

A NOTE ON THE CAPACITY OF THE BINARY PERCEPTRON

DYLAN J. ALTSCHULER AND KONSTANTIN TIKHOMIROV

ABSTRACT. Determining the capacity α_c of the Binary Perceptron is a long-standing problem. Krauth and Mézard (1989) conjectured an explicit value of α_c , approximately equal to .833, and a rigorous lower bound matching this prediction was recently established by Ding and Sun (2019). Regarding the upper bound, Kim and Roche (1998) and Talagrand (1999) independently showed that $\alpha_c < .996$, while Krauth and Mézard outlined an argument which can be used to show that $\alpha_c < .847$. The purpose of this expository note is to record a complete proof of the bound $\alpha_c < .847$. The proof is a conditional first moment method combined with known results on the spherical perceptron.

1. INTRODUCTION

Let $A \in \mathbb{R}^{\alpha N \times N}$ be a matrix with independent and identically distributed standard Gaussian entries. Consider the problem of finding $\sigma \in \{-1, +1\}^N$ with $A\sigma > 0$, entrywise. The largest value of α such that this problem is solvable is the random quantity $\alpha_* := \alpha_*(A)$, called the *capacity*. The long-standing binary perceptron problem asks for the value α_c that the capacity concentrates around.

For intuition, observe $\alpha_c \leq 1$ is a trivial bound. Indeed, define the nested random sets Z_t ,

$$Z_t := \left\{ \sigma \in \{-1, +1\}^N : A_i \sigma > 0, \quad \forall i \in 1, \dots, t \right\}.$$

For any fixed $\sigma \in Q$, we have $(A\sigma)_i > 0$ with probability one half, independently for each $i \in [t]$. So $\mathbb{E}[|Z_t|] = 2^{N-t}$, and thus for any ϵ by Markov's inequality:

$$\mathbb{P}[\alpha_* > 1 + \epsilon] = \mathbb{P}[|Z_{(1+\epsilon)N}| \geq 1] < 2^{-\epsilon N}.$$

A long-standing and notoriously difficult open problem is to rigorously verify the following conjecture.

Conjecture 1 (Krauth and Mezard [7], 1989). *The capacity of the binary perceptron concentrates around an explicit constant $\alpha_c \approx .833$.*

The value of α_c is given as the fixed point of the so-called TAP, or Thouless-Anderson-Palmer, equations. This characterization is not yet known to be well-posed due to uniqueness issues. The TAP equations can be derived from non-rigorous but powerful heuristics such as the replica method or cavity method. Current progress can be summarized as follows:

Theorem 1 (Ding and Sun [2]; Kim and Roche [6]; Talagrand [10]; Xu [13]). *With high probability, the capacity of the binary perceptron is at least α_c (i.e. .833...) and at most .9963.*

The lower bound was proven in a breakthrough of Ding and Sun. Their result is conditional on some numerical assumptions—this is not atypical in constraint satisfaction literature—and was originally proven to hold only with positive probability. The mismatch between positive probability and high probability was overcome by Xu [13] who showed the existence of a sharp threshold sequence (see also [8] for a later, different proof, and [1] for an alternative perspective). The upper-bound .9963 was shown by Kim and Roche, and a different proof (establishing an upper-bound of $1 - \epsilon$ for some small, non-explicit $\epsilon > 0$) was subsequently discovered by Talagrand. In [7, section 5], Krauth and Mezard outlined an argument which can be used to prove the following theorem:

Theorem 2. *With high probability, the capacity of the binary perceptron is at most .847.*

DYLAN J. ALTSCHULER, DEPARTMENT OF MATHEMATICAL SCIENCES, CARNEGIE MELLON UNIVERSITY.

KONSTANTIN TIKHOMIROV, DEPARTMENT OF MATHEMATICAL SCIENCES, CARNEGIE MELLON UNIVERSITY.

The proof of the theorem is based on a deep result of Shcherbina and Tirozzi [9] about the “spherical perceptron,” a related model that is far better understood. The paper [7] appeared long before the spherical perceptron was solved in [9], and for that reason the argument in [7, section 5] could not be made fully rigorous at the time. The purpose of our note is to provide a complete proof of Theorem 2 as a corollary of the main result of [9].

1.1. Spherical Perceptron. The closely related *spherical* perceptron problem asks for the largest α so that there exists $\sigma \in S^{N-1}$ with $A\sigma > 0$ entrywise. The spherical perceptron predates the binary perceptron and was originally studied as a model of the fundamental limits of machine learning. It is also significantly better understood than the binary perceptron: since σ is drawn from a convex set in the spherical case, one has access to a variety of powerful methods from convex geometry.

