_{I,1}} \right\rangle} \right| = o(1).$$

---

[11] *De Sanctis & Franz (2009). Spin glasses: statics and dynamics.*

Strategy of the proofs

From this we derive a decoupling lemma.

Lemma

Assume $[H1]-[H2]$. *For all* $I \in \mathcal{I}$

$$\mathbb{E}_\lambda \left| \mathbb{E} \frac{\left\langle \theta_{I,1} e^{-\lambda_I \theta_{I,1}} \theta_{I,2} e^{-\lambda_I \theta_{I,2}} \right\rangle}{\left\langle e^{-\lambda_I \theta_{I,1}} e^{-\lambda_I \theta_{I,2}} \right\rangle} - \left[\mathbb{E} \frac{\left\langle \theta_{I,1} e^{-\lambda_I \theta_{I,1}} \right\rangle}{\left\langle e^{-\lambda_I \theta_{I,1}} \right\rangle} \right]^2 \right| = o(1).$$

Asume that some m.o. does not concentrate. Define

$$Y_I(u) := \int_0^1 \frac{\int_0^1 \bar{\theta}_I e^{-\lambda_I \bar{\theta}_I} \, dx}{\int_0^1 e^{-\lambda_I \bar{\theta}_I} \, dx} \, dv.$$

Given a subsequence s.t. the spins of every replica converge in distribution, by A-H representation, this lemma implies that a.s. $\operatorname{Var} Y_I = 0$.

Strategy of the proofs

(!) Then, for all $I \in \mathcal{I}$, Y_I is a.s. constant.

(!) True also for limits $a_p \to 0$ and derivatives $\partial/\partial a_p$.

Observation: if $I = (a_k)$, then

$$\frac{1}{a_k} Y_I \xrightarrow{a_k \to 0} R_\infty^{(k)}.$$

We give order to K and by similar relations, the limits of all m.o. are a.s. constant in the subseq. limit. Abs!

Questions?