Symmetric Perceptron with Random Labels

Eren C. Kızıldağ (Columbia University)

Joint work with Tanay Wakhare (MIT)

SampTA 2023

July 10, 2023

Symmetric Binary Perceptron ($\mathcal{SBP}$)

Introduced by Aubin, Perkins, and Zdeborová [APZ19].

- Fix $\kappa, \alpha > 0$. Generate iid $X_i \stackrel{d}{=} \mathcal{N}(0, I_n), 1 \le i \le M$, where $M = \lfloor n\alpha \rfloor$.
- Consider (random) set

$$S_\alpha(\kappa) = \left\{\boldsymbol{\sigma} \in \Sigma_n : |\langle \boldsymbol{\sigma}, X_i \rangle| \le \kappa\sqrt{n}, \forall i\right\}, \quad \text{where} \quad \Sigma_n \triangleq \{-1, 1\}^n.$$

Motivation:

Toy NN, random CSP, average-case discrepancy...

Perceptron Model: Motivation

Toy NN, storing patterns [Wen62, Cov65]. Popular in stat phys [Gar87, GD88, Gar88].

- **Patterns** $X_i \in \mathbb{R}^n$, **activation function** $U : \mathbb{R} \to \{0, 1\}$.
- **Storage wrt** U**:** Find a $\boldsymbol{\sigma} \in \Sigma_n$ s.t. $U(\langle \boldsymbol{\sigma}, X_i \rangle) = 1, \forall i$.
- **Capacity:** Max # of stored patterns M^*. Stat phys prediction for M^*/n as $n \to \infty$.

SBP: $U(x) = \mathbb{1}\{|x| \le \kappa\sqrt{n}\}$. Asymmetric version: $U(x) = \mathbb{1}\{x > \kappa\sqrt{n}\}$.

- SBP is structurally similar to asymmetric version [BDVLZ20].
- Mathematically easier. Analogy with k-**SAT** vs **NAE**-k-**SAT**.

Perceptron Model: Motivation

Random CSP

- Each constraint $X_i \in \mathbb{R}^n$ rules out certain $\boldsymbol{\sigma} \in \Sigma_n$.
- $\alpha = M/n$ is **constraint density**.

Random CSPs: Existence of solns, sol space geometry, limits of efficient algs...

Average-Case Discrepancy Minimization

- Given $\mathcal{M} \in \mathbb{R}^{M \times n}$, compute or bound its **discrepancy** $\min_{\boldsymbol{\sigma} \in \Sigma_n} \|\mathcal{M}\boldsymbol{\sigma}\|_\infty$.
- Vast literature [Spe85, Mat99, BS20]...

SBP: A Sharp Phase Transition

Recall $X_i \stackrel{d}{=} \mathcal{N}(0, I_n), 1 \le i \le M = \lfloor n\alpha \rfloor$ iid and $S_\alpha(\kappa) = \{\boldsymbol{\sigma} \in \Sigma_n : |\langle \boldsymbol{\sigma}, X_i \rangle| \le \kappa\sqrt{n}, \forall i\}$.

Sharp Phase Transition [Perkins-Xu'21, Abbe-Li-Sly'21]

$\lim_{n\to\infty} \mathbb{P}[S_\alpha(\kappa) \neq \varnothing] = \begin{cases} 1, & \text{if } \alpha < \alpha_c(\kappa) \\ 0, & \text{if } \alpha > \alpha_c(\kappa) \end{cases}$, where $\alpha_c(\kappa) = -1/\log_2 \mathbb{P}[|\mathcal{N}(0,1)| \le \kappa]$.

$\alpha_c(\kappa)$ matches **first moment prediction**: $\mathbb{E}[|S_\alpha(\kappa)|] = o(1)$ iff $\alpha > \alpha_c(\kappa)$.

For $\alpha < \alpha_c(\kappa)$:

- [APZ19]: $\liminf_n p_\alpha(\kappa) > 0$ by 2nd **Moment Method**.
- [PX21, ALS21]: $\lim_n p_\alpha(\kappa) = 1 - o(1)$. More delicate tools.

