Journal of Machine Learning Research 27 (2026) 1-68 Submitted 5/22; Revised 6/26; Published 8/26

The Role of Pseudo-Labels in Self-Training Linear Classifiers on High-Dimensional Gaussian Mixture Data

Takashi Takahashi TAKASHI-TAKAHASHI@G.ECC.U-TOKYO.AC.JP
Institute for Physics of Intelligence
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

Editor: Daniel Hsu

Abstract

Self-training (ST) is a simple yet effective semi-supervised learning method. However, why and how ST improves generalization performance by using potentially erroneous pseudo-labels is still not well understood. To deepen the understanding of ST, we derive and analyze a sharp characterization of the behavior of iterative ST when training a linear classifier by minimizing the ridge-regularized convex loss on binary Gaussian mixtures, in the asymptotic limit where input dimension and data size diverge proportionally. The results show that ST improves generalization in different ways depending on the number of iterations. When the number of iterations is small, ST improves generalization performance by fitting the model to relatively reliable pseudo-labels and updating the model parameters by a large amount at each iteration. This suggests that ST works intuitively. On the other hand, with many iterations, ST can gradually improve the direction of the classification plane by updating the model parameters incrementally, using soft labels and small regularization. It is argued that this is because the small update of ST can extract information from the data in an almost noiseless way. However, in the presence of label imbalance, the generalization performance of ST underperforms supervised learning with true labels. To overcome this, two heuristics are proposed to enable ST to achieve nearly compatible performance with supervised learning even with significant label imbalance.

Keywords: semi-supervised learning, self-training, pseudo-labels, statistical mechanics, replica method

1. Introduction

While supervised learning (SL) is effective when a large amount of labeled data is available, obtaining human-annotated labeled data can be expensive or even difficult in applications such as image segmentation (Zou et al., 2018) or text categorization (Nigam et al., 2000). On the other hand, obtaining unlabeled data is often inexpensive and easier. Therefore, semi-supervised learning (SSL) methods, which use a combination of labeled and unlabeled data, have been widely used in these fields to alleviate the need for labeled data (Chapelle et al., 2010).

Among many SSL methods, self-training (ST) is a simple and standard SSL algorithm, a wrapper algorithm that iteratively uses a supervised learning method (Scudder, 1965; McLachlan, 1975; Lee et al., 2013). The basic concept of ST is to use the model itself to make predictions on unlabeled data points, and then treat these predictions as labels for

subsequent training. Algorithmically, it starts by training a model on the labeled data. At each iteration, ST uses the current model to assign labels to unlabeled data points. These predicted labels can be soft (a continuous distribution) or hard (a one-hot distribution) (Xie et al., 2020). The model is then retrained using these newly labeled data. See Section 2 for algorithmic details. The model obtained in the last iteration is used for production. ST is a popular SSL method because of its simplicity and general applicability. It applies to any model that can make predictions for the unlabeled data points and can be trained by a supervised learning method. The predicted labels used in each iteration are also called *pseudo-labels* (Lee et al., 2013) because the predicted labels are only pseudo-correct compared to the ground-truth labels. Although ST is a simple heuristic, it has been empirically observed that ST can find a model with better predictive performance than the model trained on labeled data alone (Lee et al., 2013; Yalniz et al., 2019; Zhang et al., 2021; Xie et al., 2020; Rizve et al., 2021; Pham et al., 2021).

Despite its widespread acceptance and practical effectiveness, it is still not well understood why and how ST improves performance by fitting the model to potentially erroneous pseudo-labels. In particular, since ST is a wrapper algorithm, it can be combined with various heuristics, and its performance depends on such implementation details. For example, we can reject unlabeled data points from the training data if the assigned pseudo-labels are not reliable enough (Rizve et al., 2021). This procedure is known as *pseudo-label selection* (PLS). Similarly, we can use a soft label with a temperature parameter as a pseudo-label instead of a naive hard label, as in knowledge distillation (Hinton et al., 2015; Xie et al., 2020). Since the use of these heuristics is common, it is important to clarify how the performance of ST depends on such algorithmic details. In this study, we aim to improve the understanding of ST in this direction by sharply characterizing the asymptotic behavior of ST, and analyzing this sharp asymptotics.

In this work, we heuristically derive a sharp characterization of the behavior of ST in a simplified setup and use this characterization to analyze why and how ST improves the performance of classifiers. Specifically, our contributions can be summarized as follows:

- We heuristically derive a sharp characterization of the behavior of ST (Prediction 1-4) when training a linear classifier by minimizing the ridge-regularized convex loss for binary Gaussian mixtures, in the asymptotic limit where input dimension and data size diverge proportionally. There, the statistical properties of the weight vector and the logits are effectively described by a low-dimensional stochastic process whose parameters are determined by a set of equations termed *self-consistent equations* in Definition 1. The derivation is based on the evaluation of the *generating functionals* defined in (8) and (16) using the replica method of statistical mechanics (Mézard et al., 1987; Charbonneau et al., 2023; Montanari and Sen, 2024).

- Using the derived asymptotic formulas, we find that ST improves generalization performance through different mechanisms depending on the total number of iterations.

  - When the available number of iterations is small, ST can improve the generalization performance by fitting the model to relatively reliable pseudo-labels and updating the model parameters by a relatively large amount at each step. In this regime, using PLS and relatively hard labels are effective in improving the performance of the classifier (Sections 4.1 and 5.1). This is an intuitive behavior and

is consistent with existing experimental results with a small number of iteration steps (Rizve et al., 2021).

– When the number of iterations is large, ST finds a model with better generalization error by accumulating small parameter updates using small regularization parameter and moderately large batches of unlabeled data (Section 4.1 and Prediction 6 and Proposition 7). In this regime, the use of soft-labels is essential to keep the size of parameter updates small. It is argued that ST shows this behavior because the small update of ST can extract information from the data in an almost noiseless way. The derivation is based on the perturbative expansion of the solution of the self-consistent equations in the regularization parameter λ_U. Moreover, we obtain a closed-form solution for the evolution of cosine similarity between the weight vector and the cluster center when using a squared loss at the continuum limit $\lambda_U \to 0$ (Prediction 8), which verifies the above picture.

In any case, when the label imbalance is small, the performance of the model obtained by ST is close to that obtained by SL with ground truth labels. However, when the imbalance is large, the performance of naive ST is quite lower than that of SL, although the generalization performance is still improved by ST compared to the model obtained from the labeled data alone. This is because the ratio between the norm of the weight and the magnitude of the bias can become significantly large at large iteration steps.

- To overcome the problems in label imbalanced cases, we introduce two heuristics; (i) *pseudo-label annealing* (Heuristic 1), which gradually changes the pseudo-labels from soft labels to hard labels as the iteration proceeds, and *bias-fixing* (Heuristic 2), which fixes the bias term to that of the initial classifier. By numerically analyzing the asymptotic formula, we demonstrate that with these two heuristics, ST can find a classifier whose performance is nearly compatible with supervised learning using true labels even in the presence of significant label imbalance.

