A theory of learning with constrained weight-distribution

Weishun Zhong
Harvard and MIT
wszhong@mit.edu

Ben Sorscher
Stanford University
bsorsch@stanford.edu

Daniel D Lee
Cornell Tech
ddl46@cornell.edu

Haim Sompolinsky
Harvard and Hebrew University
haim@fiz.huji.ac.il

Abstract

A central question in computational neuroscience is how structure determines function in neural networks. Recent large-scale connectomic studies have started to provide a wealth of structural information such as the distribution of excitatory/inhibitory cell and synapse types as well as the distribution of synaptic weights in the brains of different species. The emerging high-quality large structural datasets raise the question of what general functional principles can be gleaned from them. Motivated by this question, we developed a statistical mechanical theory of learning in neural networks that incorporates structural information as constraints. We derived an analytical solution for the memory capacity of the perceptron, a basic feedforward model of supervised learning, with constraint on the distribution of its weights. Interestingly, the theory predicts that the reduction in capacity due to the constrained weight-distribution is related to the Wasserstein distance between the cumulative distribution function of the constrained weights and that of the standard normal distribution. To test the theoretical predictions, we use optimal transport theory and information geometry to develop an SGD-based algorithm to find weights that simultaneously learn the input-output task and satisfy the distribution constraint. We show that training in our algorithm can be interpreted as geodesic flows in the Wasserstein space of probability distributions. We further developed a statistical mechanical theory for teacher-student perceptron rule learning and ask for the best way for the student to incorporate prior knowledge of the rule (i.e., the teacher). Our theory shows that it is beneficial for the learner to adopt different prior weight distributions during learning, and shows that distribution-constrained learning outperforms unconstrained and sign-constrained learning. Our theory and algorithm provide novel strategies for incorporating prior knowledge about weights into learning, and reveal a powerful connection between structure and function in neural networks.

1 Introduction

Learning and memory are thought to take place at the microscopic level by modifications of synaptic connections. Unlike learning in artificial neural networks, synaptic plasticity in the brain operates under structural biological constraints. Theoretical efforts to incorporate some of these constraints have focused largely on the degree of connectivity [17, 35] and the constraints on the sign of the synapses (Excitatory vs. Inhibitory) [4, 16], but few include additional features of synaptic weight distributions observed in the brain [11]. More generally, recent large-scale connectomic studies

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

[36, 59, 61] are beginning to provide a wealth of structural information of neuronal circuits at an unprecedented scope and level of precision, which presents a remarkable opportunity for a more refined theoretical study of learning and memory that takes into account these hitherto unavailable structural information.

Perceptron [56] is arguably the simplest model of computation by single neuron and is the fundamental building block for many modern neural networks. Despite the drastic oversimplification, studying the computational properties of (binary and analog) perceptron has been used extensively in computational neuroscience since its dawn, particularly in the cerebellum (as a model of sensory-motor association) but also in cerebral cortex (for generic associative memory functions) [43, 2, 16, 18, 15, 14]. Forming associations is considered an 'atomic' building block for generic cortical functions, and perceptron memory capacity sets a tight bound on the memory capacity in recurrently connected neuronal circuits with application to cortex and hippocampus [27, 55, 57]. Statistical mechanical analysis predicts that near capacity, an unconstrained perceptron classifying random input-output associations has normally distributed weights [29, 28, 21], see Fig.1(a). In contrast, physiological experiments suggest that biological synapses do not change their excitatory/inhibitory identity during learning (but see recent [33]). In order to take perceptron a step closer to biological realism, prior work has imposed sign constraints during learning [4, 16]. In this case, the predicted weight distribution is a delta-function centered at zero plus a half-normal distribution, see Fig.1(b). However, a wide range of connectomic studies ranging from cortical circuits in animals [38, 32, 47, 73, 61, 40, 9], to human cerebral cortex [47, 61] have shown evidence of lognormally distributed synaptic connections. As an example, Fig.1(c) shows the weight connection distribution in mouse primary auditory cortex (data adapted from [38]). Possible reasons for the ubiquitous lognormal distributions range from biological structural/developmental constraints to computational benefits [63]. Various potential mechanisms for lognormal distributions has been proposed, from multiplicative gradient updates in feedforward networks[34, 40], to mixture of additive and multiplicative plasticity rules in spiking networks[30], but the majority of these proposals lead not just to lognormal distributions but also to sparsification in the weights. Instead of adding yet another explanation to the computational origin of lognormal distribution, here we take the observed weight distribution as a prior on the network structure, and ask for its computational consequences. The goal of the paper is to present for the first time a quantitative and qualitative theory of neural network learning performance under non-Gaussian and general weight distributions (not limited to lognormal distributions).

In this paper, we combine two powerful tools: statistical mechanics and optimal transport theory, and present a theory of perceptron learning that incorporates the knowledge of both distribution and sign information as constraints, and gives accurate predictions for capacity and generalization error. Interestingly, the theory predicts that the reduction in capacity due to the constrained weight-distribution is related to the Wasserstein distance between the cumulative distribution function of the constrained weights and that of the standard normal distribution. Along with the theoretical framework, we also present a learning algorithm derived from information geometry that is capable of efficiently finding viable perceptron weights that satisfy desired distribution and sign constraints. This paper is organized as follows: in Section 2.1 we derive the perceptron capacity for classifying random input-output associations using statistical mechanics, and illustrate our theory with a simple example. In Section 3, we derive our learning algorithm using optimal transport theory, and show that distribution of weights found by the learning algorithm coincide with geodesic distributions on a Wasserstein statistical manifold, and therefore training can be interpreted as a geodesic flow. In Section 4 we analyze a parameterized family of biologically realistic weight distributions, and use our theory to predict the shape of the distribution with optimal parameters. We map out the experimental parameter landscape for the estimated distribution of synaptic weights in mammalian cortex and show that our theory's prediction for optimal distribution is close to the experimentally measured value. In Section 5 we further develop a statistical mechanical theory for teacher-student perceptron rule learning and ask for the best way for the student to incorporate prior knowledge about the weight distribution of the rule (i.e., the teacher). Our theory shows that it is beneficial for the learner to adopt different prior weight distributions during learning.

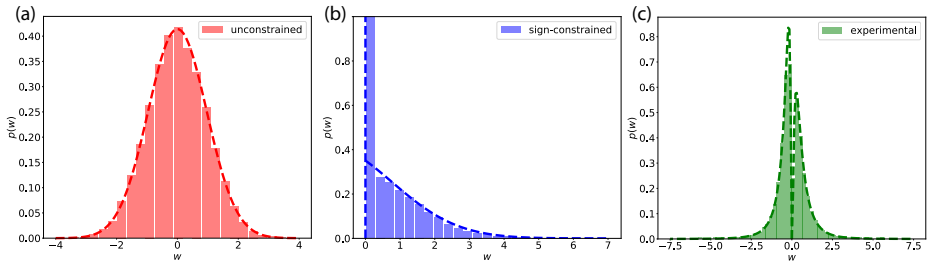

Figure 1: Theoretical and empirical synaptic weight distributions. (a)-(b) predicted distribution following perceptron learning at capacity. (a) Normal distribution when learning is unconstrained. (b) A delta-function plus a half-normal distribution when learning is sign-constrained. (c) Experimentally measured synaptic weight distribution (mouse primary auditory cortex [38]).

2 Capacity

2.1 Learning under weight distribution constraints

We begin by considering a canonical learning problem: classifying random input-output associations by a perceptron. In biological memory systems, the heavily correlated sensory data is undergoing heavy preprocessing including massive decorrelations, and previous work on brain related perceptron modeling [27, 16, 57] assumes similarly unstructured data. The data consists of pairs$\{\boldsymbol{\xi}^\mu, \zeta^\mu\}_{\mu=1}^P$, where $\boldsymbol{\xi}^\mu$ is an N-dimentional random vector drawn i.i.d. from a standard normal distribution, $p(\xi_i^\mu) = \mathcal{N}(0,1)$, and ζ^μ are random binary class labels with $p(\zeta^\mu) = \frac{1}{2}\delta(\zeta^\mu + 1) + \frac{1}{2}\delta(\zeta^\mu - 1)$. The goal is to find a hyperplane through the origin, described by a perceptron weight vector $\boldsymbol{w} \in \mathbb{R}^N$, normalized to $||\boldsymbol{w}||^2 = N$.

We call $\boldsymbol{w}$ a separating hyperplane when it correctly classifies all the examples with margin $\kappa > 0$:

$$\zeta^\mu \frac{\boldsymbol{w} \cdot \boldsymbol{\xi}^\mu}{||\boldsymbol{w}||} \geq \kappa. \tag{1}$$

We are interested in solutions $\boldsymbol{w}$ to Eqn.1 that obey a prescribed distribution constraint, $w_i \sim q(w)$, where q is an arbitrary probability density function. We further demand that $\langle w^2 \rangle_{q(w)} = 1$ to fix the overall scale of the distribution (since the task is invariant to the overall scale of w). Thus, the goal of learning is to find weights that satisfy 1 with the additional constraint that the empirical density function $\hat{q}(w) = \frac{1}{N}\sum_i^N \delta(w - w_i)$, formed by the learned weights is similar to $q(w)$, and more precisely that it converges to $q(w)$ as $N \to \infty$ (see Section 2.2 below).

Extension of this setup that includes an arbitrary number of populations each satisfying its own prescribed distribution constraints is discussed in Section 4 and in Appendix A.1.2. Note that the sign constraint is a special case of this scenario with two synaptic populations: one excitatory and one inhibitory. We further discuss the generalization of this setup to include biased inputs and sparse labels in Appendix A.1.3.

2.2 Statistical mechanical theory of capacity

We are interested in the thermodynamic limit where $P, N \to \infty$, but the load $\alpha = \frac{P}{N}$ stays $\mathcal{O}(1)$. This limit is amenable to mean-field analysis using statistical mechanics.

Following Gardner's seminal work [29, 28], we consider the fraction V of viable weights that satisfies both Eqn.1 and the distribution constraint $\hat{q} = q$, to all possible weights:

$$V = \frac{\int d\boldsymbol{w} \left[\prod_{\mu=1}^P \Theta\left(\zeta^\mu \frac{\boldsymbol{w}\cdot\boldsymbol{\xi}^\mu}{||\boldsymbol{w}||} - \kappa\right)\right] \delta(||\boldsymbol{w}||^2 - N)\delta\left(\int dk \left(\hat{q}(k) - q(k)\right)\right)}{\int d\boldsymbol{w}\delta(||\boldsymbol{w}||^2 - N)}. \tag{2}$$

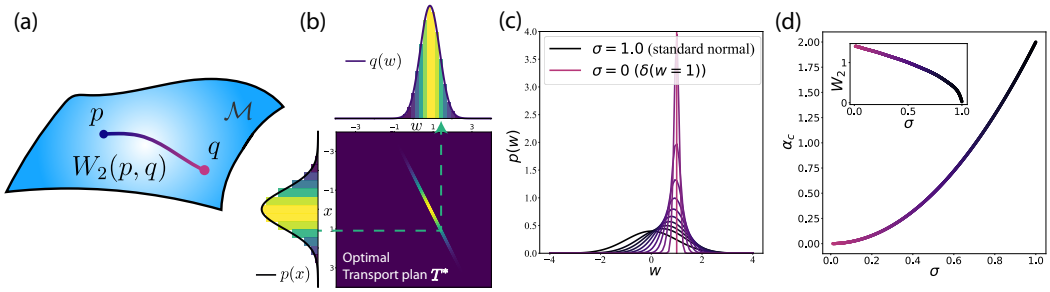

Figure 2: An illustration of optimal transport from a standard normal distribution $\mathcal{N}(0,1)$ to normal distributions with nonzero mean $\mathcal{N}(\sqrt{1-\sigma^2}, \sigma^2)$. (a) A schematic of the space $(\mathcal{M}, W_2)$ of probability distributions. (b) An example optimal transport plan from standard normal, $p(x)$, to a normal with $\sigma = 0.5$, $q(w)$. The optimal transport plan T^* is plotted in between the distributions. T^* moves $p(x)$ units of probability mass x to location w, as indicated by the dashed line, and the colors are chosen to reflect the amount of probability mass to be transported. (c) $\mathcal{N}(\sqrt{1-\sigma^2}, \sigma^2)$ interpolates between standard normal ($\sigma = 0$) to a δ-function at 1 ($\sigma = 1$). (d) Capacity $\alpha_c(\kappa = 0)$ as a function of σ. Inset shows the W_2 distance as a function of σ.

In Eqn.2, we impose the distribution constraint $\hat{q} = q$ by demanding that in the thermodynamic limit, all Fourier modes of q and $\hat{q}$ are the same , i.e., that $q(k) = \int dw e^{ikw} q(w) = \hat{q}(k) = \frac{1}{N}\sum_i^N e^{ikw_i}$,where in the last equality we have used the definition of empirical distribution. We perform a quenched average over random patterns $\boldsymbol{\xi}^\mu$ and labels ζ^μ. This amounts to calculating $\langle \log V \rangle$, which can be done using the replica trick [29, 28].

We focus on solutions with maximum margin κ at a given load α, or equivalently, the maximum load capacity $\alpha_c(\kappa)$ of separable patterns given margin κ. We proceed by assuming replica symmetry in our mean field analysis, which in general might not hold because the constraint $\hat{q} = q$ is non-convex. For all the results presented in the main text, replica symmetry solution is supported by numerical simulations. In Appendix A.5 we explore the validity of replica symmetric solutions in the case of strongly bimodal distributions and show that they fail only very close to the binary (Ising) limit.

Detailed calculations of the mean-field theory are presented in Appendix A.1.1. Our mean-field theory predicts that the reduction in capacity due to the distribution constraint is proportional to the Jacobian of the transformation from $w \sim q(w)$ to a normally distributed variable $x(w) \sim \mathcal{N}(0,1)$,

$$\alpha_c(\kappa) = \alpha_0(\kappa) \left\langle \frac{dw}{dx} \right\rangle_x^2, \tag{3}$$

where $\alpha_0(\kappa) = \left[\int_{-\kappa}^{\infty} Dt(\kappa + t)^2\right]^{-1}$ is the capacity of an unconstrained perceptron, from Gardner theory [29, 28], and $\kappa = 0$ reduces to the classical result of $\alpha_0(0) = 2$. The Jacobian factor, $\langle dw/dx \rangle_x$, can be written in terms of the constrained distribution's cumulative distribution function (CDF), $Q(w)$, and the standard normal CDF $P(x) = \frac{1}{2}\left[1 + \text{Erf}(\frac{x}{\sqrt{2}})\right]$, namely,

$$\left\langle \frac{dw}{dx} \right\rangle_x = \int_0^1 du Q^{-1}(u) P^{-1}(u). \tag{4}$$

Note that since the second moments are fixed to unity, $0 \leq \left\langle \frac{dw}{dx} \right\rangle_x \leq 1$ and it equals 1 iff $p = q$.

2.3 Geometrical interpretation of capacity

The jacobian factor Eqn.4 can be rewritten as

$$\left\langle \frac{dw}{dx} \right\rangle_x = 1 - \frac{1}{2} W_2(Q,P)^2, \tag{5}$$

where W_k ($k = 2$ in above) is the Wasserstein-k distance, given by

$$W_k(Q, P) = \left[\int_0^1 du \left(Q^{-1}(u) - P^{-1}(u) \right)^k \right]^{1/k}. \tag{6}$$

[In the following, we will make frequent use of both the probability density function (PDF), and the cumulative distribution function (CDF). We distinguish them by using upper case letters for CDFs, and lower case letters for PDFs.]

The Wasserstein distance measures the dissimilarity between two probability distributions, and is the geodesic distance between points on the manifold of probability distributions [41, 25, 20]. Therefore, we can interpret Eqn.3 as predicting that the reduction in memory capacity tracks the geodesic distance we need travel from the standard normal distribution P to the target distribution Q (Fig.2(a)).

We demonstrate Eqn.3 and Eqn.5 with an instructive example. Let's consider a parameterized family of normal distributions, with the second moment fixed to 1: $q(w) = \mathcal{N}(\sqrt{1-\sigma^2}, \sigma^2)$, see Fig.2(c). At $\sigma = 1$, $q(w)$ is the standard normal distribution and we recover the unconstrained Gardner capacity $\alpha_0(\kappa = 0) = 2$. As $\sigma \to 0$, $q(w)$ becomes a δ-function at 1 and $\alpha_c(\kappa) \to 0$ (Fig.2(c)).

As evident in this simple example, perceptron capacity is strongly affected by its weight distribution. Our theory enables prediction of the shape of the distribution with optimal parameters within a parameterized family of distributions. We apply our theory to a family of biologically plausible distributions and compare our prediction with experimentally measured distributions in Section 4.

3 Optimal transport and the DisCo-SGD learning algorithm

Eqn.3 predicts the storage capacity for a perceptron with a given weight distribution, but it does not specify a learning algorithm for finding a solution to this non-convex learning problem. Here we present a learning algorithm for perceptron learning with a given weight distribution constraint. This algorithm will also serve to test our theoretical predictions. For this purpose, we use optimal transport theory to develop an SGD-based algorithm that is able to find max-margin solutions that obey the prescribed distribution constraint. Furthermore, we show that training can be interpreted as traveling along the geodesic connecting the current empirical distribution and the target distribution.

Stochastic gradient descent (SGD) on a cross-entropy loss has been shown to asymptotically converge to max-margin solutions on separable data [62, 50]. Given data $\{\boldsymbol{\xi}^\mu, \zeta^\mu\}_{\mu=1}^P$, we use logistic regression to predict class labels from our perceptron weights, $\hat{\zeta}^\mu = \sigma(\boldsymbol{w}^t \cdot \boldsymbol{\xi}^\mu)$,where $\sigma(z) = (1 + e^{-z})^{-1}$ and $\boldsymbol{w}^t$ is the weight at the t-th update. This defines an SGD update rule :

$$w_i^{t+\delta t} \leftarrow w_i^t - \delta t \sum_\mu \xi_i^\mu (\hat{\zeta}^\mu - \zeta^\mu), \tag{7}$$

where the μ-summation goes from 1 to P for full-batch GD and goes from 1 to mini-batch size B for mini-batches SGD (see Appendix A.4 for more details). The theory of optimal transport provides a principled way of transporting each individual weight w_i^t to a new value so that overall the new set of weights satisfies the prescribed target distribution. In 1-D, the optimal transport plan T^* has a closed-form solution in terms of the current CDF P and target CDF Q [65, 3]: $T^* = Q^{-1} \circ P$, where $\circ$ denotes functional composition. We demonstrate the optimal transport map in Fig.2(b) for the instructive example discussed in Section 2.3.