SBP: A Statistical-to-Computational Gap

Gap between existential & best known algorithmic guarantees.

Random CSPs, optimization over random graphs, spin glasses...

Statistical-to-Computational Gap in SBP

- Let $\kappa \to 0$ (after $n \to \infty$). Then $\alpha_c(\kappa) \sim 1/\log(1/\kappa)$.
- $S_\alpha(\kappa) \neq \varnothing$ if $\alpha < 1/\log(1/\kappa)$. Poly-time algs work only when $\alpha = O(\kappa^2)$ [BS20].

Origins of this gap?

- Intricate geometry of sol space.
- Overlap Gap Property [GKPX22].

SBP: Solution Space Geometry and Limits of Algorithms

Theorem (Gamarnik, K., Perkins, and Xu, FOCS 2022 & COLT 2023)

- *SBP exhibits* **Ensemble multi-Overlap Gap Property** *(as $\kappa \to 0$) whp if* $\alpha = \Omega(\kappa^2 \log \frac{1}{\kappa})$.
- *For* $\alpha = \Omega\left(\kappa^2 \log \frac{1}{\kappa}\right)$, *there is* **no stable alg** *for SBP that succeeds w.p.* $O(1)$.
- *For* $\alpha = \Omega(\kappa^2)$, *there is* **no online alg** *for SBP that succeeds w.p.* $\geq \exp(-\Theta(n))$.
- *Kim-Roche algorithm [KR98] is* **stable**.

- **Stable algs** also include low-degree polynomials, and AMP.
- **Online algs** include Bansal-Spencer [BS20], our **benchmark**.

Symmetric Perceptron with Random Labels

Fix $\kappa, \alpha > 0$. Generate iid $X_i \stackrel{d}{=} \mathcal{N}(0, I_n), 1 \le i \le M$, where $M = \lfloor n\alpha \rfloor$.
Activation $U(x) = \mathbb{1}\{|x| \le \kappa\sqrt{n}\}$. Parameter $p \in [0,1]$.

- **Model I:** Let $Y_i \sim \mathrm{Bern}(p)$, $1 \le i \le M$ be iid. Set

$$S_\alpha(\kappa, p) = \{\boldsymbol{\sigma} \in \Sigma_n : Y_i = U(\langle \boldsymbol{\sigma}, X_i \rangle), \forall i\}$$

- **Model II:** Choose a $\mathcal{I} \subset \{1, \dots, M\}$ with $|\mathcal{I}| = Mp$ uar, let $Y_i = \mathbb{1}\{i \in \mathcal{I}\}$. Set

$$\widetilde{S}_\alpha(\kappa, p) = \{\boldsymbol{\sigma} \in \Sigma_n : Y_i = U(\langle \boldsymbol{\sigma}, X_i \rangle), \forall i\}$$

Case $p = 1$ corresponds to SBP.
Also captures $U(x) = \mathbb{1}\{|x| > \kappa\sqrt{n}\}$ by considering **dual** labels $1 - Y_i$.

Comparing Models I and II

- If $Y_i \sim \mathrm{Bern}(p)$ are iid, then $\big|\{i : Y_i = 1\}\big| = Mp + O(\sqrt{M})$ due to **concentration**.
- Y_i are **not independent** under Model II: for $p < 1$,

$$\mathbb{P}[j \in \mathcal{I} \mid i \in \mathcal{I}] = \binom{M-1}{Mp-1} \Big/ \binom{M}{Mp} = \frac{Mp-1}{M-1} < p.$$

Models are not exactly the same. Capacity threshold.