The remainder of the paper is organized as follows. Section 2 states the problem setup treated in this study; the assumptions on the data generation process and the algorithmic details of ST are described. Section 3 introduces the analytical framework to characterize the precise asymptotics of ST and apply it to the setup described in Section 2. We predict how the weights and the logits are statistically characterized through a small finite number of variables determined by the deterministic self-consistent equations. The comparison between our predictions and the numerical experiments is also presented here. The step-by-step derivation of the predictions is presented in Appendix A. Then, by numerically and analytically investigating the self-consistent equations, we investigate how the generalization error depends on the details of the problem setup in Section 4. Based on the findings in this section, we propose heuristics for label imbalanced cases and show their effectiveness by numerically solving the self-consistent equations in Section 5. Finally, Section 6 concludes the paper with some discussions.

1.1 Related Works

ST is a method of SSL that has a very long history (Scudder, 1965; McLachlan, 1975). Although it is relatively recent that it has been used in the context of deep learning, along

with the term pseudo-label (Lee et al., 2013), nowadays, ST is used as an important building block, especially in applications of computer vision tasks (Lee et al., 2013; Yalniz et al., 2019; Zhang et al., 2021; Xie et al., 2020; Rizve et al., 2021; Pham et al., 2021).

On the other hand, the understanding of the mechanism of ST is still very limited compared to supervised learning. Although there have been several theoretical studies in the last few years, as shown below. See also (Amini et al., 2022) for a review of recent developments.

Among the recent theoretical studies, the most closely related ones are (Oymak and Gulcu, 2020, 2021; Frei et al., 2022) which consider training linear models using ST. (Oymak and Gulcu, 2020, 2021) consider the classification of a two-component Gaussian mixture model (GMM) using the averaging estimator, which yields a Bayes-optimal classifier in a supervised setup (Dobriban and Wager, 2018; Mignacco et al., 2020a). This study sharply characterizes the behavior of this estimator in a high-dimensional setting and shows that the estimator obtained by ST is correlated with the Bayes-optimal classifier, but it is limited to the analysis of the averaging estimator and the class-balanced case. Similarly, the literature (Frei et al., 2022) considers the classification of a mixture of rotationally symmetric distributions. It studies learning linear models with ST based on the optimization of the cross-entropy or the exponential loss after supervised learning with a small labeled dataset. It is shown that ST can find a Bayes optimal classifier up to an ϵ error if $\mathcal{O}(d/\epsilon^2)$ unlabeled data points are available in ST with d the number of the input dimension. In contrast to our study, it is limited to the setup where the class labels are balanced and their result does not include the sharp characterization of the statistical behavior of the regressors.

In a similar line, (Zhang et al., 2022) considers the ST of a single hidden-layer fully connected neural network with fixed top-layer weights in a regression setup when the features are generated from a single zero-mean Gaussian distribution and the labels are generated from a realizable teacher without noise. This study shows that, under some assumptions about the initial condition and the size of the unlabeled data, ST can find the ground truth classifier with less sample complexity than without unlabeled data. Although it treats the training of a two-layer neural network, the feature model is restricted to a single Gaussian setup (not a classification on mixture models).

ST has also been studied in the context of domain adaptation. Domain adaptation aims at transferring knowledge from one source domain to a different target domain, without using labeled data from the target domain. The literature (Kumar et al., 2020) treats the gradual domain adaptation based on ST, where the classifier is a linear model. It is shown that the Bayes optimal classifier is obtained by ST even after the domain shift. Unlike our work, it assumes (i) an access to infinitely large unlabeled data at each iteration, and (ii) the initial classifier is close to Bayes optimal. Similarly, the literature (Chen et al., 2020) studies the ST of a linear model under the setting that the target domain data contains spurious features that are irrelevant to the ground truth labels. They show that ST converges to a solution that has zero regression coefficients on the spurious features. Their ST updates pseudo-labels after each SGD step, while in our work pseudo-label is fixed until the student model completely minimizes the loss. As also pointed out by (Chen et al., 2020) (see Appendix E of that literature), in practice, the student model is often trained to converge between pseudo-label updates. Therefore, deepening the understanding in our setting is also an important avenue. Finally, (Wei et al., 2021; Cai et al., 2021) analyze the learning of deep neural networks

using consistency regularization, and derive finite sample bounds on the generalization error. Although the data and classifier models are rather general, they consider a single shot learning rather than the iterative ST, and the consistency regularization loss is different from the loss function used in the conventional ST (Lee et al., 2013).

From a technical viewpoint, our work is a replica analysis of the sharp asymptotics of ST in which the input dimension and the size of the dataset diverge at the same rate. The salient feature of this proportional asymptotic regime is that the macroscopic properties of the learning results, such as the predictive distribution and the generalization error, do not depend on the details of the realization of the training data, when using convex loss. That is, the fluctuations of these quantities with respect to the training data vanish, allowing us to make sharp theoretical predictions. Analyzing such a sharp asymptotics is a topic with a long history. Around the 1990s, sharp asymptotics of many machine learning methods were studied using the replica method. Examples of the analyzed methods include linear regression (Krogh and Hertz, 1991), active learning (Seung et al., 1992), support vector machines (Dietrich et al., 1999), and two-layer neural networks (Schwarze, 1993), to name a few[1]. Although statistical mechanics techniques often have procedures whose validity is not rigorously proved yet, most of the predictions agree exactly with the experiments, and hence it is believed that their predictive power itself is credible (Talagrand, 2010; Charbonneau et al., 2023; Montanari and Sen, 2024) if it is properly used. Indeed, the replica method is used to derive various results in modern machine learning research (Gerace et al., 2020; Mignacco et al., 2020a; Canatar et al., 2021a; Zavatone-Veth and Pehlevan, 2023; Canatar et al., 2021c; D'Ascoli et al., 2020; Loureiro et al., 2022; Sorscher et al., 2022; Loureiro et al., 2021; Karakida and Akaho, 2022; Tomasini et al., 2022; Pezeshki et al., 2022; Pourkamali and Macris, 2023; Okajima et al., 2023; Ichikawa and Hukushima, 2024) as a simplified method for predicting accurate results. Also, some of the predictions are later justified by rigorous methods (Talagrand, 2010; Barbier et al., 2016; Barbier and Macris, 2019; Barbier et al., 2019). Although rigorous methods are rapidly being developed and applied to advanced problems including an iterative online optimization algorithm (Chandrasekher et al., 2023), it is still known that the replica method often yields correct predictions with less effort than rigorous approaches (Montanari and Sen, 2024). Therefore, we believe that extending the replica analysis itself is of important interest.