In order to apply T^* to transport our weights $\{w_i\}$ (omitting superscript t), we form the empirical CDF $\hat{Q}(w) = \frac{1}{N} \sum_{i=1}^N \mathbf{1}_{w_i \leq w}$, which counts how many weights w_i are observed below value w. Then the new set of weights $\{\hat{w}_i\}$ satisfying target CDF Q can be written as

$$\hat{w}_i = Q^{-1} \circ \hat{Q}(w_i). \tag{8}$$

We illustrate Eqn.8 in action in Table 1(b).

(a)

Algorithm 1: Distribution-constrained SGD (DisCo-SGD)

Initialize patterns $\boldsymbol{\xi}^\mu$, labels ζ^μ randomly, and initialize weights w_i at zeros;
while $\zeta^\mu \neq sgn(\boldsymbol{w}^T\boldsymbol{\xi}^\mu)$ *for some* μ **do**
  compute logistic regression labels: $\hat{\zeta}^\mu = \sigma(\beta \boldsymbol{w}^T\boldsymbol{\xi}^\mu)$;
  step along the gradient: $w_i \leftarrow w_i - \eta_1 \sum_\mu \xi_i^\mu(\hat{\zeta}^\mu - \zeta^\mu)$;
  $\beta \leftarrow \beta - \sum_\mu \boldsymbol{w}^T\boldsymbol{\xi}^\mu(\hat{\zeta}^\mu - \zeta^\mu)$
  compute empirical CDF $\hat{Q}(w) = \frac{1}{N}\sum_{i=1}^N \mathbf{1}_{w_i \leq w}$;
  step along the optimal transport map: $w_i \leftarrow w_i - \eta_2 \left[w_i - Q^{-1} \circ \hat{Q}(w_i)\right]$;
  if *exceeds max iters* **then**
    break;
  end
end

(b)

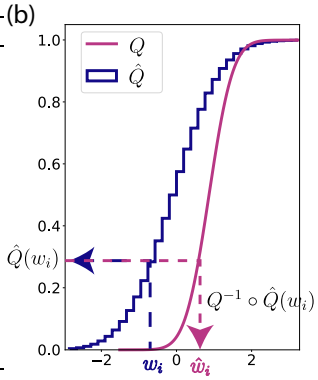

Table 1: Disco-SGD algorithm. (a) We perform alternating steps of gradient descent along the cross-entropy loss (Eqn.7), followed by steps along the optimal transport direction (Eqn.9). (b) An illustration of Eqn.8. For a given w_i, we first compute its empirical CDF value $\hat{Q}(w_i)$,then use the inverse target CDF to transport w_i to its new value, $\hat{w}_i = Q^{-1}\left(\hat{Q}(w_i)\right)$.

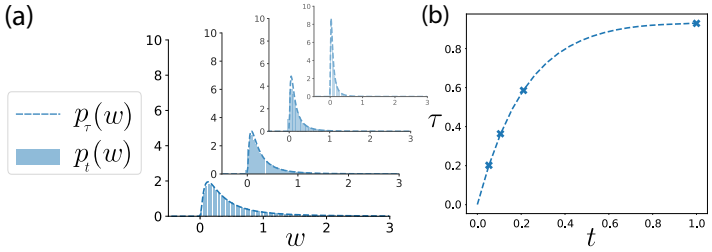

Figure 3: Intermediate distributions during learning are on the geodesic. (a) The solid histograms are the intermediate distribution p_t at different training time t from the DisCo-SGD algorithm, the dashed lines are geodesic distributions p_τ with the same W_2 distance to the target distribution Q. From right to left the training time advances, and the distributions transform further away from the δ-function initialization, and approach the target distribution (a lognormal, in this example). (b) The geodesic time τ as a function of the training time t. Location of the crosses correspond to the distributions shown in (a).

However, performing such a one-step projection strongly interferes with the cross-entropy objective, and numerically often results in solutions that do not perfectly classify the data. Therefore, it would be beneficial to have an incremental update rule based on Eqn.8:

$$w_i^{\tau+\delta\tau} \leftarrow w_i^\tau + \delta\tau\left(\hat{w}_i - w_i^\tau\right), \tag{9}$$

where we have used a different update time τ to differentiate with the cross-entropy update time t.

We present our complete algorithm in Table 1(a), which we named 'Distribution-constrained SGD' (DisCo-SGD) algorithm. In the DisCo-SGD algorithm, we perform alternating updates on Eqn.7 and Eqn.9, and identify δt and $\delta\tau$ as learning rates η_1 and η_2. Note that in logistic regression, the norm of the weight vector $||\boldsymbol{w}||$ is known to increase with training and the max-margin solution is only recovered at $||\boldsymbol{w}|| \to \infty$. In contrast, imposing a distribution constraint fixes the norm. Therefore, to allow a variable norm, in Table 1 we include a trainable parameter β in our algorithm to serve as the norm of the weight vector. This algorithm allows us to reliably discover linearly separable solutions obeying the prescribed weight distribution Q.

Interestingly, Eqn.9 takes a similar form to geodesic flows in Wasserstein space. Given samples $\{w_i\}$ drawn from the initial distribution P and $\{\hat{w}_i\}$ drawn from the final distribution Q, samples $\{w_i^\tau\}$ from intermediate distributions P_τ along the geodesic can be calculated as $w_{(i)}^\tau = (1-\tau)w_{(i)} + \tau\hat{w}_{(i)}$, where subscript (i) denotes ascending order (see more in Appendix A.2). For intermediate perceptron

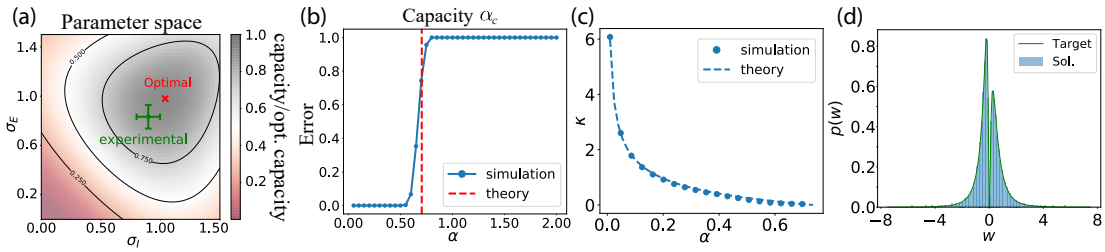

Figure 4: Biologically-realistic distribution and parameter landscape. (a) Capacity (normalized by the optimal value in the landscape) as a function of the lognormal parameters σ_E and σ_I. Experimental value is shown in green with error bars, and optimal capacity is shown in red. (b)-(d) (theory from Eqn.10 and simulations from DisCo-SGD): (b) Determination of capacity; (c) Max-margin κ at different load α, which is the same as $\alpha_c(\kappa)$; (d) Example weight distribution obtained in simulation.

weights $\boldsymbol{w}^t$ found by our algorithm, we can compute its empirical distribution p_t and compare with theoretical distribution p_τ along the geodesic with the same W_2 distance to the target distribution (see Appendix A.2 for how to calculate p_τ). In Fig.3(a), we show that indeed the empirical distributions p_t agree with the geodesic distributions p_τ at geodesic time $\tau(t)$ (Fig.3(a)). The relation between the geodesic time τ and the SGD update time t is shown in Fig.3(b). The interplay between the cross-entropy objective and the distribution constraint is manifested in the rate at which the distribution moves along the geodesic between the initial distribution and the target one.

4 Biologically-realistic distribution (E/I balanced lognormals) and experimental landscape

In order to apply our theory to the more biologically-realistic cases, we generalize our theory from a single prescribed distribution to an arbitrary number of input subpopulations each obeys its own distribution. We consider a perceptron that consists of M synaptic populations $\boldsymbol{w}^m$ indexed by m, each constrained to satisfy its own weight distribution $w_i^m \sim q_m(w^m)$. We denote the overall weight vector as $\boldsymbol{w} \equiv \{\boldsymbol{w}^m\}_{m=1}^M \in \mathbb{R}^{N\times 1}$, where the total number of weights is $N = \sum_{m=1}^M N_m$. In this case, the capacity Eqn.3 is generalized to (See Appendix A.1.2 for detailed derivation):

$$\alpha_c(\kappa) = \alpha_0(\kappa)\left[\sum_m^M g_m \left\langle \frac{dw^m}{dx} \right\rangle_x\right]^2, \tag{10}$$

where $g_m = N_m/N$ is the fraction of weights in this population. Eqn. 10 allows us to investigate the parameter space of capacity with biologically-realistic distributions and compare with the experimentally measured values. In particular, we are interested the case with two synaptic populations that models the excitatory/inhibitory synpatic weights of a biological neuron, hence, $m = E, I$. We model the excitatory/inhibitory synaptic weights as drawn from two separate lognormal distributions ($g_I = 1 - g_E$): $w_i^E \sim \frac{1}{\sqrt{2\pi}\sigma_E w^E}\exp\left\{-\frac{(\ln w^E - \mu_E)^2}{2\sigma_E^2}\right\}$ and $w_i^I \sim \frac{1}{\sqrt{2\pi}\sigma_I w^I}\exp\left\{-\frac{(\ln w^I - \mu_I)^2}{2\sigma_I^2}\right\}$.

We also demand that the mean synaptic weights satisfy the E/I balance condition [69, 70, 67, 68, 57, 48, 18] $g_E\left\langle w^E\right\rangle = g_I\left\langle w^I\right\rangle$ as is often observed in cortex connectomic experiments [5, 72, 51, 53, 8]. With the E/I balance condition and fixed second moment, the capacity is a function of the lognormal parameters σ_E and σ_I. In Fig.4(a) we map out the 2d parameter space of σ_E and σ_I using Eqn.10, and find that the optimal choice of parameters which yields the maximum capacity solution is close to the experimentally measured values in a recent connectomic studies in mouse primary auditory cortex [38].

In order to test our theory's validity on this estimated distribution of synaptic weights, we perform DisCo-SGD simulation with model parameters σ_E and σ_I fixed to their experimentally measured values. Both the capacity (Fig.4(b)), max-margin κ at different load (Fig.4(c)), and the empirical weights found by the algorithm (Fig.4(d)) are in good agreement with our theoretical prediction.

5 Generalization performance

5.1 Distribution-constrained learning as circuit inference

A central question in computational neuroscience is how underlying neural circuits determine its computation. Recently, thanks to new parallelized functional recording technologies, simultaneous recordings of the activity of hundreds of neurons in response to an ensemble of inputs are possible [1, 12]. An interesting challenge is to infer the structural connectivity from the measured input-output activity patterns. It is interesting to ask how are these stimuli-response relations related to the underlying structure of the circuit [54, 39]. In the following, we try to adress this circuit reconstruction task in a simple setup where a student perceptron tries to learn from a teacher perceptron [60, 22]. In this setup, the teacher is considered to be the underlying ground-truth neural circuit. The student is attempting to infer the connection weights of this ground-truth circuit by observing a series of input-output relations generated by the teacher. After learning is completed, one can assess the faithfulness of the inference by comparing the teacher and student. The teacher-student setup is also a well-known 'toy model' for studying generalization performance [42, 37, 44]. In this case since the learning data are generated by the teacher, the overlap between teacher and student determines the generalization performance of the learning. Here we ask to what extent prior knowledge of the teacher weight distribution helps in learning the rule and how this knowledge can be incorporated in learning. A similar motivation may arise in other contexts, in which there is a prior knowledge about the weight distribution of an unknown target linear classifier.

Let's consider the teacher perceptron, $\boldsymbol{w}_t \in \mathbb{R}^N$, drawn from some ground-truth distribution p_t. Given random inputs $\boldsymbol{\xi}^\mu$ with $p(\xi_i^\mu) = \mathcal{N}(0,1)$, we generate labels by $\zeta^\mu = \text{sgn}(\boldsymbol{w}_t \cdot \boldsymbol{\xi}^\mu / ||\boldsymbol{w}_t|| + \eta^\mu)$, where η^μ is input noise and $\eta^\mu \sim \mathcal{N}(0, \sigma^2)$. We task the student perceptron $\boldsymbol{w}_s$ to find the max-margin linear classifier for data $\{\boldsymbol{\xi}^\mu, \zeta^\mu\}_{\mu=1}^p$: $\max \kappa : \zeta^\mu \boldsymbol{w}_s \cdot \boldsymbol{\xi}^\mu \geq \kappa ||\boldsymbol{w}_s||$. Let's define the teacher-student overlap as

$$R = \frac{\boldsymbol{w}_s \cdot \boldsymbol{w}_t}{\|\boldsymbol{w}_s\| \, \|\boldsymbol{w}_t\|}, \tag{11}$$

which is a measure the faithfulness of the circuit inference. The student's generalization error is then related to the overlap by $\varepsilon_g = 1/\pi \arccos\left(R/\sqrt{1+\sigma^2}\right)$ [60, 22].

As a baseline, let's first consider a totally uninformed student (without any structural knowledge of the teacher), learning from a teacher with a given (in particular non-Gaussian) weight distribution. In this case, we can determine the overlap R (Eqn.11) as a function of load α by solving the replica symmetric mean field self-consistency equations as in [60, 22]. An example of such learning for a lognormal teacher distribution is shown in Fig.5(a) ('unconstrained') for the noiseless case ($\sigma = 0$). Note that in the presence of noise in the labels ($\sigma \neq 0$), α is bounded by $\alpha_c(\sigma)$, since max-margin learning of separable data is assumed. The case with nonzero σ is presented in Appendix A.3.4. In this unconstrained case, the student's weight distribution evolves from a Gaussian for low α to one which increasingly resembles the teacher distribution for large α (Fig.5(b)).

Next, we consider a student with information about the signs of the individual teacher weights. We can apply this knowledge as a constraint and demand that the signs of individual student weights agree with that of the teacher's. The additional sign-constraints require a modification of replica calculation in [60, 22], which we present in Appendix A.3.1. Surprisingly, we find both analytically and numerically that if the teacher weights are not too sparse, the max-margin solution generalizes poorly: after a single step of learning (with random input vectors), the overlap, R, drops substantially from its initial value (see 'sign-constrained' in Fig.5(a)). The source of the problem is that, due to the sign constraint, max-margin training with few examples yields a significant mismatch between the student and teacher weight distributions. After only a few steps of learning, half of the student's weights are set to zero, and the student's distribution, $p(w_s) = 1/2\delta(0) + 1/\sqrt{2\pi}\exp\{-w_s^2/4\}$, deviates significantly from the teacher's distribution (see more in Appendix A.3.3). The discrepancy between the teacher and student weight distributions therefore suggest that we should incorporate distribution-constraint into learning.

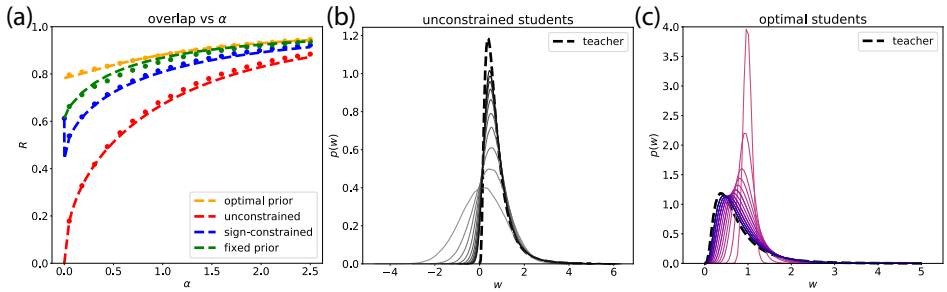

Figure 5: Compare different learning paradigms. (a) Teacher-student overlap R , or equivalently the generalization error $\varepsilon_g = 1/\pi \arccos R$, as a function of load α in different learning paradigms. Dashed lines are from theory, and dots are from simulation. Note that there is an initial drop of the overlap in sign-constrained learning due to sparsification of weights. (b)-(c) The darker color curves correspond to larger α, and dashed line is teacher distribution (same in both cases). (b) Distribution of an unconstrained student evolves from normal distribution toward the teacher distribution. (c) Optimal student prior evolves from a δ-function toward the teacher distribution.

5.2 Distribution-constrained learning outperforms unconstrained and sign-constrained learning

Let's consider the case that the student weight are constrained to some *prior* distribution $q_s(w_s)$, while the teacher obeys a distribution $p_t(w_t)$,for an arbitrary pair q_s, p_t. We can write down the Gardner volume V_g for generalization as in the capacity case (Eqn.2):

$$V_g = \frac{\int d\boldsymbol{w}_s \left[\prod_{\mu=1}^{P} \Theta\left(\text{sgn}\left(\frac{\boldsymbol{w}_t\cdot\boldsymbol{\xi}^\mu}{||\boldsymbol{w}_t||} + \eta^\mu\right)\frac{\boldsymbol{w}_s\cdot\boldsymbol{\xi}^\mu}{||\boldsymbol{w}_s||} - \kappa\right)\right]\delta(||\boldsymbol{w}_s||^2 - N)\delta\left(\int dk\,(\hat{q}(k) - q(k))\right)}{\int d\boldsymbol{w}_s \delta(||\boldsymbol{w}_s||^2 - N)}. \tag{12}$$

To obtain ensemble average of system over different realizations of the training set, we perform the quenched average of $\log V_g$ over the patterns $\boldsymbol{\xi}^\mu$ and teacher $\boldsymbol{w}_t$, and consider the thermodynamic limit of $N, P \to \infty$ and $\alpha = \frac{P}{N}$ stays $\mathcal{O}(1)$. We use the replica trick similar to [60, 22]. Overlap R (Eqn.11) can be determined as a function of load α by solving the replica symmetric mean field self-consistency equations in Appendix A.3.2. In this distribution-constrained setting, we can perform numerical simulations with DisCo-SGD algorithm (Table 1) to find such weights and compare with the predictions of our theory.