Machine Learning View

- Data $(X_i, Y_i) \in \mathbb{R}^n \times \{0,1\}$, $1 \le i \le M$, find the **best fit** $f(\cdot, \boldsymbol{\sigma}), \boldsymbol{\sigma} \in \theta$.
- Solve the empirical risk minimization:

$$\min_{\boldsymbol{\sigma}\in\theta} \widehat{\mathcal{L}}(\boldsymbol{\sigma}), \quad \text{where} \quad \widehat{\mathcal{L}}(\boldsymbol{\sigma}) = \frac{1}{M}\sum_{1\le i\le M} \ell\big(Y_i; f(X_i, \boldsymbol{\sigma})\big).$$

- Let $\theta = \Sigma_n$, $\ell(y;x) = \mathbb{1}\{y \ne x\}$ and $f(X_i, \boldsymbol{\sigma}) = U(\langle \boldsymbol{\sigma}, X_i\rangle)$.
- Satisfying sol to CSP are **interpolators** of ER:

$$S_\alpha(\kappa, p) = \big\{\boldsymbol{\sigma} \in \Sigma_n : \widehat{\mathcal{L}}(\boldsymbol{\sigma}) = 0\big\}.$$

Negative Spherical Perceptron

$Y_i\langle \boldsymbol{\sigma}, X_i\rangle \ge \kappa$, $\|\boldsymbol{\sigma}\|_2 = 1$. Rigorously studied by Montanari et al. [MZZ21].

Main Results: A Sharp Phase Transition for Expected Cardinality

Let $q(\kappa) = \mathbb{P}\left[|\mathcal{N}(0,1)| \le \kappa\right]$.

Theorem (K. and Wakhare, 2023)

Model I: *Let* $\alpha_c(\kappa, p) = -1/\log_2\left(pq(\kappa) + (1-p)(1-q(\kappa))\right)$. *Then,*

$$\mathbb{E}\left[\left|S_\alpha(\kappa,p)\right|\right] = \begin{cases} \exp(-\Theta(n)), & \text{if } \alpha > \alpha_c(\kappa,p) \\ \exp(\Theta(n)), & \text{if } \alpha < \alpha_c(\kappa,p) \end{cases},$$

Model II: *Let* $\widetilde{\alpha}_c(\kappa, p) = -1/\left(p\log_2 q(\kappa) + (1-p)\log_2(1-q(\kappa)\right)$. *Then,*

$$\mathbb{E}\left[\left|\widetilde{S}_\alpha(\kappa,p)\right|\right] = \begin{cases} \exp(-\Theta(n)), & \text{if } \alpha > \widetilde{\alpha}_c(\kappa,p) \\ \exp(\Theta(n)), & \text{if } \alpha < \widetilde{\alpha}_c(\kappa,p) \end{cases}.$$

In particular, $S_\alpha(\kappa,p) = \varnothing$ *whp for* $\alpha > \alpha_c(\kappa,p)$ *and* $\widetilde{S}_\alpha(\kappa,p) = \varnothing$ *whp for* $\alpha > \widetilde{\alpha}_c(\kappa,p)$.

Sharp Phase Transition for Expected Cardinality

Proof Sketch

- Based on **first moment method**. Fix $\boldsymbol{\sigma} \in \Sigma_n$. Then,

$$\mathbb{P}\big[\boldsymbol{\sigma} \in S_\alpha(\kappa, p)\big] = \mathbb{P}\big[Y_i = U(\langle \boldsymbol{\sigma}, X_i \rangle), \forall i\big] = \big(pq(\kappa) + (1-p)(1-q(\kappa)\big)^{\alpha n}.$$

- By linearity of expectatation,

$$\mathbb{E}\big[|S_\alpha(\kappa, p)|\big] = 2^n \cdot \big(pq(\kappa) + (1-p)(1-q(\kappa)\big)^{\alpha n} = \exp_2\left(n\left(1 - \frac{\alpha}{\alpha_c(\kappa, p)}\right)\right)$$

- $n^{-1} \log \mathbb{E}\big[|S_\alpha(\kappa, p)|\big]$ is **annealed free energy** in stat phys.
- So, $\alpha_c(\kappa, p)$, $\widetilde{\alpha}_c(\kappa, p)$ is **annealed capacity**.