There have been several works investigating iterative procedures using the replica method. The idea of analyzing the time-evolving systems using the replica method has been proposed in analyzing the physics of glassy systems (Krzakala and Kurchan, 2007; Franz and Parisi, 2013). Formally, applying this technique for only one step of the iteration is called the method of *Franz-Parisi potential* (Franz and Parisi, 1997, 1998; Parisi et al., 2020; Bandeira et al., 2022), and has been used in the context of machine learning, such as knowledge distillation (Saglietti and Zdeborova, 2022), adaptive sparse estimation (Obuchi and Kabashima, 2016), and loss-landscape analysis (Huang and Kabashima, 2014; Baldassi et al., 2016). However, it has not been used in analyzing multiple iteration procedures in machine learning except (Okajima and Takahashi, 2025; Takanami et al., 2025). Our work can be regarded as an extension of these analyses to the iterative ST.

1. See also (Opper and Kinzel, 1996; Opper, 2001; Engel and Van den Broeck, 2001) for a review of early statistical mechanics studies and their relationship with the standard statistical learning theory.

Notation	Description
μ, ν	sample indices of labeled and unlabeled data points
i, j	indices of the weight vector $\boldsymbol{w}$
$[n]$	for a positive integer n, the set $\{1, \dots, n\}$
T	total number of iteration used in ST
$t \in [T]$	index representing an iteration step in ST with a maximum number of iterations T
$\lvert\cdot\rvert$	if applied to a number, absolute value
$(\,\cdot\,)^\top$	vector/matrix transpose
$\boldsymbol{x} \cdot \boldsymbol{y}$	for vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^N$, the inner product of them: $\boldsymbol{x} \cdot \boldsymbol{v} = \sum_{i=1}^N x_i y_i$.
v_i	i-th element of a vector $\boldsymbol{v}$
$\lVert\boldsymbol{v}\rVert_p$	for a vector $\boldsymbol{v} = [v_i]_{1 \le i \le N}$, ℓ_p norm of the vector defined as $\left(\sum_{i=1}^N \lvert v_i\rvert^p\right)^{1/p}$
$\mathbf{1}_N$	an N-dimensional vector $(1, 1, \dots, 1) \in \mathbb{R}^N$
I_N	identity matrix of size $N \times N$
$\mathbb{1}(\cdot)$	indicator function
$\mathcal{N}(\mu, \sigma^2)$	Gaussian density with mean μ and variance σ^2
$D\xi$	standard Gaussian measure $e^{-\xi^2/2}/\sqrt{2\pi}d\xi$
$d\boldsymbol{x}$	with $\boldsymbol{x} \in \mathbb{R}^N$, a measure over $\mathbb{R}^N$
$d^n\boldsymbol{x}$	with $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in \mathbb{R}^N$, a measure over $\mathbb{R}^{N \times n}$
$\mathbb{E}_{X \sim p_X}[f(X)]$	Expectation regarding random variable X where p is the density function for the random variable X (lower subscript $X \sim p_X$ can be omitted if there is no risk of confusion)
$\mathbb{V}\mathrm{ar}_{X \sim p_X}[f(X)]$	Variance: $\mathbb{E}[f(X)^2] - \mathbb{E}[f(X)]^2$
$\{x_i\} \overset{\mathrm{d}}{=} X$	empirical distribution of $\{x_i\}$ is equal to the distribution of the r.v. X
$\delta_{\mathrm{d}}(\cdot)$	Dirac's delta function
$\delta_{a,b}$	for integers a, b, the Kronecker's delta: $\delta_{a,b} = \mathbb{1}(a = b)$
$\mathrm{extr}_x f(x)$	extremization with respect to x
$\partial_i \mathcal{F}$	for a bivariate function $\mathcal{F}(y, x)$, the partial derivative of $\mathcal{F}$ with respect to the i-th argument For example, $\partial_1 \mathcal{F}(Y, X) = \left.\frac{\partial \mathcal{F}}{\partial y}\right\rvert_{y=Y, x=X}$.
$f(n) = \mathcal{O}(g(n))$	Landau's O as $n \to 0$; $\lvert f(n)/g(n)\rvert < \infty$ as $n \to 0$

Table 1: Notations

1.2 Notations

Throughout the paper, we use some shorthand notations for convenience. We summarize them in Table 1.

2. Problem Setup

This section presents the problem setup and our interest in this work. The assumptions on the data generation process are first described and then the iterative ST procedure is formalized.

Let $D_L = \{(\boldsymbol{x}_\mu^{(0)}, y_\mu^{(0)})\}_{\mu=1}^{M_L}, \boldsymbol{x}_\mu^{(0)} \in \mathbb{R}^N, y_\mu^{(0)} \in \{0,1\}$ be the set of independent and identically distributed (iid) labeled data points, and let $D_U^{(t)} = \{\boldsymbol{x}_\nu^{(t)}\}_{\nu=1}^{M_U}, \boldsymbol{x}_\nu^{(t)} \in \mathbb{R}^N, t = 1, 2, \ldots, T$ be the sets of iid unlabeled data points; there are T batches of the unlabeled datasets of size M_U, thus, in total, there are TM_U unlabeled data points. This study assumes that the data points are generated from binary Gaussian mixtures whose centroids are located at $\pm\boldsymbol{v}/\sqrt{N}$ with $\boldsymbol{v} \in \mathbb{R}^N$ as a fixed vector. The covariance matrices for these two Gaussian distributions are assumed to be spherical. From the rotational symmetry of these Gaussian distributions, we can fix the direction of the vector $\boldsymbol{v}$ as $\boldsymbol{v} = (1, 1, \ldots, 1)\cdot$ without loss of generality. Furthermore, we assume that each Gaussian contains a fraction $\rho^{(t)}$ and $(1-\rho^{(t)})$ of the points with $\rho^{(t)} = \rho_L \in (0, 0.5]$ if $t = 0$ and $\rho^{(t)} = \rho_U \in (0, 0.5]$, otherwise. In this setup, the feature vectors $\boldsymbol{x}_\mu^{(0)}$ and $\boldsymbol{x}_\nu^{(t)}$ can be written as

$$\boldsymbol{x}_\mu^{(0)} = (2y_\mu^{(0)} - 1)\frac{1}{\sqrt{N}}\boldsymbol{v} + \boldsymbol{z}_\mu^{(0)}, \quad \mu = 1, 2, \ldots, M_L, \tag{1}$$

$$\boldsymbol{x}_\nu^{(t)} = (2y_\nu^{(t)} - 1)\frac{1}{\sqrt{N}}\boldsymbol{v} + \boldsymbol{z}_\nu^{(t)}, \quad \nu = 1, 2, \ldots, M_U,\ t = 1, 2, \ldots, T, \tag{2}$$

where $\boldsymbol{z}_\mu^{(0)} \sim_{\text{iid}} \mathcal{N}(0, \Delta_L I_N), \boldsymbol{z}_\nu^{(t)} \sim_{\text{iid}} \mathcal{N}(0, \Delta_U I_N), \Delta_L, \Delta_U > 0$ are the independent Gaussian noise and $y_\mu^{(0)} \sim_{\text{iid}} p_y^{(0)}, y_\nu^{(t)} \sim_{\text{iid}} p_y^{(t)}$ are the ground truth label where $p_y^{(t)}$ is defined as

$$p_y^{(t)}(y) = \begin{cases} p_{y,L}(y) \equiv \rho_L \delta_{\text{d}}(y-1) + (1-\rho_L)\delta_{\text{d}}(y), & t = 0 \\ p_{y,U}(y) \equiv \rho_U \delta_{\text{d}}(y-1) + (1-\rho_U)\delta_{\text{d}}(y), & \text{otherwise} \end{cases}$$

