# A theory of weight distribution-constrained learning

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

A central question in computational neuroscience is how does structure determine function in neural networks. Recent large-scale connectomics 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 worms, flies, fish and mammals. 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 this 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 synaptic 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. Given a parameterized family of weight distributions, our theory predicts the shape of the distribution with optimal parameters. We apply our theory to 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. We further developed a statistical mechanical theory for teacher-student perceptron rule learning and ask of 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, and shows that distribution-constrained learning outperforms unconstrained and sign-constrained learning. Our theory and algorithm provide novel strategies of 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 be rooted 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 focused largely on the degree of connectivity [10, 25] and the constraints on the sign of the synapses (Excitatory vs. Inhibitory) [3, 9], but did not include additional features of synaptic weight distributions observed in the brain. As a result, computational theories of learning and memory often predict an unrealistic weight distribution (Fig.1). More generally, recent large-scale connectomics studies

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

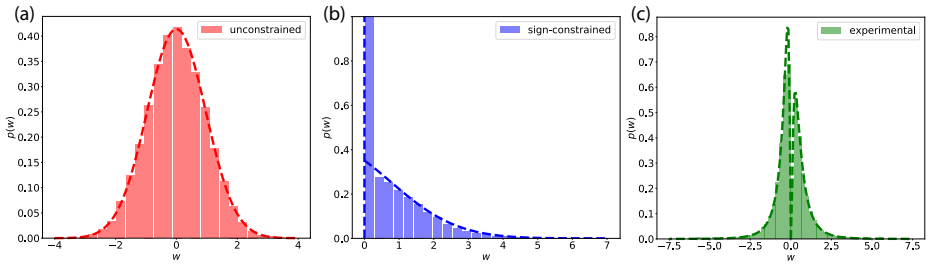


Figure 1: Theoretical and empirical synaptic weight distributions. (a)-(b) predicted distribution following perceptron learning at capacity. (a) Normal distribution when learning is unconstrained. (b) A delta-function plus a half-normal distribution when learning is sign-constrained. (c) Experimentally measured synaptic weight distribution (mouse primary auditory cortex [27]).

[26, 43, 45] 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 of these hitherto unavailable structural information.

The simplest model of neural learning is the perceptron [40]. Statistical mechanical analysis predicts that near capacity, an unconstrained perceptron classifying random input-output associations has normally distributed weights [21, 20, 14], see Fig.1(a). In contrast, physiological experiments suggest that biological synapses do not change their excitatory/inhibitory identity during learning (but see recent [24]). In order to take perceptron a step closer to biological realism, prior work has imposed sign constraints during learning [3, 9]. 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 connectomics studies ranging from cortical circuits in animals [27, 23, 33, 55, 45, 29, 7], to human cerebral cortex [33, 45] have shown evidence of log-normally distributed synaptic connections. As an example, Fig.1(c) shows the weight connection distribution in mouse primary auditory cortex (data adapted from [27]). The discrepancy between theory and current experimental result, together with the anticipated connectomics data due to recent breakthrough in connectomics technology, call for a theory of perceptron learning that incorporates the knowledge of global structural information. 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 for an arbitrary number of input synaptic subpopulations, each with its own prescribed distribution. Interestingly, the theory predicts that the reduction in capacity due to the constrained synaptic 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 satisfying 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 further 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 of 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.

## 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. The data consists of pairs$\{\boldsymbol{\xi}^\mu, \zeta^\mu\}_{\mu=1}^P$, where $\boldsymbol{\xi}^\mu$ is a $N$-dimentional random vector drawn i.i.d. from a standard normal distribution, $p(\xi_i^\mu) = \mathcal{N}(0,1)$, and $\zeta^\mu$ 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).

Extensions of this setup to include an arbitrary number of populations $\boldsymbol{w}^m$, each satisfying its own prescribed distribution constraints, $q_m(w^m)$ are 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 [21, 20], we consider the fraction $V$ of viable weights that satisfies both Eqn.1 and the distribution constraint $\hat{q} = q$, to all possible weights:

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

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 $\zeta^\mu$. This amounts to calculating $\langle \log V \rangle$, which can be done using the replica trick [21, 20].