Now we ask if the student has a *prior* on the teacher's weight distribution p_t, whether incorporating this knowledge in training will improve generalization performance. One might be tempted to conclude that the optimal prior distribution the student should adopt is always that of the teacher's, i.e., $q_s = p_t$. We call this learning paradigm 'fixed prior', and show that its generalization performance is better than that of the unconstrained and sign-constrained case (Fig.5(a)). However, instead of using a fixed prior for the student, we can in fact choose the *optimal prior* distribution p_s^* at different load α. This presents a new learning paradigm we called 'optimal prior'. In Fig.5(a), we show that choosing optimal priors at different α achieves the overall best generalization performance compared with all other learning paradigms. For a given parameterized family of distributions, our theory provides a way to analytically obtain the optimal prior p_s^* as a function of α (Fig.5(c)). Note that unlike the unconstrained case (Fig.5(b)), the optimal prior starts from a δ-function at 1 at zero α, and asymptotically approaches the teacher distribution p_t as $\alpha \to \infty$.

6 Summary and Discussion

We have developed a statistical mechanical framework that incorporates structural constraints (sign and weight distribution) into perceptron learning. The synaptic weights in our perceptron learning satisfy two key biological constraints: (1) individual synaptic signs are not affected by the learning task (2) overall synaptic weights obey a prescribed distribution. These constraints may arise also in

neuromorphic devices [31, 66]. Under the replica-symmetry assumption, we derived a novel form of distribution-constrained perceptron storage capacity, which admits a simple geometric interpretation of the reduction in capacity in terms of the Wasserstein distance between the standard normal distribution and the imposed distribution. To numerically test our analytic theory, we used tools from optimal transport and information geometry to develop an SGD-based algorithm, DisCo-SGD, in order to reliably find weights that satisfy such prescribed constraints and correctly classify the data, and showed that training with the algorithm can be interpreted as geodesic flows in the Wasserstein space of distributions. It would be interesting to compare our theory and algorithm to [7, 58] where the Wasserstein distance is used as an objective for training generative models. We applied our theory to the biologically realistic case of of excitatory/inhibitory lognormal distributions that are observed in the cortex, and found experimentally-measured parameters close to the optimal parameter values predicted by our theory. We further studied input-output rule learning where the target rule is defined in terms of a weighted sum of the inputs, and asked to what extent prior knowledge of the target distribution may improve generalization performance. Using the teacher-student perceptron learning setup, we showed analytically and numerically that distribution constrained learning substantially enhances the generalization performance. In the context of circuit inference, distribution constrained learning provides a novel and reliable way to recover the underlying circuit structure from observed input-output neural activities. In summary, our work provides new strategies of incorporating knowledge about weight distribution in neural learning and reveals a powerful connection between structure and function in neural networks. Ongoing extensions of the present work include weight distribution constraints in recurrent and deep architectures as well as testing against additional connectomic databases.

Acknowledgments

This paper is dedicated to the memory of Mrs. Lily Safra, a great supporter of brain research. The authors would like to thank Madhu Advani, Haozhe Shan, and Julia Steinberg for very helpful discussions. W.Z. acknowledges support from the Swartz Program in Theoretical Neuroscience at Harvard and the NIH grant NINDS (1U19NS104653). B.S. acknowledges support from the Stanford Graduate Fellowship. D.D.L. acknowledges support from the NIH grant NINDS (1U19NS104653). H.S. acknowledges support from the Swartz Program in Theoretical Neuroscience at Harvard, the NIH grant NINDS (1U19NS104653), and the Gatsby Charitable Foundation.

References

[1] Misha B Ahrens, Michael B Orger, Drew N Robson, Jennifer M Li, and Philipp J Keller. Whole-brain functional imaging at cellular resolution using light-sheet microscopy. *Nature methods*, 10(5):413–420, 2013. 5.1

[2] James S Albus. A theory of cerebellar function. *Mathematical biosciences*, 10(1-2):25–61, 1971. 1

[3] Luigi Ambrosio and Nicola Gigli. A user's guide to optimal transport. In *Modelling and optimisation of flows on networks*, pages 1–155. Springer, 2013. 3, A.2, A.2

[4] Daniel J Amit, KY Michael Wong, and Colin Campbell. Perceptron learning with sign-constrained weights. *Journal of Physics A: Mathematical and General*, 22(12):2039, 1989. 1

[5] Jeffrey S Anderson, Matteo Carandini, and David Ferster. Orientation tuning of input conductance, excitation, and inhibition in cat primary visual cortex. *Journal of neurophysiology*, 84(2):909–926, 2000. 4

[6] Chris Angeloni and Maria N Geffen. Contextual modulation of sound processing in the auditory cortex. *Current opinion in neurobiology*, 49:8–15, 2018.

[7] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017. 6, A.2

[8] Bassam V Atallah and Massimo Scanziani. Instantaneous modulation of gamma oscillation frequency by balancing excitation with inhibition. *Neuron*, 62(4):566–577, 2009. 4

[9] Michael Avermann, Christian Tomm, Celine Mateo, Wulfram Gerstner, and Carl CH Petersen. Microcircuits of excitatory and inhibitory neurons in layer 2/3 of mouse barrel cortex. *Journal of neurophysiology*, 107(11):3116–3134, 2012. 1, A.1.2

[10] Carlo Baldassi, Alfredo Braunstein, Nicolas Brunel, and Riccardo Zecchina. Efficient supervised learning in networks with binary synapses. *Proceedings of the National Academy of Sciences*, 104(26):11079–11084, 2007. A.5.1

[11] Boris Barbour, Nicolas Brunel, Vincent Hakim, and Jean-Pierre Nadal. What can we learn from synaptic weight distributions? *TRENDS in Neurosciences*, 30(12):622–629, 2007. 1

[12] Antal Berényi, Zoltán Somogyvári, Anett J Nagy, Lisa Roux, John D Long, Shigeyoshi Fujisawa, Eran Stark, Anthony Leonardo, Timothy D Harris, and György Buzsáki. Large-scale, high-density (up to 512 channels) recording of local circuits in behaving animals. *Journal of neurophysiology*, 111(5):1132–1149, 2014. 5.1

[13] M Bouten, L Reimers, and B Van Rompaey. Learning in the hypercube: A stepping stone to the binary perceptron. *Physical Review E*, 58(2):2378, 1998. A.5.1, A.5.1

[14] Guy Bouvier, Johnatan Aljadeff, Claudia Clopath, Célian Bimbard, Jonas Ranft, Antonin Blot, Jean-Pierre Nadal, Nicolas Brunel, Vincent Hakim, and Boris Barbour. Cerebellar learning using perturbations. *Elife*, 7:e31599, 2018. 1

[15] Nicolas Brunel. Is cortical connectivity optimized for storing information? *Nature neuroscience*, 19(5):749–755, 2016. 1

[16] Nicolas Brunel, Vincent Hakim, Philippe Isope, Jean-Pierre Nadal, and Boris Barbour. Optimal information storage and the distribution of synaptic weights: perceptron versus purkinje cell. *Neuron*, 43(5):745–757, 2004. 1, 2.1

[17] György Buzsáki and Kenji Mizuseki. The log-dynamic brain: how skewed distributions affect network operations. *Nature Reviews Neuroscience*, 15(4):264–278, 2014. 1

[18] Julio Chapeton, Tarec Fares, Darin LaSota, and Armen Stepanyants. Efficient associative memory storage in cortical circuits of inhibitory and excitatory neurons. *Proceedings of the National Academy of Sciences*, 109(51):E3614–E3622, 2012. 1, 4

[19] Jing Chen, Yifan Chen, Hao Wu, and Dinghui Yang. The quadratic wasserstein metric for earthquake location. *Journal of Computational Physics*, 373:188–209, 2018. A.2

[20] Yifan Chen and Wuchen Li. Optimal transport natural gradient for statistical manifolds with continuous sample space. *Information Geometry*, 3(1):1–32, 2020. 2.3, A.2

[21] Thomas M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, EC-14(3):326–334, 1965. 1

[22] Andreas Engel and Christian Van den Broeck. *Statistical mechanics of learning*. Cambridge University Press, 2001. 5.1, 5.1, 5.2, A.3.1

[23] Bjorn Engquist and Brittany D Froese. Application of the wasserstein metric to seismic signals. *arXiv preprint arXiv:1311.4581*, 2013. A.2

[24] Bjorn Engquist, Brittany D Froese, and Yunan Yang. Optimal transport for seismic full waveform inversion. *arXiv preprint arXiv:1602.01540*, 2016. A.2

[25] Alessio Figalli and Cédric Villani. Optimal transport and curvature. In *Nonlinear PDE's and Applications*, pages 171–217. Springer, 2011. 2.3, A.2

[26] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. Learning with a wasserstein loss. *Advances in neural information processing systems*, 28, 2015. A.2

[27] Elizabeth Gardner. Maximum storage capacity in neural networks. *EPL (Europhysics Letters)*, 4(4):481, 1987. 1, 2.1

[28] Elizabeth Gardner. The space of interactions in neural network models. *Journal of physics A: Mathematical and general*, 21(1):257, 1988. 1, 2.2, 2.2, 2.2, A.1.1

[29] Elizabeth Gardner and Bernard Derrida. Optimal storage properties of neural network models. *Journal of Physics A: Mathematical and general*, 21(1):271, 1988. 1, 2.2, 2.2, 2.2, A.1.1

[30] Matthieu Gilson and Tomoki Fukai. Stability versus neuronal specialization for stdp: long-tail weight distributions solve the dilemma. *PloS one*, 6(10):e25339, 2011. 1

[31] Geonhui Han, Chuljun Lee, Jae-Eun Lee, Jongseon Seo, Myungjun Kim, Yubin Song, Young-Ho Seo, and Daeseok Lee. Alternative negative weight for simpler hardware implementation of synapse device based neuromorphic system. *Scientific reports*, 11(1):1–7, 2021. 6

[32] Carl Holmgren, Tibor Harkany, Björn Svennenfors, and Yuri Zilberter. Pyramidal cell communication within local networks in layer 2/3 of rat neocortex. *The Journal of physiology*, 551(1):139–153, 2003. 1, A.1.2

[33] SeulAh Kim, Michael L Wallace, Mahmoud El-Rifai, Alexa R Knudsen, and Bernardo L Sabatini. Co-packaging of opposing neurotransmitters in individual synaptic vesicles in the central nervous system. *Neuron*, 2022. 1

[34] Jyrki Kivinen and Manfred K Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *information and computation*, 132(1):1–63, 1997. 1

[35] Alexei A Koulakov, Tomáš Hromádka, and Anthony M Zador. Correlated connectivity and the distribution of firing rates in the neocortex. *Journal of Neuroscience*, 29(12):3685–3694, 2009. 1

[36] Michael Kunst, Eva Laurell, Nouwar Mokayes, Anna Kramer, Fumi Kubo, António M Fernandes, Dominique Förster, Marco Dal Maschio, and Herwig Baier. A cellular-resolution atlas of the larval zebrafish brain. *Neuron*, 103(1):21–38, 2019. 1

[37] Sebastian Lee, Sebastian Goldt, and Andrew Saxe. Continual learning in the teacher-student setup: Impact of task similarity. In *International Conference on Machine Learning*, pages 6109–6119. PMLR, 2021. 5.1

[38] Robert B Levy and Alex D Reyes. Spatial profile of excitatory and inhibitory synaptic connectivity in mouse primary auditory cortex. *Journal of Neuroscience*, 32(16):5609–5619, 2012. 1, 1, 4, A.1.2, A.4

[39] Jian K Liu, Helene M Schreyer, Arno Onken, Fernando Rozenblit, Mohammad H Khani, Vidhyasankar Krishnamoorthy, Stefano Panzeri, and Tim Gollisch. Inference of neuronal functional circuitry with spike-triggered non-negative matrix factorization. *Nature communications*, 8(1):1–14, 2017. 5.1

[40] Yonatan Loewenstein, Annerose Kuras, and Simon Rumpel. Multiplicative dynamics underlie the emergence of the log-normal distribution of spine sizes in the neocortex in vivo. *Journal of Neuroscience*, 31(26):9481–9488, 2011. 1

[41] John Lott. Some geometric calculations on wasserstein space. *arXiv preprint math/0612562*, 2006. 2.3, A.2

[42] Bruno Loureiro, Cedric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mezard, and Lenka Zdeborová. Learning curves of generic features maps for realistic datasets with a teacher-student model. *Advances in Neural Information Processing Systems*, 34:18137–18151, 2021. 5.1

[43] David Marr. A theory of cerebellar cortex. *The Journal of Physiology*, 202(2):437, 1969. 1

[44] Tambet Matiisen, Avital Oliver, Taco Cohen, and John Schulman. Teacher–student curriculum learning. *IEEE transactions on neural networks and learning systems*, 31(9):3732–3740, 2019. 5.1

[45] Ludovic Métivier, Romain Brossier, Quentin Mérigot, Edouard Oudet, and Jean Virieux. Measuring the misfit between seismograms using an optimal transport distance: Application to full waveform inversion. *Geophysical Supplements to the Monthly Notices of the Royal Astronomical Society*, 205(1):345–377, 2016. A.2

[46] Ludovic Métivier, Romain Brossier, Quentin Merigot, Edouard Oudet, and Jean Virieux. An optimal transport approach for seismic tomography: Application to 3d full waveform inversion. *Inverse Problems*, 32(11):115008, 2016. A.2

[47] Gábor Molnár, Szabolcs Oláh, Gergely Komlósi, Miklós Füle, János Szabadics, Csaba Varga, Pál Barzó, and Gábor Tamás. Complex events initiated by individual spikes in the human cerebral cortex. *PLoS biology*, 6(9):e222, 2008. 1, A.1.2

[48] Gianluigi Mongillo, Simon Rumpel, and Yonatan Loewenstein. Inhibitory connectivity defines the realm of excitatory plasticity. *Nature neuroscience*, 21(10):1463–1470, 2018. 4

[49] Grégoire Montavon, Klaus-Robert Müller, and Marco Cuturi. Wasserstein training of restricted boltzmann machines. *Advances in Neural Information Processing Systems*, 29, 2016. A.2

[50] Mor Shpigel Nacson, Nathan Srebro, and Daniel Soudry. Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3051–3059. PMLR, 2019. 3

[51] Michael Okun and Ilan Lampl. Instantaneous correlation of excitation and inhibition during ongoing and sensory-evoked activities. *Nature neuroscience*, 11(5):535–537, 2008. 4

[52] RW Penney and D Sherrington. The weight-space of the binary perceptron. *Journal of Physics A: Mathematical and General*, 26(22):6173, 1993. A.5.1, A.5.1

[53] Cindy Poo and Jeffry S Isaacson. Odor representations in olfactory cortex:"sparse" coding, global inhibition, and oscillations. *Neuron*, 62(6):850–861, 2009. 4

[54] Esteban Real, Hiroki Asari, Tim Gollisch, and Markus Meister. Neural circuit inference from function to structure. *Current Biology*, 27(2):189–198, 2017. 5.1

[55] Edmund T Rolls, Alessandro Treves, and Edmund T Rolls. *Neural networks and brain function*, volume 572. Oxford university press Oxford, 1998. 1

[56] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958. 1

[57] Ran Rubin, LF Abbott, and Haim Sompolinsky. Balanced excitation and inhibition are required for high-capacity, noise-robust neuronal selectivity. *Proceedings of the National Academy of Sciences*, 114(44):E9366–E9375, 2017. 1, 2.1, 4

[58] Maziar Sanjabi, Jimmy Ba, Meisam Razaviyayn, and Jason D Lee. On the convergence and robustness of training gans with regularized optimal transport. *Advances in Neural Information Processing Systems*, 31, 2018. 6

[59] Louis K Scheffer, C Shan Xu, Michal Januszewski, Zhiyuan Lu, Shin-ya Takemura, Kenneth J Hayworth, Gary B Huang, Kazunori Shinomiya, Jeremy Maitlin-Shepard, Stuart Berg, et al. A connectome and analysis of the adult drosophila central brain. *Elife*, 9:e57443, 2020. 1

[60] Hyunjune Sebastian Seung, Haim Sompolinsky, and Naftali Tishby. Statistical mechanics of learning from examples. *Physical review A*, 45(8):6056, 1992. 5.1, 5.1, 5.2, A.3.1

[61] Alexander Shapson-Coe, Michał Januszewski, Daniel R Berger, Art Pope, Yuelong Wu, Tim Blakely, Richard L Schalek, Peter H Li, Shuohong Wang, Jeremy Maitin-Shepard, et al. A connectomic study of a petascale fragment of human cerebral cortex. *BioRxiv*, 2021. 1

[62] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018. 3

[63] Jun-nosuke Teramae and Tomoki Fukai. Computational implications of lognormally distributed synaptic weights. *Proceedings of the IEEE*, 102(4):500–512, 2014. 1

[64] Alex M Thomson, David C West, Yun Wang, and A Peter Bannister. Synaptic connections and small circuits involving excitatory and inhibitory neurons in layers 2–5 of adult rat and cat neocortex: triple intracellular recordings and biocytin labelling in vitro. *Cerebral cortex*, 12(9):936–953, 2002. A.1.2

[65] Matthew Thorpe. Introduction to optimal transport. *Lecture Notes*, 2019. 3, A.2, A.2

[66] Son Ngoc Truong and Kyeong-Sik Min. New memristor-based crossbar array architecture with 50-% area reduction and 48-% power saving for matrix-vector multiplication of analog neuromorphic computing. *JSTS: Journal of Semiconductor Technology and Science*, 14(3):356–363, 2014. 6

[67] Misha V Tsodyks and Terrance Sejnowski. Rapid state switching in balanced cortical network models. *Network: Computation in Neural Systems*, 6(2):111, 1995. 4

[68] C Van Vreeswijk and H Sompolinsky. Irregular activity in large networks of neurons. In *Les Houches*, volume 80, pages 341–406. Elsevier, 2005. 4

[69] Carl Van Vreeswijk and Haim Sompolinsky. Chaos in neuronal networks with balanced excitatory and inhibitory activity. *Science*, 274(5293):1724–1726, 1996. 4

[70] Carl van Vreeswijk and Haim Sompolinsky. Chaotic balanced state in a model of cortical circuits. *Neural computation*, 10(6):1321–1371, 1998. 4

[71] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009. A.2

[72] Michael Wehr and Anthony M Zador. Balanced inhibition underlies tuning and sharpens spike timing in auditory cortex. *Nature*, 426(6965):442–446, 2003. 4

[73] Jian-Ming Yang, Jing Zhang, Xiao-Juan Chen, Hong-Yan Geng, Mao Ye, Nicholas C Spitzer, Jian-Hong Luo, Shu-Min Duan, and Xiao-Ming Li. Development of gaba circuitry of fast-spiking basket interneurons in the medial prefrontal cortex of erbb4-mutant mice. *Journal of Neuroscience*, 33(50):19724–19733, 2013. 1, A.1.2

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section 3.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]

(b) Did you describe the limitations of your work? [Yes] See Section 2.2 and Appendix A.5.
(c) Did you discuss any potential negative societal impacts of your work? [No] We do not see any potential negative societal impacts of our work.
(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 2.2.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix A.1-A.3.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Section 3.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section A.4.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] Our theory concerns the thermodynamic limit where the observables are dominated by their typical values and fluctuations are small.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Appendix

Preliminaries

Throughout the appendix, we make frequent use of Gaussian integrals. We introduce short-hand notations $\int Dt \equiv \int \frac{dt}{\sqrt{2\pi}} e^{-t^2/2}$ and $H(x) \equiv \int_x^\infty Dt$. Also, when we do not specify the integration range it is understood that we are integrating from $-\infty$ to ∞.