The goal of ST is to obtain a classifier with a better generalization ability from

$$D = D_L \cup D_U^{(1)} \cup \ldots D_U^{(T)},$$

than the model trained with D_L only.

We focus on the ST with the linear model, i.e., the model's output $f(\boldsymbol{x})$ at an input $\boldsymbol{x}$ is a function of a linear combination of the weight $\boldsymbol{w}$ and the bias B:

$$f_{\text{model}}(\boldsymbol{x}) = \sigma\left(\frac{1}{\sqrt{N}}\boldsymbol{x}\cdot\boldsymbol{w} + B\right),$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is a link function that maps a logit to the model output; some examples are shown in Table 2. The factor $1/\sqrt{N}$ is introduced to ensure that the logit $\boldsymbol{x}\cdot\boldsymbol{w}/\sqrt{N} + B$ should have moderate magnitude at $N \gg 1$.

In the following, we denote $\boldsymbol{\theta} = (\boldsymbol{w}, B)$ for the shorthand notation of the linear model's parameter. Similarly, a pseudo-label $\hat{y}(\boldsymbol{x})$ at an input $\boldsymbol{x}$ is also given as a linear combination of the model's parameter:

$$\hat{y}(\boldsymbol{x}) = \sigma_{\text{pl}}\left(\frac{1}{\sqrt{N}}\boldsymbol{x}\cdot\boldsymbol{w} + B\right),$$

where $\sigma_{\rm pl}:\mathbb{R}\to\mathbb{R}$ is the pseudo-labeling function that maps the logit of the current model to the pseudo-label used at the next ST step. Depending on the choice of $\sigma_{\rm pl}$, the resulting pseudo-label may retain a real-valued confidence score (soft-label) or take a discrete value (hard-label). Examples are given in Table 2. In each iteration, the model is trained by minimizing the ridge-regularized convex loss functions l for the supervised learning and $l_{\rm pl}$ for the ST steps, where $l(y,f)$ denotes the loss between a ground-truth label y and a model output f, and at the subsequent ST steps, $l_{\rm pl}(\hat{y},f)$ denotes the loss between a pseudo-label $\hat{y}$ and a model output f. Our asymptotic characterization is formulated for general choices of σ, $\sigma_{\rm pl}$, l, and $l_{\rm pl}$ for which the objective function at each step is convex with respect to $\boldsymbol{\theta}^{(t)}$. Examples of these functions are given in Table 2. Additional regularity assumptions needed for particular results are stated where they are used.

Furthermore, we accept removing unlabeled data points from the training data if the assigned pseudo-labels are not reliable enough. Here we simply define the reliability based on the magnitude of the logit. Then, given a threshold $\Gamma\geq 0$ for PLS, and regularization strength $\lambda^{(t)}\geq 0$, the ST algorithm is formalized as follows:

- *Step 0: Initializing the model with the labeled data D_L.* Initialize the iteration number $t=0$ and obtain a model $\hat{\boldsymbol{\theta}}^{(0)}=(\hat{\boldsymbol{w}}^{(0)},\hat{B}^{(0)})$ by minimizing the following loss:

$$\mathcal{L}^{(0)}(\boldsymbol{\theta}^{(0)};D_L)=\sum_{\mu=1}^{M_L} l\left(y_\mu,\sigma\left(\frac{1}{\sqrt{N}}\boldsymbol{x}_\mu^{(0)}\cdot\boldsymbol{w}^{(0)}+B^{(0)}\right)\right)+\frac{\lambda^{(0)}}{2}\|\boldsymbol{w}^{(0)}\|_2^2, \tag{3}$$

  with respect to $\boldsymbol{\theta}^{(0)}$. Let $t\leftarrow 1$, and proceed to Step 1.

- *Step1: Creating pseudo-labels.* Give the pseudo-labels for unlabeled data points in $D_U^{(t)}$ so that the pseudo-label for the unlabeled data point $\boldsymbol{x}_\mu^{(t)}$ is

$$\hat{y}_\nu^{(t)}=\sigma_{\rm pl}\left(\frac{1}{\sqrt{N}}\boldsymbol{x}_\nu^{(t)}\cdot\hat{\boldsymbol{w}}^{(t-1)}+\hat{B}^{(t-1)}\right),\quad \nu=1,2,\dots,M_U.$$

- *Step2: Updating the model.* Obtain the model $\hat{\boldsymbol{\theta}}^{(t)}=(\hat{\boldsymbol{w}}^{(t)},\hat{B}^{(t)})$ by minimizing the following loss:

$$\begin{aligned}\mathcal{L}^{(t)}(\boldsymbol{\theta}^{(t)};D_U^{(t)},\hat{\boldsymbol{\theta}}^{(t-1)})=\sum_{\nu=1}^{M_U}\mathbb{1}\left(\left|\frac{1}{\sqrt{N}}\boldsymbol{x}_\nu^{(t)}\cdot\hat{\boldsymbol{w}}^{(t-1)}+\hat{B}^{(t-1)}\right|>\Gamma\sqrt{\bar{q}^{(t-1)}}\right)\\ \times l_{\rm pl}\left(\hat{y}_\nu^{(t)},\sigma\left(\frac{1}{\sqrt{N}}\boldsymbol{x}_\nu^{(t)}\cdot\boldsymbol{w}^{(t)}+B^{(t)}\right)\right)+\frac{\lambda^{(t)}}{2}\|\boldsymbol{w}^{(t)}\|_2^2,\end{aligned} \tag{4}$$

  with respect to $\boldsymbol{\theta}^{(t)}$, where $\bar{q}^{(t-1)}=\frac{1}{N}\|\hat{\boldsymbol{w}}^{(t-1)}\|_2^2$ is the normalized squared norm of the weight vector at the previous step. If $t<T$, let $t\leftarrow t+1$ and go back to Step 1.

In the following, we will use T as the total number of iterations in ST, and t as the iteration index of ST when the total number of T is fixed.