We focus on solutions with maximum margin $\kappa$ at a given load $\alpha$, or equivalently, the maximum load capacity $\alpha_c(\kappa)$ of seperable patterns given margin $\kappa$. 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 [21, 20], 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 $0 \leq \left\langle \frac{dw}{dx} \right\rangle_x \leq 1$ 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 [30, 18, 13]. 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 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 $\delta$-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 paramerized 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 travelling 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 [46, 36]. 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 :

$$w_i^{t+\delta t} \leftarrow w_i^t - \delta t \sum_\mu \boldsymbol{\xi}^\mu (\hat{\zeta}^\mu - \zeta^\mu). \tag{7}$$

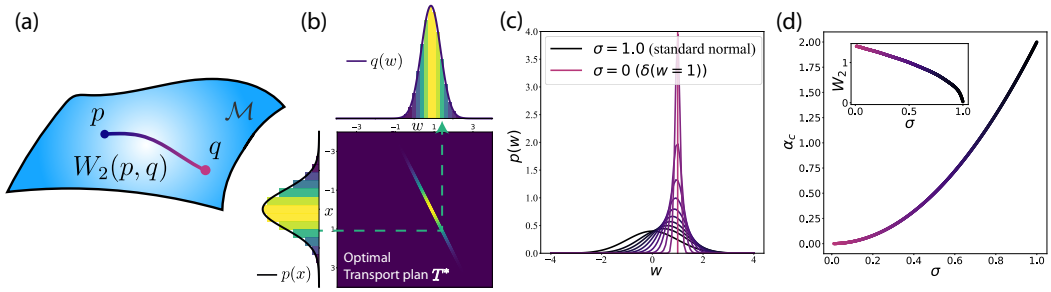


Figure 2: An illustration of optimal transport from a standard normal distribution $\mathcal{N}(0,1)$ to normal distributions with nonzero mean $\mathcal{N}(\sqrt{1-\sigma^2}, \sigma^2)$. (a) An schematic of the space $(\mathcal{M}, W_2)$ of probability distributions. (b) An example optimal transport 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 $\delta$-function at 1 ($\sigma = 1$). (d) Capacity $\alpha_c(\kappa = 0)$ as a function of $\sigma$. Inset shows the $W_2$ distance as a function of $\sigma$.

The theory of optimal transport provides a principled way of transport 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 $Q$ [47, 2]: $T^* = Q^{-1} \circ P$, 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 this Eqn.8 in action in Table 1(b).

However, performing such one-step projection strongly interferes with the cross-entropy objective, and numerically often results in solutions which 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 $\tau$ 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 $\delta t$ and $\delta\tau$ as learning rates $\eta_1$ and $\eta_2$. 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 $\beta$ 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 initial distribution $P$ and $\{\hat{w}_i\}$ drawn from final distribution $Q$, samples $\{w_i^{\tau}\}$ from intermediate distributions $P_\tau$ 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 weights $\boldsymbol{w}^t$ found by our algorithm, we can compute its empirical distribution $p_t$ and compare with theoretical distribution $p_\tau$ along the geodesic with the same $W_2$ distance to the target distribution (see Appendix A.2 for how to calculate $p_\tau$). In Fig.3(a), we show that indeed the empirical distributions $p_t$ agree with the geodesic distributions $p_\tau$ at geodesic time $\tau(t)$ (Fig.3(a)). The relation between the

(a)

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

Initialize patterns $\boldsymbol{\xi}^\mu$, labels $\zeta^\mu$ randomly, and initialize weights $w_i$ at zeros;
**while** $\zeta^\mu \neq sgn(\boldsymbol{w}^T\boldsymbol{\xi}^\mu)$ *for some* $\mu$ **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)

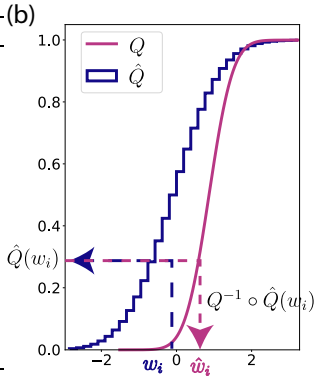


Table 1: Disco-SGD algorithm. (a) We perform alternating steps of gradient descent along the cross-entropy loss (Eqn.7), followed by steps along the optimal transport direction (Eqn.9). (b) An illustration of Eqn.8. For a given $w_i$, we first compute its empirical CDF value $\hat{Q}(w_i)$,then use the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