A.1 Capacity supplemental materials

A.1.1 Replica calculation of distribution-constrained capacity

In this section, we present the replica calculation of the distribution-constrained storage capacity of a perceptron.

As described in main text Eqn.2, we need to perform a quenched average $\langle \cdot \rangle$ over the patterns $\boldsymbol{\xi}^\mu$ and labels ζ^μ for $\log V$, which can be carried out using the replica trick, $\langle \log V \rangle = \lim_{n\to 0}(\langle V^n \rangle - 1)/n$.

Following [29, 28], we consider first integer n, and at the end perform analytic continuation of $n \to 0$. The replicated Gardner volume is:

$$V = \frac{\prod_{\alpha=1}^{n} \int d\boldsymbol{w}^{\alpha} \left[\prod_{\mu=1}^{P} \Theta\left(\zeta^{\mu} \frac{\boldsymbol{w}^{\alpha}\cdot\boldsymbol{\xi}^{\mu}}{||\boldsymbol{w}^{\alpha}||} - \kappa\right)\right] \delta(||\boldsymbol{w}^{\alpha}||^2 - N)\delta\left(\int dk\Big(\hat{q}(k) - q(k)\Big)\right)}{\prod_{\alpha=1}^{n} \int d\boldsymbol{w}^{\alpha}\delta(||\boldsymbol{w}^{\alpha}||^2 - N)} \tag{13}$$

Let's rewrite the Heaviside step function using Fourier representation of the δ-function $\delta(x) = \int_{-\infty}^{\infty} \frac{dk}{2\pi} e^{ikx}$ as (defining $z_{\alpha}^{\mu} = \zeta^{\mu} \frac{\boldsymbol{w}^{\alpha}\cdot\boldsymbol{\xi}^{\mu}}{||\boldsymbol{w}^{\alpha}||}$)

$$\Theta\left(z_{\alpha}^{\mu} - \kappa\right) = \int_{\kappa}^{\infty} d\rho_{\alpha}^{\mu} \delta(\rho_{\alpha}^{\mu} - z_{\alpha}^{\mu}) = \int_{\kappa}^{\infty} d\rho_{\alpha}^{\mu} \int \frac{dx_{\alpha}^{\mu}}{2\pi} e^{ix_{\alpha}^{\mu}(\rho_{\alpha}^{\mu} - z_{\alpha}^{\mu})}. \tag{14}$$

Note that now all the $\boldsymbol{\xi}^{\mu}, \zeta^{\mu}$ dependence is in $e^{-ix_{\alpha}^{\mu} z_{\alpha}^{\mu}}$. We perform the average with respect to $\xi_i^{\mu} \sim p(\xi_i^{\mu}) = \mathcal{N}(0,1)$ and $p(\zeta^{\mu}) = \frac{1}{2}\delta(\zeta^{\mu}+1) + \frac{1}{2}\delta(\zeta^{\mu}-1)$ (also note that $||\boldsymbol{w}^{\alpha}|| = \sqrt{N}$):

$$\begin{aligned}
\left\langle \prod_{\mu\alpha} e^{-ix_{\alpha}^{\mu} z_{\alpha}^{\mu}} \right\rangle_{\xi\eta} &= \prod_{\mu j} \left\langle \exp\left\{-\frac{i}{\sqrt{N}} \zeta^{\mu} \xi_j^{\mu} \sum_{\alpha} x_{\alpha}^{\mu} w_j^{\alpha}\right\} \right\rangle_{\xi\zeta} \\
&= \prod_{\mu i} \left\langle \exp\left\{-\frac{(\zeta^{\mu})^2}{2N} \sum_{\alpha\beta} x_{\alpha}^{\mu} x_{\beta}^{\mu} w_i^{\alpha} w_i^{\beta}\right\} \right\rangle_{\zeta} \\
&= \prod_{\mu} \exp\left\{-\frac{1}{2N} \sum_{\alpha\beta} x_{\alpha}^{\mu} x_{\beta}^{\mu} \sum_{i} w_i^{\alpha} w_i^{\beta}\right\}.
\end{aligned} \tag{15}$$

Introducing the replica overlap parameter $q_{\alpha\beta} = \frac{1}{N}\sum_i w_i^{\alpha} w_i^{\beta}$, and notice that the μ index gives P identical copies of the same integral. We can suppress the μ indices and write

$$\left\langle \prod_{\mu\alpha} \Theta\left(z_{\alpha}^{\mu} - \kappa\right) \right\rangle_{\xi\zeta} = \left[\int_{\kappa}^{\infty} \left(\prod_{\alpha} \frac{d\rho_{\alpha} dx_{\alpha}}{2\pi}\right) e^{K}\right]^{P}, \tag{16}$$

where

$$K = i\sum_{\alpha} x_{\alpha}\rho_{\alpha} - \frac{1}{2}\sum_{\alpha\beta} q_{\alpha\beta} x_{\alpha} x_{\beta} \tag{17}$$

captures all the data dependence in the quenched free energy landscape, and therefore it is called the 'energetic' part of the free energy. In contrast, the δ-functions in Eqn.13 are called 'entropic' part because they regulate what kind of weights are considered in the version space (space of viable weights).

The entropic part

$$\delta(Nq_{\alpha\beta} - \sum_i w_i^{\alpha} w_i^{\beta}) = \int \frac{d\hat{q}_{\alpha\beta}}{2\pi} \exp\left\{iN\hat{q}_{\alpha\beta} q_{\alpha\beta} - i\hat{q}_{\alpha\beta} \sum_i w_i^{\alpha} w_i^{\beta}\right\}. \tag{18}$$

Note that the normalization constraint $\delta(||\boldsymbol{w}^{\alpha}||^2 - N)$ is automatically satisfied by requiring $q_{\alpha\alpha} = 1$. Using replica-symmetric ansatz: $\hat{q}_{\alpha\beta} = -\frac{i}{2}(\Delta\hat{q}\delta_{\alpha\beta} + \hat{q}_1)$, and $q_{\alpha\beta} = (1-q)\delta_{\alpha\beta} + q$, we have

$$iN\sum_{\alpha\beta}\hat{q}_{\alpha\beta}q_{\alpha\beta}=\frac{nN}{2}\left[\Delta\hat{q}+\hat{q}_1(1-q)\right]+\mathcal{O}(n^2). \tag{19}$$

and

$$\begin{aligned}
-i\sum_{\alpha\beta}\hat{q}_{\alpha\beta}\sum_i w_i^\alpha w_i^\beta &= -\frac{1}{2}(\Delta\hat{q}+\hat{q}_1)\sum_\alpha\sum_i (w_i^\alpha)^2-\frac{1}{2}\hat{q}_1\sum_{(\alpha\beta)}\sum_i w_i^\alpha w_i^\beta \\
&= -\frac{1}{2}\Delta\hat{q}\sum_\alpha\sum_i (w_i^\alpha)^2-\frac{1}{2}\hat{q}_1\sum_i\left(\sum_\alpha w_i^\alpha\right)^2 \\
&\overset{\text{HST}}{=} -\frac{1}{2}\Delta\hat{q}\sum_\alpha\sum_i (w_i^\alpha)^2+\sqrt{-\hat{q}_1}\sum_i t_i\left(\sum_\alpha w_i^\alpha\right),
\end{aligned} \tag{20}$$

where in the last step HST denotes Hubbard-Stratonovich transformation $\int\frac{dt}{\sqrt{2\pi}}e^{-t^2/2}e^{bt}=e^{b^2/2}$ that we use to linearize the quadratic term at the cost of introducing an auxiliary Gaussian variable t to be averaged over later.

Recall that $\hat{q}(k)=\int e^{ikw}\hat{p}(w)=\frac{1}{N}\sum_i^N e^{ikw_i^\alpha}$, the distribution constraint becomes

$$\begin{aligned}
\delta\left(\int dk\,(\hat{q}(k)-q(k))\right) &= \delta\left(\int dk\left(\frac{1}{N}\sum_i^N e^{ikw_i^\alpha}-q(k)\right)\right) \\
&= \int\frac{d\hat{\lambda}_\alpha(k)}{2\pi}\exp\left\{\int dki\hat{\lambda}_\alpha(k)\left(\sum_i e^{ikw_i^\alpha}-Nq(k)\right)\right\}.
\end{aligned} \tag{21}$$

Note that $\sum_i\int dki\hat{\lambda}_\alpha(k)e^{ikw_i^\alpha}=2\pi i\sum_i\lambda_\alpha(-w_i^\alpha)$ by inverse Fourier transform. Next,

$$\begin{aligned}
-iN\int dk\hat{\lambda}_\alpha(k)q(k) &= -iN\int dk\left(\int dwe^{ikw}\lambda_\alpha(w)\right)\left(\int dw'e^{ikw'}q(w')\right) \\
&= -2\pi iN\int dwdw'\lambda_\alpha(w)q(w')\delta(w+w') \\
&= -2\pi iN\int dwq(w)\lambda_\alpha(-w).
\end{aligned} \tag{22}$$

Now we can write down the full free energy. We ignore overall constant coefficients such as 2π's and i's in the integration measure, which become irrelevant upon taking the saddle-point approximation. We also leave out the denominator of V, as it does not depend on data and is an overall constant. Note that under the replica-symmetric ansatz the replica index α gives n identical copies of the same integral and thus the replica index α can be suppressed (same for synaptic index i):

$$\langle V^n\rangle=\int dqd\hat{\lambda}(k)d\Delta\hat{q}d\hat{q}_1e^{nN(G_0+G_1)}, \tag{23}$$

where (please note that q is replica overlap, and $q(w)$ is the imposed target distribution)

$$
\begin{aligned}
G_0 &= \frac{1}{2}\Delta\hat{q} + \frac{1}{2}\hat{q}_1(1-q) - 2\pi i \int dw q(w)\lambda(-w) + \langle \log Z(t)\rangle_t \,, \\
Z(t) &= \int dw \exp\left\{2\pi i\lambda(-w) - \frac{1}{2}\Delta\hat{q}w^2 + \sqrt{-\hat{q}_1}tw\right\}.
\end{aligned} \tag{24}
$$

Note that integrals in Eqn.23 can be evaluated using saddle-point approximation in the thermodynamic limit $N \to \infty$.

Redefining $2\pi i\lambda(-w) - \frac{1}{2}\Delta\hat{q}w^2 \to -\lambda(w)$ and $-\hat{q}_1 \to \hat{q}_1$, we have

$$
\begin{aligned}
G_0 &= \frac{1}{2}\Delta\hat{q} - \frac{1}{2}\hat{q}_1(1-q) + \int dw q(w)\lambda(w) - \frac{1}{2}\Delta\hat{q}\int dw q(w)w^2 + \langle \log Z(t)\rangle_t \,, \\
Z(t) &= \int dw \exp\left\{-\lambda(w) + \sqrt{\hat{q}_1}tw\right\}.
\end{aligned} \tag{25}
$$

We seek the saddle-point solution for G_0 with respect to the order parameters $\Delta\hat{q}$, $\lambda(w)$, and $\hat{q}_1$:

$$
0 = \frac{\partial G_0}{\partial \Delta\hat{q}} \Rightarrow 1 = \int dw q(w)w^2 = \left\langle w^2\right\rangle_{q(w)} \,, \tag{26}
$$

$$
0 = \frac{\partial G_0}{\partial \lambda(w)} \Rightarrow q(w) = \left\langle \frac{1}{Z(t)} \exp\left\{-\lambda(w) + \sqrt{\hat{q}_1}tw\right\}\right\rangle. \tag{27}
$$

We observe that the saddle-point equation Eqn.26 fixes the second moment of the imposed distribution $q(w)$ to 1 and therefore can be thought of as a second moment constraint. G_0 now simplifies to

$$
G_0 = -\frac{1}{2}\hat{q}_1(1-q) + \int dw q(w)\lambda(w) + \langle \log Z(t)\rangle_t \,. \tag{28}
$$

The remaining $\hat{q}_1$ saddle-point equation is a bit more complicated,

$$
0 = \frac{\partial G_0}{\partial \hat{q}_1} = -\frac{1}{2}(1-q) + \frac{t}{2\sqrt{\hat{q}_1}}\left\langle \frac{1}{Z(t)}\int dw w \exp\left\{-\lambda(w) + \sqrt{\hat{q}_1}tw\right\}\right\rangle_t \tag{29}
$$

Integration by parts for the second term in rhs:

$$
\begin{aligned}
1-q =& \frac{1}{\sqrt{\hat{q}_1}}\int Dt \frac{1}{Z}\sqrt{\hat{q}_1}\int dw w^2 \exp\left\{-\lambda(w) + \sqrt{\hat{q}_1}tw\right\} \\
&- \frac{1}{\sqrt{\hat{q}_1}}\int Dt \frac{1}{Z^2}\sqrt{\hat{q}_1}\left(\int dw w \exp\left\{-\lambda(w) + \sqrt{\hat{q}_1}tw\right\}\right)^2 \\
=& \left\langle \left\langle w^2\right\rangle_{f(w)}\right\rangle_t - \left\langle \left\langle w\right\rangle^2_{f(w)}\right\rangle_t \,,
\end{aligned} \tag{30}
$$

where in the last step we have defined an induced distribution $f(w) = Z(t)^{-1}\exp\left\{-\lambda(w) + \sqrt{\hat{q}_1}tw\right\}$. Since the second moments are fixed to 1, we have

$$
q = \left\langle \left\langle w\right\rangle^2_{f(w)}\right\rangle_t \,, \tag{31}
$$

which gives a nice interpretation of q in terms of the average overlap of w in the induced distribution $f(w)$.

Limit $q \to 1$

We are interested in the critical load α_c where the version space (space of viable weights) shrinks to a single point, i.e., there exists only one viable solution. Since q measures the typical overlap between weight vectors in the version space, the uniqueness of the solution implies $q \to 1$ at α_c. In this limit, the order parameters $\{\hat{q}_1, \lambda(w)\}$ diverges and we need to re-derive the saddle point equations Eqn.27 and Eqn.31 in terms of the undiverged order parameters $\{u, r(w)\}$:

$$\hat{q}_1 = \frac{u^2}{(1-q)^2}; \qquad \lambda(w) = \frac{r(w)}{1-q}. \tag{32}$$

Now G_0 becomes

$$G_0 = \frac{1}{1-q}\left\{-\frac{1}{2}u^2 + \int dw q(w) r(w) + (1-q)\left\langle \log Z(t)\right\rangle_t\right\}, \tag{33}$$

and

$$Z(t) \;=\; \int dw \exp \frac{1}{1-q}\left\{-r(w) + utw\right\}. \tag{34}$$

We can perform a saddle-point approximation for the w integral in $Z(t)$ at the saddle value w such that $r'(w) = ut$:

$$Z(t) = \exp\left\{\frac{-r(w) + utw}{1-q}\right\}. \tag{35}$$

Then

$$G_0 = \frac{1}{1-q}\left\{-\frac{1}{2}u^2 + \int dw q(w) r(w) - \left\langle r(w)\right\rangle_t + u\left\langle tw\right\rangle\right\}. \tag{36}$$

Let's use integration by parts to rewrite

$$\begin{aligned} \int dw q(w) r(w) &= -\int Q(w) r'(w) dw \\ \left\langle r(w)\right\rangle_t &= \int \frac{dt}{\sqrt{2\pi}} e^{-t^2/2} r(w) = -\int P(t) r'(w) dw, \end{aligned} \tag{37}$$

where $Q(w)$ is the CDF of the imposed distribution $q(w)$ and $P(t) = \frac{1}{2}\left[1 + \text{Erf}(\frac{t}{\sqrt{2}})\right]$ is the normal CDF.

Now the saddle-point equation

$$0 = \frac{\partial G_0}{\partial r'(w)} \Rightarrow Q(w) = P(t) \tag{38}$$

determines $w(t)$ implicitly. The u equation gives

$$0 = \frac{\partial G_0}{\partial u} \Rightarrow u = \left\langle tw\right\rangle_t = \left\langle \frac{dw}{dt}\right\rangle_t \tag{39}$$

where in the last equality we have used integration by parts. Using Eqn.38-39 G_0 is simplified to

$$G_0 = \frac{1}{2(1-q)} \left\langle \frac{dw}{dt} \right\rangle_t^2 . \tag{40}$$

The energetic part

We would like to perform a similar procedure as shown above, to Eqn.17 using the replica-symmetric ansatz. We observe that the effect of the distribution constraint is entirely captured in G_0 and therefore G_1 is unchanged compared with the standard Gardner calculation of perceptron capacity. We reproduce the calculation here for completeness.