We assume positive regularization in the ST steps to avoid a trivial solution $\hat{\theta}^{(t)}=\hat{\theta}^{(t-1)}$ for $t=1,2,\dots$.To see this, consider $\lambda^{(1)}=\cdots=\lambda^{(T)}=0$ and a soft pseudo-label setting

Table 2: Examples of the combinations of $\sigma, \sigma_{\rm pl}, l$, and $l_{\rm pl}$.

$\sigma(x)$	$\sigma_{\rm pl}(x)$	$l(p,q)$	$l_{\rm pl}(p,q)$
x	x	$\frac{1}{2}(p-q)^2$	$\frac{1}{2}(p-q)^2$
$1/(1+e^{-x})$	$1/(1+e^{-\gamma x}), \gamma > 0$	$-p\log q - (1-p)\log(1-q)$	$-p\log q - (1-p)\log(1-q)$

in which the pseudo-label is generated by the same output map as the model, and $l_{\rm pl}$ is minimized when its two arguments are equal. If we set $\boldsymbol{\theta}^{(t)} = \hat{\boldsymbol{\theta}}^{(t-1)}$, then the linear model gives the same logit on every unlabeled example as the previous model did. Hence the model output exactly reproduces the pseudo-label, and every pseudo-label loss term is minimized. Therefore, $\hat{\boldsymbol{\theta}}^{(t-1)}$ is a minimizer of the unregularized ST objective at step t. If this minimizer is unique, the update gives $\hat{\boldsymbol{\theta}}^{(t)} = \hat{\boldsymbol{\theta}}^{(t-1)}$, and by iteration $\hat{\boldsymbol{\theta}}^{(0)} = \hat{\boldsymbol{\theta}}^{(1)} = \cdots = \hat{\boldsymbol{\theta}}^{(T)}$. If the minimizer is not unique, the unregularized update itself is ambiguous. We therefore restrict our analysis to $\lambda^{(t)} > 0$ for $t = 1, \ldots, T$.

The iterative ST procedure above is simplified in two ways. First, the labeled data points are not used in steps 1 and 2. Second, the algorithm does not use the same unlabeled data points in each iteration. These two aspects simplify the following analysis, which allows us to obtain a concise analytical result. Although the above ST has these differences from the commonly used procedures, we can still obtain non-trivial results as in the literature (Oymak and Gulcu, 2020, 2021).

2.1 Generalization Error

Aside from the sharp characterization of the statistical properties of the estimators $\hat{\boldsymbol{\theta}}^{(t)}$, we are interested in the generalization performance of the ST algorithm that is evaluated by the generalization error defined as

$$\bar{\epsilon}_{\rm g}^{(t)} = \mathbb{E}_{(\boldsymbol{x},y)}\left[\mathbb{1}\left[y \neq \hat{y}_{\rm pred}(\boldsymbol{x}; \hat{\boldsymbol{\theta}}^{(t)})\right]\right], \tag{5}$$

$$\hat{y}_{\rm pred}(\boldsymbol{x}; \hat{\boldsymbol{\theta}}^{(t)}) = \mathbb{1}\left[\frac{1}{\sqrt{N}}\hat{\boldsymbol{w}}^{(t)} \cdot \boldsymbol{x} + \hat{B}^{(t)} > 0\right],$$

where $(\boldsymbol{x}, y)$ follows the same data generation process with the labeled data points:

$$\boldsymbol{x} = (2y-1)\frac{1}{\sqrt{N}}\boldsymbol{v} + \boldsymbol{z},$$
$$y \sim p_{y,L}, \quad \boldsymbol{z} \sim \mathcal{N}(0, \Delta_L I_N),$$

For evaluating this, we consider the large system limit where $N, M_L, M_U \to \infty$, keeping their ratios as $(M_L/N, M_U/N) = (\alpha_L, \alpha_U) \in (0,\infty) \times (0,\infty)$. In this asymptotic limit, ST's behavior can be sharply characterized by the replica analysis presented in the next section. We term this asymptotic limit as *large system limit* (LSL). Hereafter, $N \to \infty$ represents LSL as a shorthand notation to avoid cumbersome notation.

As reported in (Mignacco et al., 2020a), when the feature vectors are generated according the spherical Gaussians as assumed above, the generalization error (5) can be described by

the bias $\hat{B}^{(t)}$ and two macroscopic quantities that characterize the geometrical relations between the estimator and the centroid of the Gaussians $\boldsymbol{v}$:

$$\bar{\epsilon}_{\mathrm{g}}^{(t)} = \rho_U H\left(\frac{\bar{m}^{(t)} + \hat{B}^{(t)}}{\sqrt{\Delta_U \bar{q}^{(t)}}}\right) + (1-\rho_U) H\left(\frac{\bar{m}^{(t)} - \hat{B}^{(t)}}{\sqrt{\Delta_U \bar{q}^{(t)}}}\right), \tag{6}$$

$$H(x) = \int_x^\infty D\xi, \qquad D\xi \equiv \frac{e^{-\xi^2/2}}{\sqrt{2\pi}}\, d\xi,$$

$$\bar{q}^{(t)} = \frac{1}{N}\|\hat{\boldsymbol{w}}^{(t)}\|_2^2, \quad \bar{m}^{(t)} = \frac{1}{N}\hat{\boldsymbol{w}}^{(t)} \cdot \boldsymbol{v}. \tag{7}$$

Hence, evaluation of $\hat{B}^{(t)}, \bar{q}^{(t)}$ and $\bar{m}^{(t)}$ is crucial in our analysis.

The main question we want to investigate is how the statistical property of the estimators $\hat{\boldsymbol{\theta}}^{(t)}$ and the generalization error depends on the regularization parameters $\lambda^{(0)}, \lambda^{(1)}, \ldots, \lambda^{(T)}$, the number of iterations T, and the properties of the data, such as the degree of the label imbalance ρ_L, ρ_U.

3. Sharp Asymptotics of ST

The first major technical contribution of this work is the development of a theoretical framework for sharply characterizing the behavior of ST using the replica method. We first rewrite ST as a limit of probabilistic inference in a statistical mechanics formulation in subsection 3.1. Then, we propose the theoretical framework for analyzing the ST by using the replica method in subsection 3.2, and show the first main analytical result for characterizing the behavior of ST in LSL in subsection 3.3. As usual in the replica method, it contains the non-rigorous steps due to the continuation to zero replica numbers. See Section 3.2.1 for details. Some comparisons with the numerical experiments are presented in subsection 3.4. Detailed calculations are presented in Appendix A.