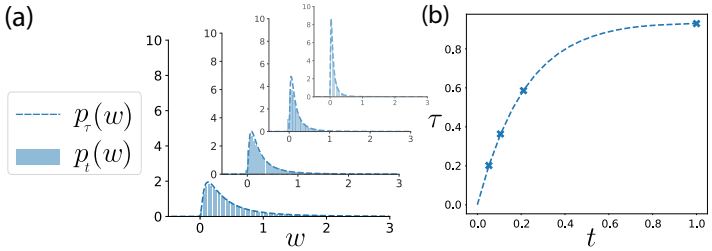


Figure 3: Intermediate distributions during learning are on the geodesic. (a) The solid histograms are the intermediate distribution $p_t$ at different training time $t$ from the DisCo-SGD algorithm, the dashed lines are geodesic distributions $p_\tau$ with the same $W_2$ distance to the target distribution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

geodesic time $\tau$ 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 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):

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

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 experi-

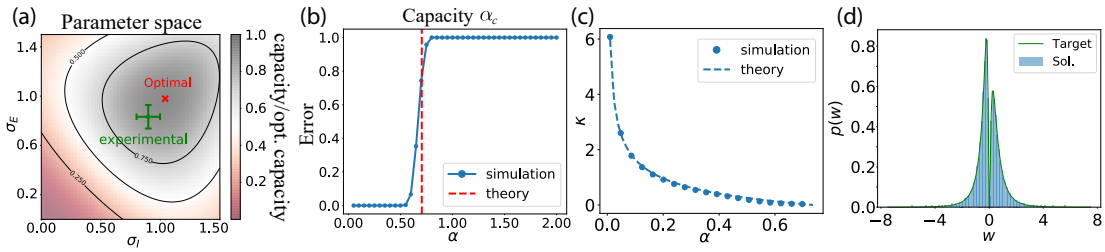


Figure 4: Biologically-realistic distribution and parameter landscape. (a) Capacity (normalized by the optimal value in the landscape) as a function of the lognormal parameters $\sigma_E$ and $\sigma_I$. Experimental value is shown in green with error bars, and optimal capacity is shown in red. (b)-(d)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

mentally 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 ($g_I = 1 - g_E$): $w_i^E \sim \frac{g_E}{\sqrt{2\pi}\sigma_E w^E} \exp\left\{-\frac{(\ln w^E - \mu_E)^2}{2\sigma_E^2}\right\}$ and $w_i^I \sim \frac{g_I}{\sqrt{2\pi}\sigma_I w^I} \exp\left\{-\frac{(\ln w^I - \mu_I)^2}{2\sigma_I^2}\right\}$.

We also demand that the mean synaptic weights satisfy the E/I balance condition [51, 52, 49, 50, 41, 34, 11] $g_E \left\langle w^E \right\rangle = g_I \left\langle w^I \right\rangle$ as is often observed in cortex connectomics experiments [4, 54, 37, 38, 6]. With the E/I balance condition and fixed second moment, the capacity is a function of the log-normal parameters $\sigma_E$ and $\sigma_I$. In Fig.4(a) we map out the 2d parameter space of $\sigma_E$ and $\sigma_I$ 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 connectomics studies in mouse primary auditory cortex [27].

In order to test our theory's validity on this estimated distribution of synaptic weights, we perform DisCo-SGD simulation with model parameters $\sigma_E$ and $\sigma_I$ fixed to their experimentally measured values. Both the capacity (Fig.4(b)), max-margin $\kappa$ 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 does 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, 8]. 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 [39, 28]. In the following, we try to answer this question in a simple setup where a student perceptron tries to learn from a teacher perceptron [44, 15]. 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. 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 $\eta^\mu$ 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