Under the replica-symmetric ansatz $q_{\alpha\beta} = (1-q)\delta_{\alpha\beta} + q$, Eqn.17 becomes

$$\begin{aligned} K &= i \sum_\alpha x_\alpha \rho_\alpha - \frac{1-q}{2} \sum_\alpha x_\alpha^2 - \frac{q}{2} \left(\sum_\alpha x_\alpha \right)^2 \\ &\overset{\text{HST}}{=} i \sum_\alpha x_\alpha \rho_\alpha - \frac{1-q}{2} \sum_\alpha x_\alpha^2 - it\sqrt{q} \sum_\alpha x_\alpha . \end{aligned} \tag{41}$$

where we have again used the Hubbard-Stratonovich transformation to linearize the quadratic piece. Performing the Gaussian integrals in x_α (define $\alpha = \frac{P}{N}$),

$$nG_1 = \alpha \log \left[\left\langle \int_\kappa^\infty \frac{d\rho}{\sqrt{2\pi(1-q)}} \exp\left\{ -\frac{(\rho + t\sqrt{q})^2}{2(1-q)} \right\} \right\rangle_t^n \right] . \tag{42}$$

At the limit $n \to 0$,

$$nG_1 = \alpha n \left\langle \log \left[\int_\kappa^\infty \frac{d\rho}{\sqrt{2\pi(1-q)}} \exp\left\{ -\frac{(\rho + t\sqrt{q})^2}{2(1-q)} \right\} \right] \right\rangle_t . \tag{43}$$

Perform the Gaussian integral in ρ and define $\tilde{\kappa} = \frac{\kappa + t\sqrt{q}}{\sqrt{1-q}}$, we have

$$G_1 = \alpha \int Dt \log H(\tilde{\kappa}). \tag{44}$$

At the limit $q \to 1, \alpha \to \alpha_c$, $\int_{-\infty}^\infty Dt$ is dominated by $\int_{-\kappa}^\infty Dt$, and $H(\tilde{\kappa}) \to \frac{1}{\sqrt{2\pi}\tilde{\kappa}} e^{-\tilde{\kappa}^2/2}$. The $\mathcal{O}\left(\frac{1}{1-q}\right)$ (leading order) contribution gives

$$G_1 = -\frac{1}{2(1-q)} \alpha_c \int_{-\kappa}^\infty Dt(\kappa + t)^2. \tag{45}$$

Let $G = G_0 + G_1$. As $n \to 0$, $\langle V^n \rangle = e^{n(NG)} \to 1 + n\,(NG)$, and $\langle \log V \rangle = \lim_{n\to 0} \frac{\langle V^n \rangle - 1}{n} = NG$.

Combining with Eqn.40 (relabel $t \leftrightarrow x$ to distinguish between the two auxiliary Gaussian variables), we have

$$\langle \log V \rangle = \frac{N}{2(1-q)} \left[\left\langle \frac{dw}{dx} \right\rangle_x^2 - \alpha_c \int_{-\kappa}^\infty Dt(\kappa + t)^2 \right] \tag{46}$$

Capacity α_c is reached when Eqn.13 goes to zero. We arrive at the distribution-constrained capacity

$$\alpha_c(\kappa) = \alpha_0(\kappa) \left\langle \frac{dw}{dx} \right\rangle_x^2, \tag{47}$$

where $\alpha_0(\kappa) = \left[\int_{-\kappa}^{\infty} Dt(\kappa+t)^2\right]^{-1}$ is the unconstrained capacity.

Instructive Examples

(1) Standard normal distribution $w \sim \mathcal{N}(0,1)$.

In this case $w = x$ and $\alpha_c(\kappa) = \alpha_0(\kappa)$.

(2) Normal distribution with nonzero mean $w \sim \mathcal{N}(\mu, \sigma^2)$. This is the example discussed in the main text Fig.1.

In this case $w = \mu + \sigma x$ and $\mu^2 + \sigma^2 = 1$ due to the second moment constraint Eqn.26. Then $\alpha_c(\kappa) = \sigma^2 \alpha_0(\kappa)$.

(3) Lognormal distribution $w \sim \frac{1}{\sqrt{2\pi}w} \exp\left\{-\frac{(\ln w - \mu)^2}{2\sigma^2}\right\}$.

In this case $w = e^{\mu + \sigma x}$ where $\mu = -\sigma^2$. $\alpha_c(\kappa) = \sigma^2 e^{-\sigma^2} \alpha_0(\kappa)$.

Geometrical interpretation

Note that although the Jacobian factor $\left\langle \frac{dw}{dx} \right\rangle_x$ takes a simple form, in practice sometimes it might not be the most convenient form to use. Integrating by parts ($p(x) = \mathcal{N}(0,1)$),

$$\left\langle \frac{dw}{dx} \right\rangle_x = \int dx p(x) w x \tag{48}$$

Now define $u = P(x)$ so that $du = p(x)dx$ and $w = Q^{-1}(P(x)) = Q^{-1}(u)$, we can express the Jacobian in terms of the CDFs

$$\left\langle \frac{dw}{dx} \right\rangle_x = \int_0^1 du \left(Q^{-1}(u) P^{-1}(u)\right) \tag{49}$$

Furthermore,

$$\begin{aligned} \left\langle \frac{dw}{dx} \right\rangle_x &= \frac{1}{2}\left[\int_0^1 du \left(Q^{-1}(u)\right)^2 + \int_0^1 du \left(P^{-1}(u)\right)^2 - \int_0^1 du \left(Q^{-1}(u) - P^{-1}(u)\right)^2\right] \\ &= \frac{1}{2}\left[2 - W_2(P,Q)^2\right], \end{aligned} \tag{50}$$

where we have used second moments equal to 1 and the definition of the Wasserstein-k distance in the second equality. Therefore, we have arrived at the geometric interpretation of the Jacobian term

$$\left\langle \frac{dw}{dx} \right\rangle_x = 1 - \frac{1}{2} W_2(P,Q)^2. \tag{51}$$

A.1.2 Theory for an arbitrary number of synaptic subpopulations

In this section, we generalize our theory in the above section to the set up of a perceptron with M synaptic populations indexed by m, $\boldsymbol{w}^m$, such that each w_i^m satisfies its own distributions constraints $w_i^m \sim q_m(w^m)$. We denote the overall weight vector as $\boldsymbol{w} \equiv \{\boldsymbol{w}^m\}_{m=1}^M \in \mathbb{R}^{N\times 1}$, where the total number of weights is $N = \sum_{m=1}^M N_m$. The replica overlap now becomes $q_{\alpha\beta} = \frac{1}{N}\sum_m^M \sum_i^{N_m} w_i^{m\alpha} w_i^{m\beta}$. The distribution constraint becomes (see Eqn.21 for the case of $M = 1$)

$$\prod_m \delta\left(\int dk^m \left(\frac{1}{N_m}\sum_i^{N_m} e^{ik^m w_i^{m\alpha}} - q_m(k^m)\right)\right). \tag{52}$$

We introduce $\hat{q}_{\alpha\beta}, \lambda_m(k)$ to write the δ-functions into Fourier representations, and use replica-symmetric ansatz $\hat{q}_{\alpha\beta} = -\frac{i}{2}(\Delta\hat{q}\delta_{\alpha\beta} + \hat{q}_1)$, and $q_{\alpha\beta} = (1-q)\delta_{\alpha\beta} + q$ as before. After similar manipulations that lead to Eqn.25, the entropic part of the free energy becomes ($g_m = N_m/N$ is the fraction of weights in m-th population)

$$\begin{aligned} G_0 =& \frac{1}{2}\Delta\hat{q} - \frac{1}{2}\hat{q}_1(1-q) + \sum_m g_m \int dw^m q_m(w^m)\lambda_m(w^m) \\ &- \frac{1}{2}\Delta\hat{q}\sum_m g_m \int dw^m q_m(w^m)\left(w^m\right)^2 + \sum_m g_m \left\langle \log Z_m(t)\right\rangle_t, \\ Z_m(t) =& \int dw^m \exp\left\{-\lambda_m(w^m) + \sqrt{\hat{q}_1} t w^m\right\}. \end{aligned} \tag{53}$$

Now the second moment constraint $0 = \partial G_0/\partial\Delta\hat{q}$ (Eqn.26) becomes the weighted sum of second moments from each population:

$$1 = \sum_m g_m \int dw^m q_m(w^m)\left(w^m\right)^2 = \sum_m g_m \left\langle \left(w^m\right)^2\right\rangle_{q_m}. \tag{54}$$

We take the $q \to 1$ limit as before:

$$\hat{q}_1 = \frac{u^2}{(1-q)^2}; \qquad \lambda_m(w^m) = \frac{r_m(w^m)}{1-q}. \tag{55}$$

Use saddle-point approximation for $Z_m(t)$ and integrate by parts as in Eqn.35-37, the entropic part becomes

$$G_0 = \frac{1}{1-q}\left\{-\frac{1}{2}u^2 + \sum_m g_m r_m'(w^m)\left[P(x) - Q_m(w^m)\right] + u\sum_m g_m \left\langle t w^m\right\rangle_t\right\}. \tag{56}$$

Now the saddle-point equation for order parameters $r_m'(w^m)$ and u gives

$$\begin{aligned} P(x) &= Q_m(w^m) \\ u &= \sum_m g_m \left\langle t w^m\right\rangle_t = \sum_m g_m \left\langle \frac{dw^m}{dt}\right\rangle_t. \end{aligned} \tag{57}$$

Therefore,

$$G_0 = \frac{1}{2(1-q)}\left[\sum_m g_m \left\langle \frac{dw^m}{dt}\right\rangle_t\right]^2. \tag{58}$$

The energetic part (Eqn.35) remains unchanged and thus (relabel $t \leftrightarrow x$)

$$\alpha_c(\kappa) = \alpha_0(\kappa)\left[\sum_m g_m \left\langle \frac{dw^m}{dx}\right\rangle_x\right]^2. \tag{59}$$

E/I balanced lognormals

Now we specialize to the biologically realistic E/I balanced lognormal distributions described in the main text. We are interested the case with two synaptic populations $m = E, I$ that models the excitatory/inhibitory synpatic weights of a biological neuron. $w_i^E \sim \frac{1}{\sqrt{2\pi}\sigma_E w^E} \exp\left\{-\frac{(\ln w^E - \mu_E)^2}{2\sigma_E^2}\right\}$ and $w_i^I \sim \frac{1}{\sqrt{2\pi}\sigma_I w^I} \exp\left\{-\frac{(\ln w^I - \mu_I)^2}{2\sigma_I^2}\right\}$. Let's denote the E/I fractions as $g_E = r$ and $g_I = 1 - r$. The CDF of the lognormals are given by

$$Q_m(w^m) = H\left[\frac{1}{\sigma_m}\left(\mu_m - \ln w^m\right)\right]. \tag{60}$$

The corresponding inverse CDF is

$$Q_m^{-1}(u) = \exp\left\{\mu_m - \sigma_m H^{-1}(u)\right\}. \tag{61}$$

The capacity is therefore

$$\begin{aligned}\alpha_c &= \alpha_0 \left[\sum_m g_m \int_0^1 du Q_m^{-1}(u) P^{-1}(u)\right]^2 \\ &= \alpha_0 \left[r \int_0^1 du H^{-1}(u) \exp\left\{\mu_E - \sigma_E H^{-1}(u)\right\} + (1-r) \int du H^{-1}(u) \exp\left\{\mu_I - \sigma_I H^{-1}(u)\right\}\right]^2.\end{aligned} \tag{62}$$

This model has five parameters $\{r, \sigma_E, \sigma_I, \mu_E, \mu_I\}$. We use values of r reported in experiments (the ratio between of E. connections found and I. connections found).

We also have two constraints. The E/I balanced constraint $g_E \left\langle w^E \right\rangle_{q_E} = g_I \left\langle w^I \right\rangle_{q_I}$:

$$re^{\mu_E + \frac{1}{2}\sigma_E^2} = (1-r)e^{\mu_I + \frac{1}{2}\sigma_I^2}, \tag{63}$$

and the second moment constraint $1 = \sum_m g_m \left\langle (w^m)^2 \right\rangle_{q_m}$:

$$1 = re^{2(\mu_E + \sigma_E^2)} + (1-r)e^{2(\mu_I + \sigma_I^2)}. \tag{64}$$

Therefore there are two free parameters left and we choose to express μ_E and μ_I in terms of the rest:

$$\begin{aligned}\mu_I &= -\frac{1}{2}\sigma_I^2 - \ln(1-r) - \frac{1}{2}\ln\left[\frac{e^{\sigma_I^2}}{1-r} + \frac{e^{\sigma_E^2}}{r}\right] \\ \mu_E &= -\frac{1}{2}\sigma_E^2 - \ln r - \frac{1}{2}\ln\left[\frac{e^{\sigma_I^2}}{1-r} + \frac{e^{\sigma_E^2}}{r}\right].\end{aligned} \tag{65}$$

The parameter landscape is plotted against the two free parameters σ_E and σ_I. Here we report comparisons across different experiments [38, 9, 32, 47, 64, 73] similar to main text Fig.4 (Fig.4 (a) is included here for reference). Note that despite the apparently different shape of distributions, all the experimentally measured parameter values are within the first quantile of the optimal values predicted by our theory.

A.1.3 Capacity for biased inputs and sparse label

In this section, we generalized our theory in Section A.1.1 to the set up of nonzero-mean input patterns $\boldsymbol{\xi}^\mu$ and sparse labels ζ^μ:

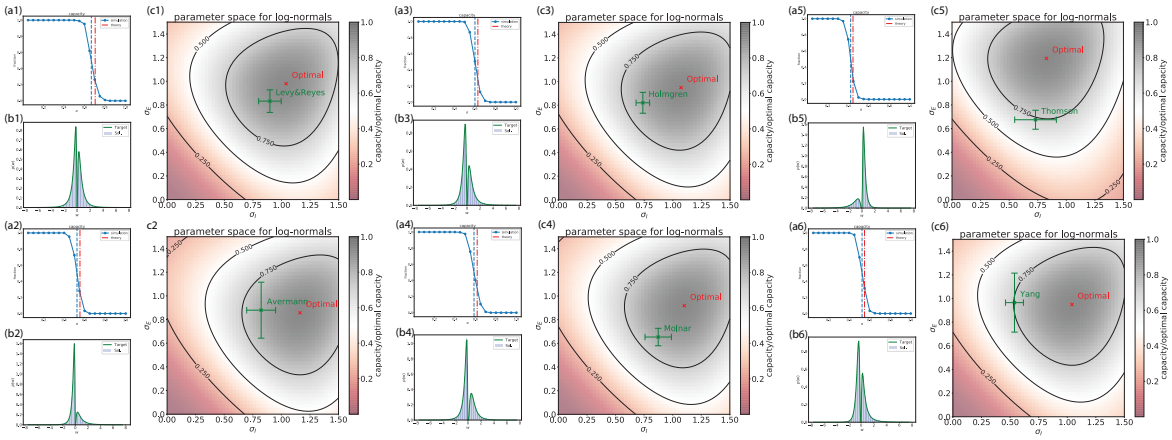

Figure 6: Additional parameter landscape for the biologically-realistic distribution. (a)-(b) (theory from main text Eqn.10 and simulations from DisCo-SGD): (a) Determination of capacity; (b) Example of weight distribution obtained in simulation. (c) Capacity (normalized by the optimal value in the landscape) as a function of the lognormal parameters σ_E and σ_I. Experimental value is shown in green with error bars, and optimal capacity is shown in red.

$$
\begin{aligned}
p(\xi_i^\mu) =& \mathcal{N}(m, 1-m^2) \\
p(\zeta^\mu) =& f\delta(\zeta^\mu - 1) + (1-f)\delta(\zeta^\mu + 1).
\end{aligned}
\tag{66}
$$

In this case, we need to include a bias in the perceptron $\hat{\zeta}^\mu = \text{sgn}(\frac{\boldsymbol{w}\cdot\boldsymbol{\xi}^\mu}{\|\boldsymbol{w}\|} - b)$ to be able to correctly classify patterns in general.

Note that $m = 0$ and $f = 1/2$ reduces to the case in Section A.1.1. We observe due to the multiplicative relation between the Jacobian term and the original Gardner capacity in Eqn.47, entropic effects (such as distribution constraints and sign-constraints) factors with the energetic effects (such as the nonzero mean inputs and sparse labels), and they don't interfere with each other. Therefore, the calculations for nonzero mean inputs and sparse labels are identical with the original Gardner case. Here we only reproduce the calculation for completeness. Readers already familiar with this calculation should skip this part.