$n^{-1}\mathbb{E}\big[\log |S_\alpha(\kappa, p)|\big]$ is **quenched free energy**. **Harder** to study.

Model with Independent Labels Have Higher Annealed Capacity

- **Model I**: IID labels, $\alpha_c(\kappa, p) = -1/\log_2\bigl(pq(\kappa) + (1-p)(1-q(\kappa))\bigr)$.
- **Model II**: Dependent labels, $\widetilde{\alpha}_c(\kappa, p) = -1/\bigl(p\log_2 q(\kappa) + (1-p)\log_2(1-q(\kappa))\bigr)$.
- As $x \mapsto \log_2 x$ is **concave**, **Jensen's inequality** yields $\alpha_c(\kappa, p) \geq \widetilde{\alpha}_c(\kappa, p)$.

Model I with iid labels has **higher** annealed capacity.

Capacity vs dependence structure for other random CSPs?

Main Results: Universality for Annealed Capacity

Annealed capacity do **not** depend on distributional details.

Theorem (K. and Wakhare, 2023)

$\alpha_c(\kappa, p)$ and $\widetilde{\alpha}_c(\kappa, p)$ remains the same if $X_i = (X_i(j) : 1 \le j \le n)$ has iid coordinates with

$$\mathbb{E}[X_i(1)] = 0, \quad \mathbb{E}[X_i(1)^2] > 0, \quad \textit{and} \quad \mathbb{E}[|X_i(1)^3|] < \infty.$$

- Proof based on **Berry-Esseen Theorem**.
- **Related result**: [GKPX22] establish universality for Ensemble-m-OGP in $\mathtt{SBP}$.

A Sharp Phase Transition Conjecture

Large $\mathbb{E}\left[\left|S_\alpha(\kappa, p)\right|\right]$ does not mean $S_\alpha(\kappa, p) \neq \varnothing$ whp.

1st moment **prediction** for SBP is **correct** [PX21, ALS21].

Conjecture

$\exists \kappa^* > 0$ *such that for every* $\kappa < \kappa^*$ *and* $p \in [0, 1]$,

$$\lim_{n\to\infty} \mathbb{P}[S_\alpha(\kappa, p) \neq \varnothing] = \begin{cases} 0, & \text{if } \alpha > \alpha_c(\kappa, p) \\ 1, & \text{if } \alpha < \alpha_c(\kappa, p), \end{cases}$$

$$\lim_{n\to\infty} \mathbb{P}[\widetilde{S}_\alpha(\kappa, p) \neq \varnothing] = \begin{cases} 0, & \text{if } \alpha > \widetilde{\alpha}_c(\kappa, p) \\ 1, & \text{if } \alpha < \widetilde{\alpha}_c(\kappa, p). \end{cases}$$

For $p = 0$, moment method works only for $\kappa < \kappa^* \approx 0.817$ [APZ19]. **RSB** for $\kappa > \kappa^*$.

Main Results: An Evidence Towards Sharp PT Conjecture

Theorem (K. and Wakhare, 2023)

$\forall \kappa > 0$, $\exists p_\kappa^* < 1$ *such that the following holds. Fix any* $p \in [p_\kappa^*, 1]$, $\alpha < \widetilde{\alpha}_c(\kappa, p)$. *Then,*

$$\liminf_{n\to\infty} \mathbb{P}\left[\widetilde{S}_\alpha(\kappa, p) \ne \varnothing\right] > 0.$$

$\forall \kappa \in (0, 0.817)$, $\exists p_\kappa^{**} > 0$ *such that the following holds. Fix any* $p \in [0, p_\kappa^{**}]$, $\alpha < \widetilde{\alpha}_c(\kappa, p)$. *Then,*

$$\liminf_{n\to\infty} \mathbb{P}\left[\widetilde{S}_\alpha(\kappa, p) \ne \varnothing\right] > 0.$$

- Covers p close to 1 (SBP) and close to 0 (u-function binary perceptron).
- Based on 2nd **moment method** [AM02, APZ19].
- Contingent on an assumptionn regarding a real function [DS19, APZ19, PX21].