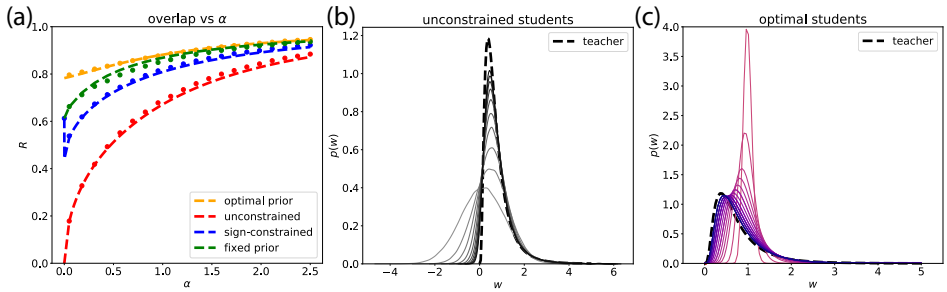


Figure 5: Compare different learning paradigms. (a) Teacher-student overlap $R$ as a function of load $\alpha$ in different learning paradigms. Dashed lines are from theory, and dots are from simulation. Note that there is an initial drop of the overlap in sign-constrained learning due to sparsification of weights. (b)-(c) The darker color curves correspond to larger $\alpha$, 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 $\delta$-function toward the teacher distribution.

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 $\varepsilon = 1/\pi \arccos R/\sqrt{1+\sigma^2}$ [44, 15].

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 $\alpha$ by solving the replica symmetric mean field self-consistency equations as in [44, 15]. 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 ($\sigma \neq 0$), $\alpha$ is bounded by $\alpha_c(\sigma)$ , since max-margin learning of separable data is assumed. The case with nonzero $\sigma$ is presented in Appendix A.3.4. In this unconstrained case, the student's weight distribution evolves from a gaussian for low $\alpha$ to one which increasingly resembles the teacher distribution for large $\alpha$ (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 [44, 15], 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.

### 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):

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

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 replica trick similar to [44, 15]. Overlap $R$ (Eqn.11) can be determined as a function of load $\alpha$ 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 $\alpha$. This presents a new learning paradigm we called 'optimal prior'. In Fig.5(a), we show that choosing optimal priors at different $\alpha$ achieves the overall best generalization performance compared with all other learning paradigms. For a given parameterizd family of distributions, our theory provides a way to analytically obtain the optimal prior $p_s^*$ as a function of $\alpha$ (Fig.5(c)). Note that unlike the unconstrained case (Fig.5(b)), the optimal prior starts from a $\delta$-function at 1 at zero $\alpha$, and asymptotically approaches the teacher distribution $p_t$ as $\alpha \to \infty$.

# 6 Summary and Discussion

We have developed a statistical mechanical framework of incorporating 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 [22, 48]. Under 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 to the imposed distribution. To numerically test our analytic theory, we use tools from optimal transport and information geometry to develop an SGD-based algorithm, DisCo-SGD, to realiably find weights that satisfy such prescribed constraints and correctly classify the data, and show 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 [5, 42] where Wasserstein distance is used as an objective for training generative models. We apply our theory to the biologically realistic case of of excitatory/inhibitory log-normal distributions that are observed in the cortex, and found experimentally measured parameters close to the optimal parameter values predicted by our theory. We further study input-output rule learning where the target rule is defined in terms of a weighted sum of the inputs, and ask to what extent prior knowledge of the target distribution may improve generalization performance. Using the teacher-student perceptron learning setup, we show 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.

# 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] 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

[3] 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

[4] 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

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

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

[7] 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

[8] 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

[9] 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

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

[11] 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. 4

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

[13] 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

[14] 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

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

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

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

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

[19] 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.

[20] 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

[21] 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

[22] 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

[23] 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

[24] 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

[25] 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

[26] 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

[27] 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

[28] 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

[29] 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

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

[31] 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.

[32] 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.

[33] 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

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

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

[36] 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

[37] 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

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

[39] 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

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

[41] 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. 4

[42] 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

[43] 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

[44] 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

[45] 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

[46] 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

[47] Matthew Thorpe. Introduction to optimal transport. *Lecture Notes*, 2019. 3

[48] 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

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

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

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

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

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

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

[55] 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

# 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

## A.1 Capacity supplemental materials

### A.1.1 Replica calculation of distribution-constrained capacity

### A.1.2 Theory for an arbitrary number of synaptic subpopulations

### A.1.3 Capacity for biased inputs and sparse label

## A.2 Optimal transport theory

## A.3 Generalization supplemental materials

### A.3.1 Replica calculation of generalization with sign-constraint

### A.3.2 Replica calculation of generalization with distribution-constraint

### A.3.3 Sparsification of weights in sign-constraint leanring

### A.3.4 Noisy teacher

## A.4 DisCo-SGD simulations

## A.5 Replica symmetry breaking

### A.5.1 Bimodal distributions

### A.5.2 Sparse distributions