The analog of Eqn.15 reads (define the local fields as $h_i^\mu = \sum_\alpha x_\alpha^\mu w_i^\alpha$)

$$
\begin{aligned}
\prod_{\mu\alpha} \left\langle e^{-\frac{i}{\sqrt{N}} x_\alpha^\mu \zeta^\mu \boldsymbol{\xi}^\mu \cdot \boldsymbol{w}^\alpha} \right\rangle_{\xi\zeta} =& \prod_{\mu i} \left\langle \exp\left\{ -\frac{i}{\sqrt{N}} \zeta^\mu \xi_i^\mu h_i^\mu \right\} \right\rangle_{\xi\zeta} \\
=& \prod_{\mu i} \left\langle \exp\left\{ -\frac{im}{\sqrt{N}} \zeta^\mu h_i^\mu - \frac{1}{2N}(1-m^2)\left(h_i^\mu\right)^2 \right\} \right\rangle_{\zeta} \\
=& \prod_{\mu} \left\langle \exp\left\{ -im\zeta^\mu \sum_\alpha x_\alpha^\mu M_\alpha - \frac{1-m^2}{2} \sum_{\alpha\beta} x_\alpha^\mu x_\beta^\mu q_{\alpha\beta} \right\} \right\rangle_{\zeta},
\end{aligned}
\tag{67}
$$

where in the second equality we have carried out the Gaussian integral in $\boldsymbol{\xi}^\mu$ and in the third equality we introduced the order parameters

$$
q_{\alpha\beta} = \frac{1}{N} \sum_i w_i^\alpha w_i^\beta, \qquad M_\alpha = \frac{1}{\sqrt{N}} \sum_i w_i^\alpha. \tag{68}
$$

Now the full energetic term becomes

$$
\begin{aligned}
&\left\langle \Theta\left(\frac{1}{\sqrt{N}}\zeta^\mu \boldsymbol{\xi}^\mu\cdot\boldsymbol{w}^\alpha - b\zeta^\mu-\kappa\right)\right\rangle_{\xi\zeta}\\
&=\prod_\mu\left\langle\int_{\kappa+b\zeta^\mu}^\infty\frac{d\lambda_\alpha^\mu}{2\pi}\int dx_\alpha^\mu\exp\left\{-im\zeta^\mu\sum_\alpha x_\alpha^\mu M_\alpha-\frac{1-m^2}{2}\sum_{\alpha\beta}x_\alpha^\mu x_\beta^\mu q_{\alpha\beta}\right\}\right\rangle_\zeta\\
&=f\prod_\mu\int_{\kappa+b}^\infty\frac{d\lambda_\alpha^\mu}{2\pi}\int dx_\alpha^\mu\exp\left\{i\sum_\alpha x_\alpha^\mu\left(\lambda_\alpha^\mu-mM_\alpha\right)-\frac{1-m^2}{2}\sum_{\alpha\beta}x_\alpha^\mu x_\beta^\mu q_{\alpha\beta}\right\}\\
&+(1-f)\prod_\mu\int_{\kappa-b}^\infty\frac{d\lambda_\alpha^\mu}{2\pi}\int dx_\alpha^\mu\exp\left\{i\sum_\alpha x_\alpha^\mu\left(\lambda_\alpha^\mu+mM_\alpha\right)-\frac{1-m^2}{2}\sum_{\alpha\beta}x_\alpha^\mu x_\beta^\mu q_{\alpha\beta}\right\}\\
&=f\prod_\mu\int_{\frac{\kappa+b-mM_\alpha}{\sqrt{1-m^2}}}^\infty\frac{d\lambda_\alpha^\mu}{2\pi}\int dx_\alpha^\mu\exp\left\{i\sum_\alpha x_\alpha^\mu\lambda_\alpha^\mu-\frac{1}{2}\sum_{\alpha\beta}x_\alpha^\mu x_\beta^\mu q_{\alpha\beta}\right\}\\
&+(1-f)\prod_\mu\int_{\frac{\kappa-b+mM_\alpha}{\sqrt{1-m^2}}}^\infty\frac{d\lambda_\alpha^\mu}{2\pi}\int dx_\alpha^\mu\exp\left\{i\sum_\alpha x_\alpha^\mu\lambda_\alpha^\mu-\frac{1}{2}\sum_{\alpha\beta}x_\alpha^\mu x_\beta^\mu q_{\alpha\beta}\right\}.
\end{aligned}
$$

Now G_1 becomes

$$
G_1=\frac{1}{1-q}\left\{f\int_{\frac{\kappa-b+mM}{\sqrt{1-m^2}}}^\infty Dt\left(t+\frac{\kappa+b-mM}{\sqrt{1-m^2}}\right)^2+(1-f)\int_{\frac{-\kappa-b-mM}{\sqrt{1-m^2}}}^\infty Dt\left(t+\frac{\kappa-b+mM}{\sqrt{1-m^2}}\right)^2\right\}. \tag{69}
$$

Note that the hat-variables $\hat{M}_\alpha$ conjugated with M_α are in subleading order to $\hat{q}_{\alpha\beta}$ in the thermodynamic limit, and therefore G_0 is unchanged. Let $v=M-b/m$, we have now the capacity

$$
\alpha_c(\kappa)=\left\langle\frac{dw}{dx}\right\rangle_x^2\left[f\int_{\frac{-\kappa+mv}{\sqrt{1-m^2}}}^\infty Dt\left(t+\frac{\kappa-mv}{\sqrt{1-m^2}}\right)^2+(1-f)\int_{\frac{-\kappa-mv}{\sqrt{1-m^2}}}^\infty Dt\left(t+\frac{\kappa+mv}{\sqrt{1-m^2}}\right)^2\right]^{-1}, \tag{70}
$$

where the order parameter v needs to be determined from the saddle-point equation

$$
f\int_{\frac{-\kappa+mv}{\sqrt{1-m^2}}}^\infty Dt\left(t+\frac{\kappa-mv}{\sqrt{1-m^2}}\right)=(1-f)\int_{\frac{-\kappa-mv}{\sqrt{1-m^2}}}^\infty Dt\left(t+\frac{\kappa+mv}{\sqrt{1-m^2}}\right). \tag{71}
$$

In Fig.7 we numerically solve $\alpha_c(\kappa)$ for different values of m and f.

A.2 Optimal transport theory

In recent years, Wasserstein distances has found diverse applications in fields ranging from machine learning [7, 26, 49] to geophysics [23, 24, 19, 45, 46]. In optimal transport theory, the Wasserstein-k distance arise as the minimal cost one needs to pay in transporting one probability distribution to another, when the moving cost between probability masses are measured by the L_k norm [71]. When one equips the probability density manifold with the Wasserstein-2 distance as metric, it becomes the Wasserstein space, a Riemannian manifold of real-valued distributions with a constant nonnegative sectional curvature [41, 25, 20]. Note that in our statistical mechanical theory main text Eqn.3-5,

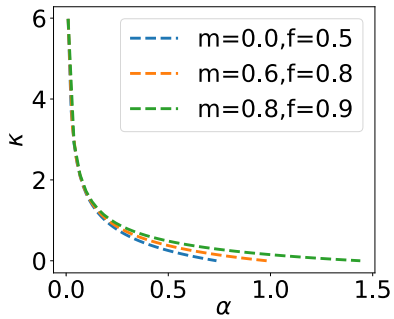

Figure 7: $\alpha_c(\kappa)$ for different values of input mean m and label sparsity f. Note that the blue curve corresponds to the vanilla case shown in main text Fig.4(c).

the Wasserstein-2 distance naturally arises in the mean-field limit without assuming any a priori transportation cost.

Here we briefly review the theory of optimal transport. Intuitively, optimal transport concerns the problem of finding the shortest path of morphing one distribution into another. In the following, we will use the *Monge* formulation [65, 3].

Given probability distributions P and Q with supports X and Y, we say that $T : X \to Y$ is a transport map from P to Q if the *push-forward* of P through T, $T_\# P$, equals Q:

$$Q = T_\# P \equiv P(T^{-1}(Y)). \tag{72}$$

Eqn.72 can be understood as moving probability masses $x \in X$ from distribution P to $y \in Y$ according to transportation map T, such that upon completion the distribution over Y becomes Q.

We are interested in finding a transportation plan that minimizes the transportation cost as measured by some distance function $d : X \times Y \to \mathbb{R}$:

$$C(T; d) = \int_X d(T(x), x) p(x) dx \qquad \text{s.t. } T_\# P = Q. \tag{73}$$

The transportation plan that minimizes Eqn.73 is called the optimal transport plan $T^* = \text{argmin}_T C(T; d)$. When the distance function d is chosen to be the L_k norm, the minimal cost becomes the Wasserstein-k distance:

$$W_k(P, Q) = \inf_T C(T; L_k)|_{T_\# P = Q}. \tag{74}$$

In 1-dimension, the Wasserstein-k distance has a closed form given by main text Eqn.6, and the optimal transport map has an explicit formula in terms of the CDFs: $T^* = Q^{-1} \circ P$. An example of the optimal transport map and how it moves probability masses between distributions is given in Fig.8 for transport between $p(w) = \mathcal{N}(0, 1)$ and $q(w) = \frac{1}{\sqrt{2\pi}\sigma w} \exp\left\{\frac{(\ln w - \mu)^2}{2\sigma^2}\right\}$. Note that in this case, the optimal transport plan is simply $T^*(x) = e^{\mu + \sigma x}$.

Now consider the manifold $\mathcal{M}$ of real-valued probability distributions, where points on this manifold are probability measures that admits a probability density function. When endowed with the W_k metric, $(\mathcal{M}, W_k)$ becomes a metric space and is in particular a geodesic space [65, 3]. We can explicitly construct the geodesics connecting points on $\mathcal{M}$. We parameterize the geodesic by the *geodesic time* $\tau \in [0, 1]$. Then given T^* an optimal transport plan, the intermediate probability distributions along the geodesic take the following form [65]:

$$P_\tau = ((1 - \tau)\text{Id} + \tau T^*)_\# P \tag{75}$$

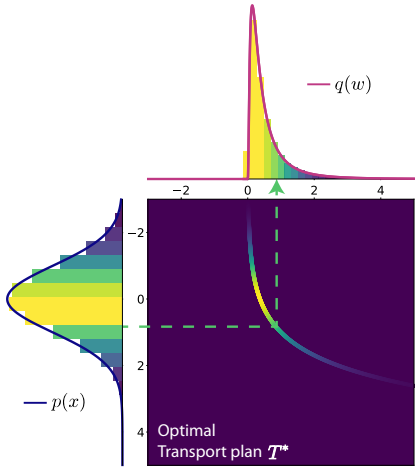

Figure 8: An example optimal transport plan from standard normal, $p(x)$, to a lognormal distribution $q(w)$. The optimal transport plan T^* is plotted in between the distributions. T^* moves $p(x)$ units of probability mass x to location w, as indicated by the dashed line, and the colors are chosen to reflect the amount of probability mass to be transported.

where Id is the identity map and P_τ is a constant speed geodesic connecting $P_{\tau=0} = P$ and $P_{\tau=1} = Q$.

For the discrete case, we can describe the sample $\{w_i^\tau\}$ from P_τ in a simple manner in terms of the samples $\{w_i\}$ drawn from P and $\{\hat{w}_i\}$ drawn from Q. We can arrange the samples in the ascending order, or equivalently, forming their order statistics $\left\{x_{(i)} : x_{(1)} \leq ... \leq x_{(N)}\right\}$, which can be thought of as atoms in a discrete measure. Then in terms of the order statistics,

$$w_{(i)}^\tau = (1-\tau)w_{(i)} + \tau\hat{w}_{(i)} \tag{76}$$

Upon infinitetesimal change in the geodesic time, $\tau \to \tau + \delta\tau$, the geodesic flow becomes

$$w_{(i)}^{\tau+\delta\tau} = w_{(i)}^\tau + \delta\tau\left(\hat{w}_{(i)} - w_{(i)}\right) \tag{77}$$

Specializing to the case discussed in main text Section 3, $w_{(i)} = w_{(i)}^{\tau=0}$ is the initialization for the perceptron weight and therefore just a constant, we can promoted it $w_{(i)} \to w_{(i)}^\tau$ to fix the overall scale in the perceptron weight, then we arrive at main text Eqn.9.

A.3 Generalization supplemental materials

A.3.1 Replica calculation of generalization with sign-constraint

In this section, we calculate the sign-constraint teacher-student setup. To ease notation, let's denote the teacher perceptron $\boldsymbol{w_t} \equiv \boldsymbol{w}^0$ and the (replicated) student perceptron $\boldsymbol{w}_s^a \equiv \boldsymbol{w}^a$. Given random inputs $\boldsymbol{\xi}^\mu$ with $p(\xi_i^\mu) = \mathcal{N}(0,1)$, we generate labels by $\zeta^\mu = \text{sgn}(\boldsymbol{w}^0 \cdot \boldsymbol{\xi}^\mu / ||\boldsymbol{w}^0|| + \eta^\mu)$, where η^μ is input noise and $\eta^\mu \sim \mathcal{N}(0, \sigma^2)$. Let's denote the signs of the teacher perceptron as $s_i = \text{sgn}(w_i^0)$. The student perceptron's weights are constrained to have the same sign as that of the teacher's, so we insert $\Theta(s_i w_i^a)$ in the Gardner volume to enforce this constraint (we leave out the denominator part of V as it does not depend on data and is an overall constant):

$$\langle V^n \rangle_{\xi\eta w^0} = \prod_{\alpha=1}^{n} \left\langle \int_{-\infty}^{\infty} \frac{d\boldsymbol{w}^a}{\sqrt{2\pi}} \prod_{\mu=1}^{p} \Theta\left(\text{sgn}\left(\frac{\boldsymbol{w^0}\cdot\boldsymbol{\xi}^\mu}{||\boldsymbol{w^0}||} + \eta^\mu\right)\frac{\boldsymbol{w^a}\cdot\boldsymbol{\xi}^\mu}{||\boldsymbol{w^a}||} - \kappa\right) \prod_i^N \Theta(s_i w_i^a) \right\rangle_{\xi\eta w^0}. \tag{78}$$

We observe that upon redefining $s_i w_i^a \to w_i^a, s_i \xi_i^\mu \to \xi_i^\mu$, we can absorb the sign-constraints into the integration range of w from $[-\infty, +\infty]$ to $[0, \infty]$:

$$\langle V^n \rangle_{\xi\eta w^0} = \prod_{\alpha=1}^{n} \left\langle \int_0^\infty \frac{d\boldsymbol{w}^a}{\sqrt{2\pi}} \prod_{\mu=1}^{p} \Theta\left(\text{sgn}\left(\frac{\boldsymbol{w^0} \cdot \boldsymbol{\xi}^\mu}{||\boldsymbol{w^0}||} + \eta^\mu \right) \frac{\boldsymbol{w^a} \cdot \boldsymbol{\xi}^\mu}{||\boldsymbol{w^a}||} - \kappa \right) \right\rangle_{\xi\eta w^0}. \tag{79}$$

Therefore, sign constraint amounts to restricting all the weights to be positive. In the following, we denote $\int_0^\infty$ as $\int_c$.

Let's define the local fields as

$$h_\mu^a = \frac{\boldsymbol{w^a} \cdot \boldsymbol{\xi}^\mu}{\sqrt{N}}; \qquad h_\mu^0 = \frac{\boldsymbol{w^0} \cdot \boldsymbol{\xi}^\mu}{\sqrt{N}} + \eta^\mu \tag{80}$$

We leave the average over teacher w^0 to the very end.

$$\begin{aligned}
\langle V^n \rangle_{\xi\eta} &= \prod_{\mu a} \int_c \frac{d\boldsymbol{w}^a}{\sqrt{2\pi}} \int dh_\mu^a \Theta\Big(\text{sgn}(h_\mu^0) h_\mu^a - \kappa \Big) \left\langle \delta\left(h_\mu^a - \frac{\boldsymbol{w^a} \cdot \boldsymbol{\xi}^\mu}{\sqrt{N}} \right) \right\rangle_{\xi\eta} \\
&= \int_c (\prod_{a=1}^{n} \frac{d\boldsymbol{w}^a}{\sqrt{2\pi}}) \int \prod_{\mu a} \frac{dh_\mu^a d\hat{h}_\mu^a}{2\pi} \int \prod_{\mu} \frac{dh_\mu^0 d\hat{h}_\mu^0}{2\pi} \prod_{\mu a} \Theta\Big(\text{sgn}(h_\mu^0) h_\mu^a - \kappa \Big) \\
&\times \left\langle \exp\left\{ \sum_{\mu a} \left(i\hat{h}_\mu^a h_\mu^a - i\hat{h}_\mu^a \frac{\boldsymbol{w^a} \cdot \boldsymbol{\xi}^\mu}{\sqrt{N}} \right) + \sum_{\mu} \left(i\hat{h}_\mu^0 h_\mu^0 - i\hat{h}_\mu^0 \frac{\boldsymbol{w^0} \cdot \boldsymbol{\xi}^\mu}{\sqrt{N}} - i\hat{h}_\mu^0 \eta^\mu \right) \right\} \right\rangle_{\xi\eta} \\
&= \int_c (\prod_{a=1}^{n} \frac{d\boldsymbol{w}^a}{\sqrt{2\pi}}) \int \prod_{\mu a} \frac{dh_\mu^a d\hat{h}_\mu^a}{2\pi} \int \prod_{\mu} \frac{dh_\mu^0 d\hat{h}_\mu^0}{2\pi} \prod_{\mu a} \Theta\Big(\text{sgn}(h_\mu^0) h_\mu^a - \kappa \Big) \\
&\times \exp\left\{ \sum_{\mu a} i\hat{h_\mu^a} h_\mu^a + \sum_{\mu} i\hat{h}_\mu^0 h_\mu^0 \right\} \\
&\times \prod_{\mu} \exp\left\{ -\frac{1}{2N} \left[\sum_{a,b} \hat{h}_\mu^a \hat{h}_\mu^b \sum_i w_i^a w_i^b + N \left(\hat{h}_\mu^0 \right)^2 + 2 \sum_a \hat{h}_\mu^a \hat{h}_\mu^0 \sum_i w_i^a w_i^0 \right] \right\},
\end{aligned} \tag{81}$$

where in the last step we perform the average over noise $\eta^\mu \sim \mathcal{N}(0, \sigma^2)$ and patterns $p(\xi_i^\mu) = \mathcal{N}(0, 1)$, and make use of the normalization conditions $\sum_i (w_i^0)^2 = N$ and $\sum_i (w_i^a)^2 = N$.