Recall for the binary perceptron, one can track the size of $|Z_t|$, the number of remaining solutions after adding t constraints. For the spherical perceptron it makes sense to track the volume of the solution set instead. Since volume scales exponentially in dimension, it is natural to actually consider the logarithm of the volume. Since this can possibly be negative infinity, it is further sensible to truncate the volume at some super-exponential cutoff. This motivates the study of the following random quantity, sometimes called the free energy:

$$F(A) := \frac{1}{N} \max \left\{ \ln \left(\frac{\text{Vol} \left(B(0, \sqrt{N}; \mathbb{R}^N) \cap \{\sigma : A\sigma > 0\} \right)}{\text{Vol} \left(B(0, \sqrt{N}; \mathbb{R}^N) \right)} \right), -N^2 \right\}. \quad (1)$$

The celebrated prediction of Gardner [3, 4] (later in collaboration with Derrida [5]) from 1987, rigorously established by Shcherbina and Tirozzi in 2002 [9], gives an exact formula for the free energy F .

Theorem 3 (Gardner-Derrida formula for the Spherical Perceptron [9]). *Let $F := F(A)$, where A is an $\alpha N \times N$ matrix with iid standard Gaussian entries.*

- $\text{Var}(F) \rightarrow 0$ as $N \rightarrow \infty$.
- For any $\alpha > 2$, $F \rightarrow -\infty$ as $N \rightarrow \infty$.
- For any $\alpha < 2$, letting $u \in \mathbb{R}$ be distributed as a standard normal, we have:

$$\lim_{N \rightarrow \infty} \mathbb{E}[F] = \text{GD}(\alpha) := \min_{q \in [0, 1]} \alpha \mathbb{E} \left[\ln H \left(\frac{u\sqrt{q}}{\sqrt{1-q}} \right) \right] + \frac{1}{2} \frac{q}{1-q} + \frac{1}{2} \ln(1-q), \quad (2)$$

where $H(\sigma) = \mathbb{P}[Z \geq \sigma]$ for Z a standard normal random variable.

We refer the reader to the monographs of Talagrand [11, 12] for detailed exposition.

1.2. Proof of Main Result. Let A be an $\alpha N \times N$ matrix, for $\alpha = .847$. We wish to show that with high probability, there does not exist $\sigma \in \{-1, +1\}^N$ with $A\sigma > 0$ entrywise. As before, we proceed by Markov’s inequality. We start with the following decoupling trick. Let O be a random rotation uniform from the Haar measure. By basic symmetry considerations,

$$\mathbb{E}[|Z_{\alpha N}|] = \sum_{\sigma \in \{-1, +1\}^N} \mathbb{P}[A\sigma > 0] = \sum_{\sigma \in \{-1, +1\}^N} \mathbb{P}[A(O\sigma) > 0]. \quad (3)$$

Observe that $O\sigma$ is uniformly distributed on the sphere of radius \sqrt{N} . Recalling Eq. (1)

$$\mathbb{P}[A(O\sigma) > 0 \mid A] = \frac{\text{Vol} \left(B(0, \sqrt{N}; \mathbb{R}^N) \cap \{\sigma : A\sigma > 0\} \right)}{\text{Vol} \left(B(0, \sqrt{N}; \mathbb{R}^N) \right)} \leq e^{NF(A)}.$$

By Theorem 3, the event $\mathcal{G} := \{A : |F(A) - \text{GD}(\alpha)| < \epsilon\}$ has probability tending to one for any $\epsilon > 0$. We claim that for ϵ sufficiently small (e.g. $\epsilon = .001$), it holds $\mathcal{G} \subset \{A : F(A) < -\log 2 - .001\}$.

Proposition 1. $\text{GD}(.847) < -\log 2 - .002$.

Assuming the proposition, the theorem readily follows. By Markov’s inequality:

$$\mathbb{P}[|Z_{\alpha N}| \geq 1] \leq o(1) + \mathbb{P}[|Z_{\alpha N}| \geq 1, \mathcal{G}] \leq o(1) + \mathbb{E}[|Z_{\alpha N}| \mathbb{1}(\mathcal{G})].$$

But, the event \mathcal{G} implies $F(A) < -\log(2) - .001$. Then Eq. (3) yields:

$$\mathbb{E}[|Z_{\alpha N}| \mathbb{1}(\mathcal{G})] \leq \sum_{\sigma \in \{-1, +1\}^N} e^{-N \log 2 - .001N}.$$

In total, $\mathbb{P}[|Z_{\alpha N}| \geq 1] = o(1)$ where $\alpha = .847$. Equivalently stated, the capacity is at most α with high probability. All that remains is the proof of the proposition.