Proof Idea

Based on **second moment method**.

Let $Z = \left|\widetilde{S}_\alpha(\kappa, p)\right|$. **Goal:** $\liminf_{n\to\infty} \mathbb{P}[Z \geq 1] > 0$.

Paley-Zygmund Inequality

$$\mathbb{P}[Z \geq 1] = \mathbb{P}[Z > 0] \geq \frac{\mathbb{E}[Z]^2}{\mathbb{E}[Z^2]}.$$

To prove: $\mathbb{E}[Z^2] = \Theta(\mathbb{E}[Z]^2)$. Laplace's method [AM02].

Future Directions

- **Sharp PT** analogous to SBP [PX21, ALS21].
- Interplay between **capacity** and **dependence structure**.
- Other perceptron models, *e.g.* **spherical** case or different activations.
- **Polynomial-time** algs for finding a $\boldsymbol{\sigma} \in S_\alpha(\kappa, p)$.
- Limits of algs. Solution space geometry and OGP.

References I

- Emmanuel Abbe, Shuangping Li, and Allan Sly, *Proof of the contiguity conjecture and lognormal limit for the symmetric perceptron*, arXiv preprint arXiv:2102.13069 (2021).
- Dimitris Achlioptas and Cristopher Moore, *The asymptotic order of the random k-sat threshold*, The 43rd Annual IEEE Symposium on Foundations of Computer Science, 2002. Proceedings., IEEE, 2002, pp. 779–788.
- Benjamin Aubin, Will Perkins, and Lenka Zdeborová, *Storage capacity in symmetric binary perceptrons*, Journal of Physics A: Mathematical and Theoretical **52** (2019), no. 29, 294003.
- Carlo Baldassi, Riccardo Della Vecchia, Carlo Lucibello, and Riccardo Zecchina, *Clustering of solutions in the symmetric binary perceptron*, Journal of Statistical Mechanics: Theory and Experiment **2020** (2020), no. 7, 073303.
- Nikhil Bansal and Joel H. Spencer, *On-line balancing of random inputs*, Random Structures and Algorithms **57** (2020), no. 4, 879–891 (English (US)).

References II

Thomas M Cover, *Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition*, IEEE transactions on electronic computers (1965), no. 3, 326–334.

Jian Ding and Nike Sun, *Capacity lower bound for the Ising perceptron*, Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing, 2019, pp. 816–827.

Elizabeth Gardner, *Maximum storage capacity in neural networks*, EPL (Europhysics Letters) **4** (1987), no. 4, 481.

________, *The space of interactions in neural network models*, Journal of physics A: Mathematical and general **21** (1988), no. 1, 257.

Elizabeth Gardner and Bernard Derrida, *Optimal storage properties of neural network models*, Journal of Physics A: Mathematical and general **21** (1988), no. 1, 271.

References III

- David Gamarnik, Eren C Kızıldağ, Will Perkins, and Changji Xu, *Algorithms and barriers in the symmetric binary perceptron model*, 2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS), IEEE, 2022, pp. 576–587.
- Jeong Han Kim and James R Roche, *Covering cubes by random half cubes, with applications to binary neural networks*, Journal of Computer and System Sciences **56** (1998), no. 2, 223–252.
- Jiri Matousek, *Geometric discrepancy: An illustrated guide*, vol. 18, Springer Science & Business Media, 1999.
- Andrea Montanari, Yiqiao Zhong, and Kangjie Zhou, *Tractability from overparametrization: The example of the negative perceptron*, arXiv preprint arXiv:2110.15824 (2021).

References IV

- Will Perkins and Changji Xu, *Frozen 1-RSB structure of the symmetric Ising perceptron*, Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing, 2021, pp. 1579–1588.
- Joel Spencer, *Six standard deviations suffice*, Transactions of the American mathematical society **289** (1985), no. 2, 679–706.
- James G Wendel, *A problem in geometric probability*, Mathematica Scandinavica **11** (1962), no. 1, 109–111.