Now let's define order parameters

$$q_{ab} = \frac{1}{N} \sum_i w_i^a w_i^b, \qquad R_a = \frac{1}{N} \sum_i w_i^a w_i^0. \tag{82}$$

We introduce conjugate variables $\hat{q}_{ab}$ and $\hat{R}_a$ to write the δ-functions into its Fourier representations, and after some algebraic manipulations we can bring the Gardner volume into the following form ($\alpha \equiv p/N$):

$$\langle \langle V^n \rangle \rangle_{\xi,z} = \int (\prod_a d\hat{q}_1^a)(\prod_{ab} dq^{ab} d\hat{q}^{ab})(\prod_a dR^a d\hat{R}^a) e^{nNG}, \tag{83}$$

where ($\bar{h}_\mu^0 = \gamma h_\mu^0; \quad \gamma = 1/\sqrt{1+\sigma^2}$)

$$
\begin{aligned}
nG =& nG_0 + \alpha n G_E \\
nG_0 =& -\frac{1}{2}\sum_{ab}\hat{q}^{ab}q^{ab} - \sum_a \hat{R}^a R^a + n\left\langle \ln Z\right\rangle_{w^0}, \\
Z =& \int_c \left(\prod_a \frac{dw_i^a}{\sqrt{2\pi}}\right)\exp\left\{\frac{1}{2}\sum_a \hat{q}_1^a (w_i^a)^2 + \frac{1}{2}\sum_{a\neq b}\hat{q}^{ab}w_i^a w_i^b + \sum_a \hat{R}^a w_i^a w_i^0\right\}, \\
nG_1 =& \ln\int\prod_a \frac{d\hat{h}^a dh^a}{2\pi}\int D\bar{h}^0 \prod_a \Theta\left(\text{sgn}(\frac{\bar{h}^0}{\gamma})h^a - \kappa\right) \\
&\times \exp\left\{i\sum_a \hat{h}^a h^a - i\gamma\bar{h}^0\sum_a h^a R^a - \frac{1}{2}\sum_a(\hat{h}^a)^2[1-(\gamma R^a)^2] - \frac{1}{2}\sum_{a\neq b}\hat{h}^a\hat{h}^b(q^{ab}-\gamma^2 R^a R^b)\right\}.
\end{aligned}
\tag{84}
$$

The energetic part G_1 is the same as the unconstrained case in [60, 22]. After standard manipulations, we have

$$
G_1 = 2\int DtH\left(-\frac{\gamma R t}{\sqrt{q-\gamma^2 R^2}}\right)\ln H\left(\frac{\kappa-\sqrt{q}t}{\sqrt{1-q}}\right). \tag{85}
$$

Entropic part

In this subsection, we perform the integrals in the entropic part, and we will see novel terms coming from the constraint on the student's integration range.

We start by assuming a replica-symmetric solution for the auxiliary variables introduced in the Fourier decomposition of the δ-functions,

$$
\hat{R}^a = \hat{R}; \qquad \hat{q}^{ab} = \hat{q} + (\hat{q}_1 - \hat{q})\delta_{ab}; \qquad \hat{q}_1^a = \hat{q}_1; \qquad m_i^a = m_i; \qquad \hat{m}_i^a = \hat{m}_i, \tag{86}
$$

and $q_{ab} = (1-q)\delta_{ab} + q$.

Then the entropic part,

$$
\begin{aligned}
Z =& \int\left(\prod_a \frac{dw_i^a}{\sqrt{2\pi}}\right)\exp\left\{\frac{1}{2}(\hat{q}_1-\hat{q})\sum_a (w_i^a)^2 + \hat{R}w_i^0\sum_a w_i^a + \frac{1}{2}\hat{q}(\sum_a w_i^a)^2\right\} \\
\overset{\text{HST}}{=}& \int Dt\int_c (\prod_a \frac{dw_i^a}{\sqrt{2\pi}})\exp\left\{\frac{1}{2}(\hat{q}_1-\hat{q})\sum_a (w_i^a)^2 + (\hat{R}w_i^0 + t\sqrt{\hat{q}})\sum_a w_i^a\right\},
\end{aligned}
\tag{87}
$$

where we have introduced Gaussian variable t to linearize quadratic term as usual. Now the integral becomes n identical copies and we can drop the replica index a,

$$
G_0 = -\frac{1}{2}\hat{q}_1 + \frac{1}{2}\hat{q}q - \hat{R}R + \langle \ln Z\rangle_{t,w^0}. \tag{88}
$$

We can bring the log term into the form of an induced distribution $f(w)$,

$$
\begin{aligned}
Z =& \int_0^\infty \frac{dw}{\sqrt{2\pi}}\exp\left[-f(w)\right] \\
f(w) =& \frac{1}{2}(\hat{q}-\hat{q}_1)w^2 - (\hat{R}w^0 + t\sqrt{\hat{q}})w
\end{aligned}.
\tag{89}
$$

Under saddle-point approximation, we obtain a set of mean field self-consistency equations for the order parameters:

$$\begin{aligned} 0 &= \frac{\partial G_0}{\partial \hat{q}_1} \Rightarrow 1 = \left\langle \left\langle w^2 \right\rangle_f \right\rangle_{t,w^0} \\ 0 &= \frac{\partial G_0}{\partial \hat{R}} \Rightarrow R = \left\langle w^0 \left\langle w \right\rangle_f \right\rangle_{t,w^0}, \\ 0 &= \frac{\partial G_0}{\partial \hat{q}} \Rightarrow q = \left\langle \left\langle w \right\rangle_f^2 \right\rangle_{t,w^0} \end{aligned} \tag{90}$$

$$\begin{aligned} 0 &= \frac{\partial G_1}{\partial q} \Rightarrow \hat{q} = -2\alpha \partial_q G_1 \\ 0 &= \frac{\partial G_1}{\partial R} \Rightarrow \hat{R} = \alpha \partial_R G_1 \end{aligned} . \tag{91}$$

$q \to 1$ **limit**

In this limit the order parameter diverges, and we define the new set of undiverged order parameters as

$$\hat{R} = \frac{\tilde{R}}{1-q}; \qquad \hat{q} = \frac{\tilde{q}^2}{(1-q)^2}; \qquad \hat{q} - \hat{q}_1 = \frac{\Delta}{1-q}. \tag{92}$$

Then

$$\begin{aligned} f(w) =& \frac{1}{1-q}\left[\frac{1}{2}\Delta w^2 - (\tilde{R}w^0 + t\tilde{q})w\right] \\ =& \frac{1}{1-q}\left[\frac{1}{2}\Delta\left(w - \frac{1}{\Delta}(\tilde{R}w^0 + t\tilde{q})\right)^2 - \frac{1}{2\Delta}(\tilde{R}w^0 + t\tilde{q})^2\right]. \end{aligned} \tag{93}$$

Then $\langle w \rangle_f = \frac{1}{\Delta}\left(\tilde{R}w^0 + t\tilde{q}\right)$, and the integral over the auxiliary variable is dominated by values of t such that $\tilde{R}w^0 + t\tilde{q} > 0$. In the following, we denote $\left\langle [g(t)]_+ \right\rangle_t$ as integrating over range of t such that $g(t) > 0$. Then the self-consistency equations Eqn.90 take a compact form (after rescaling order parameters $\tilde{R} \to \tilde{R}\Delta$, $\tilde{q} \to \tilde{q}\Delta$)

$$\begin{aligned} 1 =& \frac{1}{\Delta}\left\langle \Theta(\tilde{R}w^0 + t\tilde{q}) \right\rangle_{t,w^0} \\ 1 =& \left\langle \left[\tilde{R}w^0 + t\tilde{q}\right]_+^2 \right\rangle_{t,w^0} \\ R =& \left\langle w^0 \left[\tilde{R}w^0 + t\tilde{q}\right]_+ \right\rangle_{t,w^0} \end{aligned} , \tag{94}$$

Eqn.90 becomes ($\tilde{\kappa} = \kappa/\sqrt{1-\gamma^2 R^2}$)

$$\begin{aligned} \tilde{R}\Delta =& \frac{\alpha\gamma}{\sqrt{2\pi}}\sqrt{1-\gamma^2R^2}\int_{-\tilde{\kappa}}^{\infty} Dt\left(\tilde{\kappa} + t\right) \\ \frac{\Delta}{2}\left(2 - \tilde{q}^2\Delta - 2\tilde{R}R\right) =& \alpha\int_{-\infty}^{\kappa} DtH\left(-\frac{\gamma R t}{\sqrt{1-\gamma^2R^2}}\right)(\kappa - t)^2 \end{aligned} . \tag{95}$$

The free energy is (recall that $\gamma = 1/\sqrt{1+\sigma^2}$)

$$G = \frac{1}{2(1-q)}\left(\Delta - \tilde{q}^2 - 2\tilde{R}R + \frac{1}{\Delta}\left\langle \left[\tilde{R}w^0 + t\tilde{q}\right]_+^2 \right\rangle_{t,w^0}\right) - \alpha \int_{-\infty}^{\kappa} DtH\left(-\frac{\gamma R t}{\sqrt{1-\gamma^2 R^2}}\right)(\kappa - t)^2. \tag{96}$$

A.3.2 Replica calculation of generalization with distribution-constraint

In this subsection, we will consider the case where student weights are constrained to some *prior* distribution $q_s(w_s)$, while the teacher obeys a distribution $p_t(w_t)$,for an arbitrary pair q_s, p_t. We can write down the Gardner volume V_g for generalization as in the capacity case (main text Eqn.2):

$$V_g = \frac{\int d\boldsymbol{w}_s \left[\prod_{\mu=1}^{P} \Theta\left(\text{sgn}\left(\frac{\boldsymbol{w}_t\cdot\boldsymbol{\xi}^\mu}{||\boldsymbol{w}_t||} + \eta^\mu\right)\frac{\boldsymbol{w}_s\cdot\boldsymbol{\xi}^\mu}{||\boldsymbol{w}_s||} - \kappa\right)\right]\delta(||\boldsymbol{w}_s||^2 - N)\delta\left(\int dk\left(\hat{q}(k) - q(k)\right)\right)}{\int d\boldsymbol{w}_s \delta(||\boldsymbol{w}_s||^2 - N)}. \tag{97}$$

We treat the distribution constraint $q_s(w)$ similar to Section A.1.1. The entropic part of the free energy becomes

$$\begin{aligned} G_0 =& -\frac{1}{2}\hat{q}_1 + \frac{1}{2}\hat{q}q - \hat{R}R + \int_{-\infty}^{\infty} dw q_s(w)\lambda(w) + \langle \ln Z\rangle_{t,w_t} \\ Z =& \int \frac{dw}{\sqrt{2\pi}} \exp\left[-f(w)\right] \\ f(w) =& \frac{1}{2}(\hat{q} - \hat{q}_1)w^2 - (\hat{R}w_t + t\sqrt{\hat{q}})w + \lambda(w) \end{aligned} . \tag{98}$$

At the limit $q \to 1$, we make the following ansatz

$$\hat{R} = \frac{\tilde{R}}{1-q}; \quad \hat{q} = \frac{u^2}{(1-q)^2}; \quad \hat{q} - \hat{q}_1 = \frac{\Delta}{1-q}; \quad \lambda(w) = \frac{r(w)}{1-q}. \tag{99}$$

Then

$$\begin{aligned} G_0 =& \frac{1}{(1-q)}\left(-\frac{1}{2}u^2 + \frac{1}{2}\Delta - \tilde{R}R + \int dw q_s(w) r(w)\right) + \langle \ln Z\rangle_{t,w_t} \\ f(w) =& \frac{1}{1-q}\left(\frac{1}{2}\Delta w^2 - (\tilde{R}w_t + ut)w + r(w)\right) \end{aligned} \tag{100}$$

We can absorb $\frac{1}{2}\Delta w^2$ into the definition of $r(w)$, $\frac{1}{2}\Delta w^2 + r(w) \to r(w)$, and $0 = \partial G_0/\partial\Delta$ gives the second moment constraint, $1 = \int dw q_s(w) w^2$.

Then,

$$\begin{aligned} G_0 =& \frac{1}{(1-q)}\left(-\frac{1}{2}u^2 - \tilde{R}R + \int dw q(w) r(w)\right) + \langle \ln Z\rangle_{t,w_t} \\ f(w) =& \frac{1}{1-q}\left(r(w) - (\tilde{R}w_t + ut)w\right) \end{aligned} . \tag{101}$$

Next, we perform a saddle-point approximation on the log-term in G_0,

$$Z = \int \frac{dw}{\sqrt{2\pi}} \exp\left[-f(w)\right] \approx \exp\left[-f(w_s)\right], \tag{102}$$

where w_s is the saddle-point value for the weight, and is determined implicitly by

$$r'(w_s) = \tilde{R}w_t + ut. \tag{103}$$

Note that $r'(w_s)$ is now an induced random variable from random variables w_t and t. For later convenience, we rescale $r'(w_s)$ to define a new random variable z,

$$z \equiv u^{-1}r'(w_s) = t + u^{-1}\tilde{R}w_t \equiv t + \varepsilon w_t, \tag{104}$$

where we have also defined

$$\varepsilon \equiv u^{-1}\tilde{R}. \tag{105}$$

The induced distribution on z is then

$$\tilde{p}(z) = \int Dt \int dw_t p(w_t)\delta(z - t - \varepsilon w_t). \tag{106}$$

Now the entropic part becomes

$$G_0 = \frac{1}{(1-q)}\left(-\frac{1}{2}u^2 - \tilde{R}R + \int dw q_s(w)r(w) + \langle(\tilde{R}w_t + ut)w_s\rangle_{t,w_t} - \langle r(w_s)\rangle_{t,w_t}\right). \tag{107}$$

Integrate by parts,

$$\int dw q(w)r(w) = -\int dw Q(w)r'(w), \tag{108}$$

$$\begin{aligned}
\langle r(w_s)\rangle_{t,w_t} &= \int Dt dw_t p_t(w_t)r(w_s) \\
&= \int dz\delta(z - t - \varepsilon w_t)\int Dt dw_t p_t(w_t)r(w_s) \\
&= \int dz\tilde{p}(z)r(w_s) \\
&= -\int dz\tilde{P}(z)r'(w_s)
\end{aligned}. \tag{109}$$

Now $0 = \partial G/\partial r'(w_s)$ gives

$$Q(w_s) = \tilde{P}(z). \tag{110}$$

which implicitly determines $w_s(z)$.

Next,

$$0 = \frac{\partial G}{\partial u} \Rightarrow u = \langle w_s(z)t\rangle_{t,w_t}, \tag{111}$$

$$0 = \frac{\partial G}{\partial \tilde{R}} \Rightarrow R = \langle w_s(z)w_t\rangle_{t,w_t}. \tag{112}$$

The free energy then simplifies to

$$G = \frac{u^2}{2\,(1-q)} + \alpha G_1. \tag{113}$$

The energetic part as $q \to 1$ becomes (same as the unconstrained and sign-constrained case)

$$G_1 = -\frac{1}{1-q}\int_{-\infty}^{\kappa} DtH\left(-\frac{\gamma Rt}{\sqrt{1-\gamma^2R^2}}\right)(\kappa - t)^2. \tag{114}$$

The remaining two saddle point equations are (1) the vanishing log-Gardner volume and (2) $0 = \partial G/\partial R$:

$$\frac{1}{2}u^2 = \alpha\int_{-\infty}^{\kappa} DtH\left(-\frac{\gamma Rt}{\sqrt{1-\gamma^2R^2}}\right)(\kappa - t)^2, \tag{115}$$

$$\varepsilon u = \alpha\gamma\sqrt{\frac{2}{\pi}}\sqrt{1-\gamma^2R^2}\int_{-\tilde{\kappa}}^{\infty} Dt\Big(\tilde{\kappa}+t\Big). \tag{116}$$

In summary, the order parameters $\{R, \kappa, u, \varepsilon\}$ can be determined from a set of self-consistency equations:

$$\begin{aligned} u =& \langle w_s(z)t\rangle_{t,w_t} \\ R =& \langle w_s(z)w_t\rangle_{t,w_t} \\ \frac{1}{2}u^2 =& \alpha\int_{-\infty}^{\kappa} DtH\Big(-\frac{\gamma Rt}{\sqrt{1-\gamma^2R^2}}\Big)(\kappa-t)^2, \\ \varepsilon u =& \frac{2\alpha\gamma}{\sqrt{2\pi}}\sqrt{1-\gamma^2R^2}\int_{-\tilde{\kappa}}^{\infty} Dt\Big(\tilde{\kappa}+t\Big) \end{aligned} \tag{117}$$

where we have introduced $\tilde{\kappa} = \kappa/\sqrt{1-\gamma^2R^2}$, an auxiliary normal variable $t \sim \mathcal{N}(0,1)$, and an induced random variable $z \equiv t + \varepsilon w_t$ with induced distribution

$$\tilde{p}(z) = \int Dt \int dw_t p_t(w_t)\delta(z-t-\varepsilon w_t). \tag{118}$$

Note that $w_s(z)$ can be determined implicitly by equating the CDF of the induced variable z and the distribution that the student is constrained to:

$$Q(w_s) = \tilde{P}(z). \tag{119}$$

Examples

(1) Lognormal distribution

In the following, we solve $w_s(z)$ explicitly from the CDF equation $Q(w_s) = \tilde{P}(z)$. For a lognormal teacher,

$$p_t(w_t) = \frac{1}{w_t}\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(\ln w_t-\mu)^2}{2\sigma^2}\right\}. \tag{120}$$

The second moment constraint implies $\mu = -\sigma^2$.

The induced CDF of z is

$$\tilde{P}(z) = \int_{-\infty}^{z} dz' \int_{-\infty}^{\infty} Dt \int_{0}^{\infty} dw_t p_t(w_t)\delta(z'-t-\varepsilon w_t). \tag{121}$$

Let $x = (\ln w - \mu)/\sigma$,

$$\begin{aligned} \tilde{P}(z) =& \int_{-\infty}^{z} dz' \int_{-\infty}^{\infty} Dt \int_{-\infty}^{\infty} Dx\delta(z'-t-\varepsilon e^{\mu+\sigma x}) \\ =& \int_{-\infty}^{\infty} DxH(\varepsilon e^{\mu+\sigma x}-z) \end{aligned}. \tag{122}$$

Now the CDF of w_s is

$$Q_s(w_s) = \int_{-\infty}^{w_s} q_s(w)dw = H\left(-\frac{\ln w_s-\mu}{\sigma}\right). \tag{123}$$

Therefore, equating $\tilde{P}(z)$ and $Q_s(w_s)$:

$$\int_{-\infty}^{\infty} DxH(\varepsilon e^{\mu+\sigma x} - z) = H\left(-\frac{\ln w_s - \mu}{\sigma}\right), \tag{124}$$

We can solve for $w_s(z)$ by (recall $z \equiv t + \varepsilon w_t$)

$$w_s(z) = \exp\left\{\mu + \sigma H^{-1}\left(\int DxH(z - \varepsilon e^{\mu+\sigma x})\right)\right\}. \tag{125}$$

Or in terms of error functions

$$w_s(z) = \exp\left\{\mu + \sqrt{2}\sigma \text{erf}^{-1}\left(\int Dx \text{erf}\left(\frac{\varepsilon e^{\mu+\sigma x} - z}{\sqrt{2}}\right)\right)\right\}. \tag{126}$$

We can also calculate the initial overlap (before any learning):

$$R_0 = \langle \boldsymbol{w}_t \cdot \boldsymbol{w}_s \rangle_{p_t q_s} = e^{2\mu+\sigma^2} = e^{-\sigma^2}. \tag{127}$$

(2) Uniform distribution

Assuming that both the teacher and the student have a uniform distribution in range $[0, \sigma]$.