Proof of Proposition. Since $\text{GD}(\alpha)$ is given by a minimization problem and we seek to establish an upper-bound, it suffices to check any particular value of q . Define

$$\text{GD}(\alpha, q) := \alpha \mathbb{E} \left[\ln H \left(\frac{u\sqrt{q}}{\sqrt{1-q}} \right) \right] + \frac{1}{2} \frac{q}{1-q} + \frac{1}{2} \ln(1-q).$$

Take $q = 1/2$. This choice is informed by numerically computing the true minimizer to be $q \approx .504$, and also by the fact that taking $q = 1/2$ trivializes the computation of GD . Indeed, letting ϕ and Φ denote the Gaussian PDF and CDF respectively, integration by parts yields:

$$\mathbb{E}[\ln H(u)] = \int_{\mathbb{R}} \phi(u) \log H(u) du = - \int_{\mathbb{R}} (1 - \Phi(u)) \frac{\phi(u)}{1 - \Phi(u)} du = -1.$$

Thus,

$$\begin{aligned} \text{GD}(.847, .5) &:= .847 \mathbb{E}[\ln H(u)] + \frac{1}{2} \left(1 + \ln \left(\frac{1}{2} \right) \right) \\ &= -.847 + (.15342 \dots) \\ &< -\ln(2) - .002. \end{aligned}$$

□

REFERENCES

- [1] ALTSCHULER, D. J. Zero-one laws for random feasibility problems. *arXiv preprint arXiv:2309.13133* (2023).
- [2] DING, J., AND SUN, N. Capacity lower bound for the Ising perceptron. In *STOC'19—Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing* (2019), ACM, New York, pp. 816–827.
- [3] GARDNER, E. Maximum storage capacity in neural networks. *Europhysics letters* 4, 4 (1987), 481.
- [4] GARDNER, E. The space of interactions in neural network models. *J. Phys. A* 21, 1 (1988), 257–270.
- [5] GARDNER, E., AND DERRIDA, B. Optimal storage properties of neural network models. *J. Phys. A* 21, 1 (1988), 271–284.
- [6] KIM, J. H., AND ROCHE, J. R. On the optimal capacity of binary neural networks: Rigorous combinatorial approaches. In *Proceedings of the eighth annual Conference on Learning Theory* (1995), pp. 240–249.
- [7] KRAUTH, W., AND MÉZARD, M. Storage capacity of memory networks with binary couplings. *Journal de Physique* 50, 20 (1989), 3057–3066.
- [8] NAKAJIMA, S., AND SUN, N. Sharp threshold sequence and universality for Ising perceptron models. In *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)* (2023), SIAM, Philadelphia, PA, pp. 638–674.
- [9] SHCHERBINA, M., AND TIROZZI, B. On the volume of the intersection of a sphere with random half spaces. *C. R. Math. Acad. Sci. Paris* 334, 9 (2002), 803–806.
- [10] TALAGRAND, M. Intersecting random half cubes. vol. 15. 1999, pp. 436–449. Statistical physics methods in discrete probability, combinatorics, and theoretical computer science (Princeton, NJ, 1997).
- [11] TALAGRAND, M. *Mean field models for spin glasses. Volume I*. Springer-Verlag, Berlin, 2011.
- [12] TALAGRAND, M. *Mean field models for spin glasses. Volume II*. Springer, Heidelberg, 2011.
- [13] XU, C. Sharp threshold for the Ising perceptron model. *Ann. Probab.* 49, 5 (2021), 2399–2415.

APPENDIX

The following is MATLAB code to numerically optimize the Gardner-Derrida formula. It is not explicitly used in the main body of this article. Nonetheless, we offer it to the interested reader since it was helpful for exploration.

```
% Grid search over possible values of the order parameter "q" and the
    constraint density "alpha".
qlist = .001:.001:.999;
a = .846:.00005:.847;

% Each possible value of "q" certifies an upper-bound on the Gardner-
    Derrida free energy formula. Track all "candidate" upper bounds.
GDcand = zeros(length(alpha),length(qlist));

gpdf = @(u) exp(-u.^2./2) ./ sqrt(2*pi) ; % Gaussian PDF
H = @(x) (1 - erf(x ./sqrt(2))) ./ 2; % Complement of Gaussian CDF

% evaluate GD(alpha, q).
for i = 1:length(qlist)
    q = qlist(i);

    % in order to avoid log(0), add a perturbation "eps" (of size
        2^{-52}) to the argument for numerical stability.
    I = @(u) gpdf(u) .* log(eps('double') + H(u .* sqrt(q) ./ sqrt
        (1-q)) );

    % array of candidate solutions (changed logarithm to base 2)
    GDcand(:,i) = (alpha .* integral(I,-10,10) + q./(2.*(1-q)) + log
        (1-q) ./ 2) ./ log(2);
end
close all
GD = min(GDcand,[],2) ; % returns col vector GD(alpha)
plot(alpha, GD)

xlabel('\alpha')
ylabel('GD')
yline(-1, 'r', 'GD = -1')
xline(.84655, 'k', '\alpha = .84655')
```