3.1 Statistical Mechanics Formulation of ST

3.1.1 Chain of Boltzmann Distributions

Let us start with rewriting the ST as a statistical mechanics problem. We introduce probability densities $p^{(0)}, p^{(t)}, t = 1, 2, \ldots, T$, which are termed the *Boltzmann distributions* following the custom of statistical mechanics, as follows. For $\beta^{(t)} > 0$ and $t = 0, 1, \ldots, T$, the Boltzmann distributions are defined as

$$p^{(0)}(\boldsymbol{\theta}^{(0)}|D_L) = \frac{1}{Z^{(0)}} e^{-\beta^{(0)}\mathcal{L}^{(0)}(\boldsymbol{\theta}^{(0)};D_L)},$$

$$p^{(t)}(\boldsymbol{\theta}^{(t)}|D_U^{(t)}, \boldsymbol{\theta}^{(t-1)}) = \frac{1}{Z^{(t)}} e^{-\beta^{(t)}\mathcal{L}^{(t)}(\boldsymbol{\theta}^{(t)};D_U^{(t)},\boldsymbol{\theta}^{(t-1)})}, \quad t = 1, 2, \ldots, T,$$

where $Z^{(0)}$ and $Z^{(t)}, t = 1, 2, \ldots, T$ are the normalization constants:

$$Z^{(0)}(D_L; \beta^{(0)}) = \int e^{-\beta^{(0)}\mathcal{L}^{(0)}(\boldsymbol{\theta}^{(0)};D_L)} d\boldsymbol{\theta}^{(0)},$$

$$Z^{(t)}(D_U^{(t)}, \boldsymbol{\theta}^{(t-1)}; \beta^{(t)}) = \int e^{-\beta^{(t)}\mathcal{L}^{(t)}(\boldsymbol{\theta}^{(t)};D_U^{(t)},\boldsymbol{\theta}^{(t-1)})} d\boldsymbol{\theta}^{(t)}, \quad t = 1, 2, \ldots, T.$$

Recall that $\mathcal{L}^{(0)}$ and $\mathcal{L}^{(t)}, t = 1, 2, \ldots, T$ are the loss functions used in each step of ST. This chain of the Boltzmann distributions defines a Markov process over $\{\boldsymbol{\theta}^{(t)}\}_{t=0}^{T}$ conditioned by the data D:

$$p_{\mathrm{ST}}(\{\boldsymbol{\theta}^{(t)}\}_{t=0}^{T}|D) \equiv p^{(0)}(\boldsymbol{\theta}^{(0)}|D_L)\prod_{t=1}^{T} p^{(t)}(\boldsymbol{\theta}^{(t)}|D_U^{(t)}, \boldsymbol{\theta}^{(t-1)}).$$

By successively taking the limit $\beta^{(0)} \to \infty, \beta^{(1)} \to \infty, \ldots$, the Boltzmann distributions $p^{(0)}, p^{(1)}, \ldots$, converge to the Delta functions at $\hat{\boldsymbol{\theta}}^{(0)}, \hat{\boldsymbol{\theta}}^{(1)}, \ldots$, respectively. Thus analyzing ST is equivalent to analyzing the Boltzmann distributions at this limit[2]. Hereafter the limit without the upper subscript $\beta \to \infty$ represents taking all of the successive limits $\beta^{(t)} \to \infty, t = 0, 1, \ldots, T$ as a shorthand notation. Furthermore, we will omit the arguments $D_L, D_U^{(t)}, \boldsymbol{\theta}^{(t)}, \beta^{(t)}$ when there is no risk of confusion to avoid cumbersome notation.

3.1.2 Generating Functional

For evaluating the behavior of the weights $\hat{w}_i^{(t)}$ and the biases $\hat{B}^{(t)}$, we introduce the *generating functional*. Let $g_w : \mathbb{R}^{T+1} \to \mathbb{R}$ and $g_B : \mathbb{R}^{T+1} \to \mathbb{R}$ be arbitrary test functions of the trajectories of a weight coordinate and the bias, respectively, such that the expectations and derivatives below are well defined. We also introduce auxiliary real-valued source parameters $\epsilon_w, \epsilon_B \in \mathbb{R}$, which are set to zero after differentiation and are used to extract the desired expectations. Then, the generating functional defined as [3]

$$\Xi_{\mathrm{ST}}(\epsilon_w, \epsilon_B) = \lim_{N,\beta\to\infty} \mathbb{E}_{\{\boldsymbol{\theta}^{(t)}\}_{t=0}^{T} \sim p_{\mathrm{ST}}, D}\left[e^{\epsilon_w g_w(\{w_i^{(t)}\}_{t=0}^{T}) + \epsilon_B g_B(\{B^{(t)}\}_{t=0}^{T})}\right]. \tag{8}$$

More specifically, the average quantities $\mathbb{E}_D[g_w(\{\hat{w}_i^{(t)}\})]$ and $\mathbb{E}_D[g_B(\{\hat{B}^{(t)}\})]$ can be evaluated by using the following formulae:

$$\mathbb{E}_D\left[g_w\left(\{\hat{w}_i^{(t)}\}_{t=0}^{T}\right)\right] = \lim_{\epsilon_w,\epsilon_B\to 0} \frac{\partial}{\partial \epsilon_w}\Xi_{\mathrm{ST}}(\epsilon_w, \epsilon_B), \tag{9}$$

$$\mathbb{E}_D\left[g_B\left(\{\hat{B}^{(t)}\}_{t=0}^{T}\right)\right] = \lim_{\epsilon_w,\epsilon_B\to 0} \frac{\partial}{\partial \epsilon_B}\Xi_{\mathrm{ST}}(\epsilon_w, \epsilon_B). \tag{10}$$

2. As the aim of this study is not to provide rigorous analysis but to provide theoretical insights by using a non-rigorous heuristic of statistical mechanics, we assume that the exchange of limits and integrals is possible throughout the study without further justification.
3. Readers familiar with statistical mechanics might be interested in the similarity between this and the generating functional of the Martin-Siggia-Rose formalism (Martin et al., 1973), or alternatively called the path integral formulation, used in the dynamical mean-field theory (DMFT), which has recently been used to analyze the learning dynamics (Agoritsas et al., 2018; Mignacco et al., 2020b, 2021; Bordelon and Pehlevan, 2022, 2023a,b; Pehlevan and Bordelon, 2023). Our generating functional is basically the same as that used in the DMFT. However, in our setup, to incorporate the update rule based on the implicit solution of the optimization problem (4), it is necessary to treat the Boltzmann distributions with nontrivial normalization constants using the replica method. This is in contrast to conventional setups of DMFT. In a conventional setup of DMFT, the expression of the parameters at the next time step is explicitly given, and thus the time evolution can be explicitly written using delta functions, which allows us to consider an average directly over the data without the replica method. In this sense, our analysis can be viewed as a DMFT analysis of the implicit update rule using the replica method.