The second moment constraint fixes $\sigma = \sqrt{3}$.

We can solve (as in the lognormal example above),

$$w_s(z) = \frac{1}{\varepsilon}\int_{-\infty}^{z} dz' \left(H(z' - \varepsilon\sigma) - H(z')\right). \tag{128}$$

(3) Half-normal distribution

Assuming that both the teacher and the student has a half-normal distribution $\frac{2}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{w^2}{2\sigma^2}\right\}$.

The second moment constraint fixes $\sigma = 1$, and

$$w_s(z) = \sigma H^{-1}\left\{\frac{1}{2} - \int_{-\infty}^{\frac{z}{\sqrt{1+\sigma^2\varepsilon^2}}} DtH(-\sigma\varepsilon t)\right\}. \tag{129}$$

Arbitrary number of synaptic subpopulations

Just like in the case of Section A.1.2, we can generalize our theory above to incorporate distribution constraints with an arbitrary number of synaptic subpopulations. Let's consider a student perceptron with M synaptic populations indexed by m, $\boldsymbol{w}^m$, such that each w_i^m satisfies its own distributions constraints $w_i^m \sim Q_m(w^m)$. We denote the overall weight vector as $\boldsymbol{w} \equiv \{\boldsymbol{w}^m\}_{m=1}^M \in \mathbb{R}^{N\times 1}$. The total number of weights is $N = \sum_{m=1}^M N_m$, and we denote the fractions as $g_m = N_m/N$. Since the derivation is similar to that of Section A.1.2 and Section A.3.2, we will only present the results here.

As before, the order parameters $\{R, \kappa, u, \varepsilon\}$ can be determined from a set of self-consistency equations:

$$
\begin{aligned}
u =& \sum_m g_m \langle w^m(z) t \rangle_{t,w_t} \\
R =& \sum_m g_m \langle w^m(z) w_t \rangle_{t,w_t} \\
\frac{1}{2}u^2 =& \alpha \int_{-\infty}^{\kappa} DtH\Big(-\frac{\gamma R t}{\sqrt{1-\gamma^2 R^2}} \Big)(\kappa - t)^2 \\
\varepsilon u =& \frac{2\alpha\gamma}{\sqrt{2\pi}} \sqrt{1-\gamma^2 R^2} \int_{-\tilde{\kappa}}^{\infty} Dt \Big(\tilde{\kappa} + t \Big)
\end{aligned}
\tag{130}
$$

where $\tilde{\kappa} = \kappa/\sqrt{1-\gamma^2 R^2}$, $t \sim \mathcal{N}(0,1)$. and an induced random variable $z \equiv t + \varepsilon w_t$ with induced distribution the same as Eqn.118.

Note that every $w^m(z)$ can be determined by equating the CDF of the induced variable z and the m-th distribution that $w^m(z)$ is constrained to:

$$
Q_m(w^m) = \tilde{P}(z). \tag{131}
$$

A.3.3 Sparsification of weights in sign-constraint learning

For unconstrained weights, max-margin solutions are considered beneficial for generalization particularly for small size training sets. As a first step toward biological plausibility, one can try to constraint the sign of individual weights during learning (e.g., excitatory or inhibitory). In the generalization error setup, we can impose a constraint that the teacher and student have the same set of weight signs. Surprisingly, we find both analytically and numerically that if the teacher weights are not too sparse, the max-margin solution generalizes poorly: after a single step of learning (with random input vectors), the overlap, R, drops substantially from its initial value R_0 (by a factor of $\sqrt{2}$ for a half-Gaussian teacher, see the blue curves in Fig.9(a).

We can verify this by calculating R_0 in two different ways. As an example, in the following we consider the case where both the teacher and student have half-normal distributions.

(1) By definition, the overlap is $R = \frac{\boldsymbol{w}_s \cdot \boldsymbol{w}_t}{\|\boldsymbol{w}_s\| \|\boldsymbol{w}_t\|}$. Since $\boldsymbol{w}_s$ and $\boldsymbol{w}_t$ are uncorrelated before learning ($\alpha = 0$), the initial overlap is then $R_0 = \frac{\langle w_s \rangle \langle w_t \rangle}{||\boldsymbol{w}_s||\boldsymbol{w}_t||} = \frac{2}{\pi}$;

(2) Take the $\alpha \to 0$ limit in Eqn.90 and Eqn.91 and calculate $R_{0+} = \lim_{\alpha \to 0+} R(\alpha) = \frac{\sqrt{2}}{\pi}$.

Therefore, in this example $R_{0+} = R_0/\sqrt{2}$.

The source of the problem is that due to the sign constraint, max-margin training with few examples yields a significant mismatch between the student and teacher weight distributions. After only a few steps of learning, half of the student's weights are set to zero, and the student's distribution, $p(w_s) = \frac{1}{2}\delta(0) + \frac{1}{\sqrt{2\pi}} \exp\{-\frac{w_s^2}{4}\}$, deviates significantly from the teacher's half-normal distribution (Fig.9(b)).

A.3.4 Noisy teacher

We generate examples $\{\boldsymbol{\xi}^\mu, \zeta^\mu\}_{\mu=1}^P$ from a teacher perceptron, $\boldsymbol{w}_t \in \mathbb{R}^N$: $\zeta^\mu = \text{sgn}(\boldsymbol{w}_t \cdot \boldsymbol{\xi}^\mu / ||\boldsymbol{w}_t|| + \eta^\mu)$, where η^μ is input noise and $\eta^\mu \sim \mathcal{N}(0, \sigma^2)$. In this subsection we present additional numerical results for the case when $\sigma \neq 0$. As in previous sections, we define the noise level parameter $\gamma = 1/\sqrt{1+\sigma^2}$.

Our theory's prediction is confirmed by numerical simulation for a wide range of teacher noise level γ and teacher weight distributions $P_t(w_t)$. We find that distribution-constrained learning performs consistently better all the way up to capacity (capacity in this framework is due to teacher noise). For illustration, in Fig.10 we show theory and simulation for fixed prior learning of three different teacher distributions: uniform, half-normal, and lognormal.

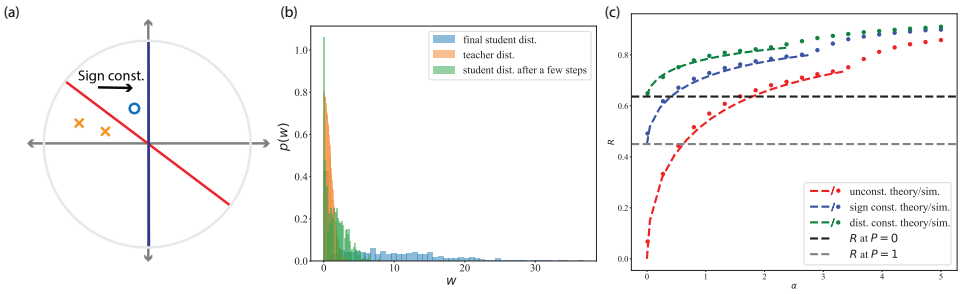

Figure 9: Sparsification of weights in sign-constraint learning. (a) An illustration of weight sparsification. In this schematic, the perceptron lives on this 1-dimensional circle and $N = 2$. Red line denotes the hyperplane orthogonal to the perceptron weight before sign-constraint, crosses and circles indicate examples in different classes. Sign-constraint pushes the weights to the first quadrant, which zeros half of the weights on average. Blue line indicates the hyperplane obtained after the sign-constraint. (b) Sparsification of weights due to max-margin training. After only a few iterations, nearly half of the student weights are set to zero, and the distribution deviates significantly from the teacher's distribution. (c) Teacher-student overlap as a function of load α for different learning paradigms. Dashed lines are from theory, and dots are from simulation. Note the horizontal dashed lines show the initial drop in overlap from zero example and to just a single example. In this case teacher has nonzero noise, $\gamma = 0.85$.

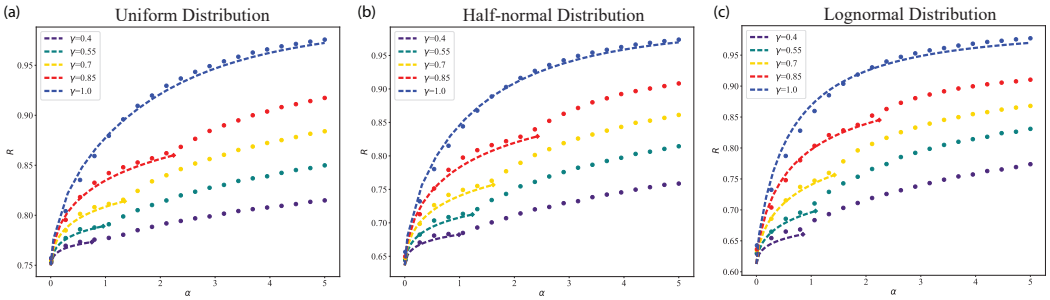

Figure 10: Generalization (measured by overlap) performance for different distributions and different noise levels in fixed prior learning. From left to right: uniform, half-normal, and lognormal distribution. In all cases the student is constrained to have the same distribution as that of the teacher's. Dashed lines are from theory and dots are from DisCo-SGD simulation.

A.4 DisCo-SGD simulations

Avoid vanishing gradients

Note that we often observe a vanishing gradient in DisCo-SGD when we choose a constant learning rate η_1, and in such cases the algorithm tends to find poor margin κ which deviates from the max-margin value predicted from the theory. We find that scaling η_1 with the standard deviation of the gradient solves this problem:

$$\eta_1 = \eta_1^0 / \text{std}\left(\sum_\mu \xi_i^\mu (\hat{\zeta}^\mu - \zeta^\mu)\right), \tag{132}$$

where the standard deviation is computed across the synaptic index i and η_1^0 is a constant.

Mini-batches

For the capacity simulations, we always use full-batch in the SGD update, so it is in fact simply gradient descent. However, in the case of generalization, we find that training with mini-batches

improves the generalization performance, since it acts as an source of stochasticity during training. In main text Fig.5 we use mini-batch size $B = 0.8P$ (80% of examples are used for each SGD update).

When we vary teacher's noise level, we find that scaling B with γ improves the quality of the solutions, as measured by the generalization performance (or equivalently, the teacher-student overlap). Generally, the more noisy the teacher is, the smaller the mini-batches should be. This is because smaller mini-batch size corresponds to higher stochasticity, which helps overcoming higher teacher noise.

Parameters

All the capacity simulations are performed with the following parameters $N = 1000, \eta_1^0 = 0.01, \eta_2 = 0.6, t_{max} = 10000$, where t_{max} is the maximum number of iterations of the DisCo-SGD algorithm.

All results are averaged over 300 realizations.

In main text Fig.4, the experimental [38] parameters are $g_E = 45.8\%, \sigma_E = 0.833, \sigma_I = 0.899$.

In main text Fig.5(a): We show the teacher-student overlap as a function of α. Dots are simulations performed with series of student distribution from $\sigma_s = 0.1$ to $\sigma_s = 1.4$, where the teacher distribution sits in the middle of this range, $\sigma_t = 0.7$. Each such simulation is performed with fixed σ_s and varying load $\alpha \in [0.05, 2.5]$. In main text Fig.5(b): we show the empirical weight distributions found by unconstrained perceptron learning for $\alpha \in [0.05, 10]$. In main text Fig.5(c) we show optimal student distribution for $\alpha \in [0.05, 2.5]$. Note that optimal prior learning approaches the teacher distribution much faster than unconstrained learning.

All the generalization DisCo-SGD simulations are performed with the same parameter as in the capacity DisCo-SGD simulations, but with two additional parameter: teacher's noise level γ and SGD mini-batch size B.

For the simulations in Fig.10 we use

$\gamma = 0.4, B = 0.2P; \gamma = 0.55, B = 0.4P; \gamma = 0.7, B = 0.6P; \gamma = 0.85, B = 0.8P; \gamma = 1.0, B = P$ (noiseless case).

A.5 Replica symmetry breaking

A.5.1 Bimodal distributions

In deriving the capacity formula, we have assumed replica-symmetry (RS). It is well-known that replica-symmetry breaking occurs in the Ising perceptron [52, 13], so it is natural to ask to what extent our theory holds when approaching the Ising limit. Let's consider a bimodal distribution with a mixture of two normal distributions with non-zero mean centered around zero,

$$p(w) = \frac{1}{2}\mathcal{N}(-\mu, \sigma) + \frac{1}{2}\mathcal{N}(\mu, \sigma)$$

The second moment constraint requires $\mu^2 + \sigma^2 = 1$.

We can gradually decrease the Gaussian width σ, or equivalently $\mu = \sqrt{1 - \sigma^2}$ (which we call 'separation' in the following) and compare the capacity theoretically predicted by the RS theory and numerically found by the DisCo-SGD algorithm.

In Fig.11 we can see that the simulation agrees well with the RS theory until one gets very close to the Ising limit ($\mu = 1$). To understand finite size effects, we extrapolate to the infinite size limit ($N \to \infty$) in Fig.12, and found that the deviation from RS theory has a sharp transition near $\mu = 1$, marking the breakdown of the RS theory.

Ising perceptron

It is also interesting to compare our distribution-constrained RS theory to the unconstrained RS theory. In this Ising limit,

$$q(w) = \frac{1}{2}\delta(w - 1) + \frac{1}{2}\delta(w + 1), \tag{133}$$

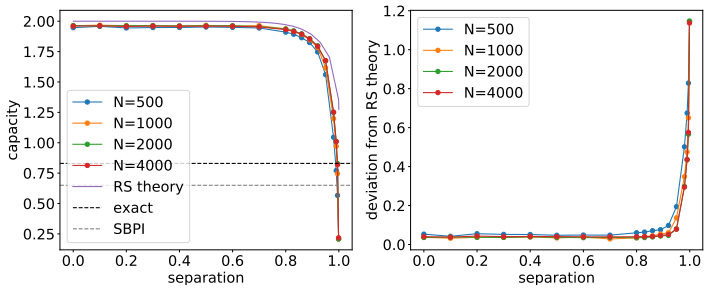

Figure 11: Left: Capacity as a function of separation for different size perceptrons. Dots are from DisCo-SGD simulations and the 'RS theory' line is from our theory. Exact values for Ising perceptron and state-of-the-art numerical values are included as well. Right: Deviation from the RS theory as a function of separation. This is the same as subtracting the simulation values from the theoretical predictions in the left figure.

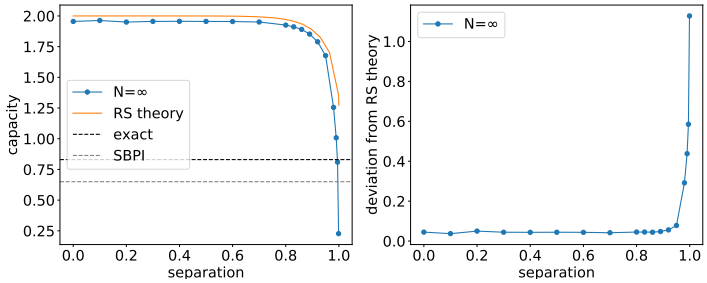

Figure 12: Finite size effects. Left/Right: we extrapolate simulation values in Fig.11 Left/Right to infinite N.

and CDF

$$Q(w) = \frac{1}{2}\Theta(w-1) + \frac{1}{2}\Theta(w+1). \tag{134}$$

Equating $Q(w)$ with the normal CDF $P(x)$ and solve for $w(x)$, we find $w(x) = \text{sgn}(x)$. Then $dw/dx = 2\delta(x)$ and $\left\langle \frac{dw}{dx} \right\rangle_x = \frac{2}{\sqrt{2\pi}}$. Therefore,

$$\lim_{Ising} \alpha_c(\kappa = 0) = \frac{4}{\pi}, \tag{135}$$

which is exactly the same as the prediction from the unconstrained RS theory [52, 13]. In contrast, the exact capacity of Ising perceptron with replica-symmetry breaking is $\alpha_c \approx 0.83$. For comparison, we have included these values in Fig.12(a), as well as the capacity found by the state-of-the-art supervised learning algorithm (Stochastic Belief Propagation, SBPI [10]) for Ising perceptron.

A.5.2 Sparse distributions

For a teacher with sparse distribution, $p(w_t) = (1-\rho)\delta(w_t) + \frac{\rho}{\sqrt{2\pi}\sigma_t w_t} \exp\left\{-\frac{(\ln w_t - \mu_t)^2}{2\sigma_t^2}\right\}$. We found that the simulations start to deviate from the theory, and the reason might be due to replica symmetry breaking. In Fig.13, we use the optimal prior learning paradigm similar to main text Fig.5. We see that our RS theory no longer gives accurate prediction of overlap in this case.

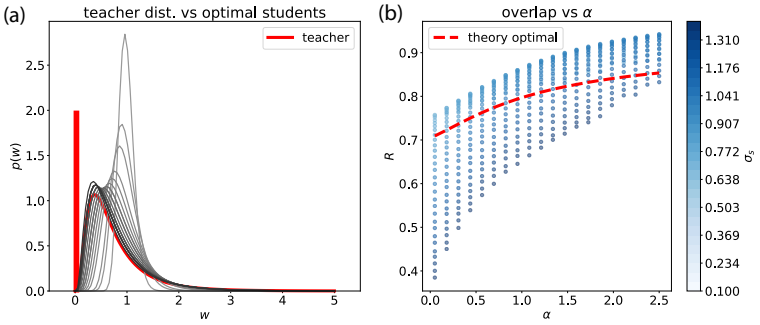

Figure 13: Optimal student prior distribution as a function of α. (a) Gray curves correspond to a series of optimal student distributions as a function of α, with the darker color representing larger α. Red is teacher distribution. (b) Overlap as a function of α for different student priors. Red dashed line is the optimal overlap calculated from our replica-symmetric theory. Dots are from DisCo-SGD simulations. For the same α, different color dots represent different overlaps obtained from simulations with different σ_s.