3.2 Replica Method for ST

The evaluation of the generating functional (8) is technically difficult because the average over $\{\boldsymbol{\theta}^{(t)}\}_{t=0}^{T-1}$ and D requires averaging the inverse of the normalization constants $(Z^{(0)}(D_L))^{-1}\prod_{t=1}^{T-1}(Z^{(t)}(D_U^{(t)},\boldsymbol{\theta}^{(t-1)}))^{-1}$ in the Boltzmann distributions:

$$\Xi_{\mathrm{ST}}(\epsilon_w,\epsilon_B) = \lim_{N,\beta\to\infty}\mathbb{E}_D\Bigg[\int e^{\epsilon_w g_w(\{w_i^{(t)}\}_{t=0}^T)+\epsilon_B g_B(\{B^{(t)}\}_{t=0}^T)}\frac{1}{Z^{(0)}(D_L)}e^{-\beta^{(0)}\mathcal{L}^{(0)}(\boldsymbol{\theta}^{(0)};D_L)} \times\prod_{t=1}^{T}\frac{1}{Z^{(t)}(D_U^{(t)},\boldsymbol{\theta}^{(t-1)})}e^{-\beta^{(t)}\mathcal{L}^{(t)}(\boldsymbol{\theta}^{(t)};D_U^{(t)},\boldsymbol{\theta}^{(t-1)})}d\boldsymbol{\theta}^{(0)}\dots d\boldsymbol{\theta}^{(T)}\Bigg], \tag{11}$$

where integration over $\theta^{(0)},\dots,\theta^{(T)}$ in Eq. (11) applies to the whole integrand, which is split over multiple lines for readability. To resolve this difficulty, we use the replica method as follows. This method rewrites Ξ_{ST} using the identity $(Z^{(t)})^{-1} = \lim_{n_t\to 0}(Z^{(t)})^{n_t-1}$ as

$$\begin{aligned}\Xi_{\mathrm{ST}} &= \lim_{n_0,\dots,n_T\to 0}\phi^{(T)}_{n_0,\dots,n_T},\\ \phi^{(T)}_{n_0,\dots,n_T} &= \lim_{N,\beta\to\infty}\mathbb{E}_D\Bigg[\int e^{\epsilon_w g_w(\{w_i^{(t)}\}_{t=0}^T)+\epsilon_B g_B(\{B^{(t)}\}_{t=0}^T)}\\ &\quad\times\left(Z^{(0)}(D_L)\right)^{n_0-1}e^{-\beta^{(0)}\mathcal{L}^{(0)}(\boldsymbol{\theta}^{(0)};D_L)}\\ &\quad\times\prod_{t=1}^{T}\left(Z^{(t)}(D_U^{(t)},\boldsymbol{\theta}^{(t-1)})\right)^{n_t-1}e^{-\beta^{(t)}\mathcal{L}^{(t)}(\boldsymbol{\theta}^{(t)};D_U^{(t)},\boldsymbol{\theta}^{(t-1)})}d\boldsymbol{\theta}^{(0)}\dots d\boldsymbol{\theta}^{(T)}\Bigg].\end{aligned}$$

Although the evaluation of $\phi^{(T)}_{n_0,\dots,n_T}$ for $n_t\in\mathbb{R}$ is difficult, this expression has the advantage explained next. For positive integers $n_0,\dots,n_T = 1,2,\dots$, it has an appealing expression:

$$\begin{aligned}\phi^{(T)}_{n_0,\dots,n_T} &= \lim_{N,\beta\to\infty}\mathbb{E}_D\Bigg[\int e^{\epsilon_w g_w(\{w_{1,i}^{(t)}\}_{t=0}^T)+\epsilon_B g_B(\{B_1^{(t)}\}_{t=0}^T)}\prod_{a_0=1}^{n_0}e^{-\beta^{(0)}\mathcal{L}^{(0)}(\boldsymbol{\theta}_{a_t}^{(0)};D_L)}\\ &\quad\times\prod_{t=1}^{T}\prod_{a_t=1}^{n_t}e^{-\beta^{(t)}\mathcal{L}^{(t)}(\boldsymbol{\theta}_{a_t}^{(t)};D_U^{(t)},\boldsymbol{\theta}_1^{(t-1)})}d^{n_0}\boldsymbol{\theta}^{(0)}\dots d^{n_T}\boldsymbol{\theta}^{(T)}\Bigg],\end{aligned} \tag{12}$$

where $d^{n_t}\boldsymbol{\theta}^{(t)}, t = 0,1,\dots,T$ are the shorthand notations for $d\boldsymbol{\theta}_1^{(t)}\dots d\boldsymbol{\theta}_{n_t}^{(t)}$, and $\{a_t\}_{t=0}^T$ are indices to distinguish the additional variables introduced by the replica method, hereafter we will refer to this type of index as the *replica index*. Note that the index 1 that appears in $\epsilon_w g_w(\{w_{1,i}^{(t)}\}_{t=0}^T)+\epsilon_B g_B(\{B_1^{(t)}\}_{t=0}^T)$ is also the replica index. The augmented system (12), which we refer to as the *replicated system*, is much easier to handle than the original generating functional because all of the factors to be evaluated are now explicit. Thus, we can consider the average over D *before* conducting the integral (or the optimization at $\beta\to\infty$) over $\{\boldsymbol{\theta}_{c_t}\}_{t=0}^T$. In the following, utilizing this formula, we obtain an analytical expression for $n_t\in\mathbb{N}$. Subsequently, under appropriate symmetry assumption, we extrapolate that expression to the limit as $n_t\to 0$.

In the following, we briefly sketch the treatment of the replicated system (12). The readers who are interested in the result can skip Section 3.2.2 and proceed to Section 3.3.

3.2.1 A Remark on the Non-Rigorous Aspect of the Replica Method

This analytical continuation $n_t \to 0$ from $n_t \in \mathbb{N}$ is the procedure that has not yet been formulated in a mathematically rigorous manner. In fact, the sequence of the replicated system (12) for integer n_t alone may not uniquely determine the expression for the replicated system at real $n_t \in \mathbb{R}$. Therefore, this analytic continuation does not have a mathematically precise formulation, although the evaluation of $\phi^{(T)}_{n_0,\dots,n_T}$ for positive integers is a well-defined problem. It is merely an extrapolation based on guessing the form of the function. This aspect is what renders the replica method a non-rigorous technique. However, it is empirically known that extrapolating the results obtained in $n_t \in \mathbb{N}$ yields the correct result for convex optimization problems, in the sense that the same result is later obtained by the other mathematical techniques (Barbier et al., 2019; Gabrié et al., 2018; Charbonneau et al., 2023). Therefore, though the replica analysis includes a procedure whose mathematical validity is not proved, we believe that this does not have a serious impact on our analysis since the cost function at each step of self-training is convex.

3.2.2 Handling of the Replicated System

The important observation is that the Gaussian noise terms in (12) only appear through the following vectors:

$$\begin{aligned}
\boldsymbol{u}^{(0)}_\mu &= \left(\frac{1}{\sqrt{N}} \boldsymbol{w}^{(0)}_1 \cdot \boldsymbol{z}^{(0)}_\mu, \frac{1}{\sqrt{N}} \boldsymbol{w}^{(0)}_2 \cdot \boldsymbol{z}^{(0)}_\mu, \dots, \frac{1}{\sqrt{N}} \boldsymbol{w}^{(0)}_{n_0} \cdot \boldsymbol{z}^{(0)}_\mu \right)^\top \in \mathbb{R}^{n_0}, \\
\boldsymbol{u}^{(t)}_\nu &= \left(\frac{1}{\sqrt{N}} \boldsymbol{w}^{(t-1)}_1 \cdot \boldsymbol{z}^{(t)}_\nu, \frac{1}{\sqrt{N}} \boldsymbol{w}^{(t)}_1 \cdot \boldsymbol{z}^{(t)}_\nu, \frac{1}{\sqrt{N}} \boldsymbol{w}^{(t)}_2 \cdot \boldsymbol{z}^{(t)}_\nu, \dots, \frac{1}{\sqrt{N}} \boldsymbol{w}^{(t)}_{n_t} \cdot \boldsymbol{z}^{(t)}_\nu \right)^\top \in \mathbb{R}^{n_t+1}, \\
& \quad t = 1, 2, \dots, T,
\end{aligned}$$

which follows independently the multivariate Gaussians as

$$\begin{aligned}
\boldsymbol{u}^{(0)}_\mu &\sim_{\text{iid}} \mathcal{N}(0, \Sigma^{(0)}), \quad \Sigma^{(0)} = \Delta_L \times \begin{bmatrix} Q^{(0)}_{11} & \cdots & Q^{(0)}_{1n_0} \\ \vdots & \ddots & \vdots \\ Q^{(0)}_{n_0 1} & \cdots & Q^{(0)}_{n_0 n_0} \end{bmatrix}, \\
Q^{(0)}_{c_0 d_0} &\equiv \frac{1}{N} \boldsymbol{w}^{(0)}_{c_0} \cdot \boldsymbol{w}^{(0)}_{d_0}, \quad c_0, d_0 = 1, 2, \dots, n_0,
\end{aligned}$$

and

$$\begin{aligned}
\boldsymbol{u}^{(t)}_\nu &\sim_{\text{iid}} \mathcal{N}(0, \Sigma^{(t)}), \quad \Sigma^{(t)} = \Delta_U \times \left[\begin{array}{c|ccc} Q^{(t-1)}_{11} & R^{(t)}_1 & \cdots & R^{(t)}_{n_t} \\ \hline R^{(t)}_1 & Q^{(t)}_{11} & \cdots & Q^{(t)}_{1n_t} \\ \vdots & \vdots & \ddots & \vdots \\ R^{(t)}_{n_t} & Q^{(t)}_{n_t 1} & \cdots & Q^{(t)}_{n_t n_t} \end{array}\right] \\
Q^{(t)}_{c_t d_t} &= \frac{1}{N} \boldsymbol{w}^{(t)}_{c_t} \cdot \boldsymbol{w}^{(t)}_{d_t}, \quad R^{(t)}_{c_t} = \frac{1}{N} \boldsymbol{w}^{(t-1)}_1 \cdot \boldsymbol{w}^{(t)}_{c_t}, \quad c_t, d_t = 1, 2, \dots, n_t,
\end{aligned}$$

for a fixed set of $\{\boldsymbol{w}^{(t)}_{c_t}\}_{c_t=1}^{n_t}, t = 0, \dots, T$. Since each data point is independent, the integral over $(\boldsymbol{u}^{(0)}_\mu, y^{(0)}_\mu)$ and $(\boldsymbol{u}^{(t)}_\nu, y^{(t)}_\nu)$ can be taken independently once $\{\boldsymbol{\theta}^{(t)}\}$ are fixed. As a

result, the generating functional does not depend on the index μ and ν, hence, we can omit it.

Above observation indicates that the replicated system (12) depends on $\{\boldsymbol{w}_{c_t}^{(t)}\}_{c_t=1}^{n_t}, t = 0, \ldots, T$ only through their inner products, such as

$$\begin{aligned}
&\frac{1}{N}\boldsymbol{w}_{c_t}^{(t)} \cdot \boldsymbol{w}_{d_t}^{(t)}, \quad c_t, d_t = 1, 2, \ldots, n_t,\ t = 0, 1, \ldots T, \\
&\frac{1}{N}\boldsymbol{w}_{c_t}^{(t)} \cdot \boldsymbol{v}, \quad c_t = 1, \ldots, n_t,\ t = 0, 1, \ldots, T, \\
&\frac{1}{N}\boldsymbol{w}_{1}^{(t-1)} \cdot \boldsymbol{w}_{c_t}^{(t)}, \quad c_t = 1, 2, \ldots, n_t,\ t = 1, \ldots, T,
\end{aligned}$$

which capture the geometric relations between the estimators and the centroid of clusters $\boldsymbol{v}$. We refer to them as *order parameters*. Thus, by introducing the auxiliary variables through the trivial identities of the delta functions

$$\begin{aligned}
1 &= \prod_{t=0}^{T} \prod_{1 \le c_t \le d_t \le n_t} N \int \delta_{\mathrm{d}}\left(NQ_{c_t d_t}^{(t)} - \boldsymbol{w}_{c_t}^{(t)} \cdot \boldsymbol{w}_{d_t}\right) dQ_{c_t d_t}^{(t)}, \\
1 &= \prod_{t=0}^{T} \prod_{c_t=1}^{n_t} N \int \delta_{\mathrm{d}}\left(Nm_{c_t}^{(t)} - \boldsymbol{w}_{c_t}^{(t)} \cdot \boldsymbol{v}\right) dm_{c_t}^{(t)}, \\
1 &= \prod_{t=1}^{T} \prod_{c_t=1}^{n_t} N \int \delta_{\mathrm{d}}\left(NR_{c_t}^{(t)} - \boldsymbol{w}_{1}^{(t-1)} \cdot \boldsymbol{w}_{c_t}^{(t)}\right) dR_{c_t}^{(t)},
\end{aligned}$$

and their Fourier representations

$$\begin{aligned}
\delta_{\mathrm{d}}\left(NQ_{c_t d_t}^{(t)} - \boldsymbol{w}_{c_t}^{(t)} \cdot \boldsymbol{w}_{d_t}\right) &= \frac{1}{4\pi} \int e^{-(NQ_{c_t d_t}^{(t)} - \boldsymbol{w}_{c_t}^{(t)} \cdot \boldsymbol{w}_{d_t}) \frac{\tilde{Q}_{c_t d_t}^{(t)}}{2}} d\tilde{Q}_{c_t d_t}^{(t)}, \\
\delta_{\mathrm{d}}\left(Nm_{c_t}^{(t)} - \boldsymbol{w}_{c_t}^{(t)} \cdot \boldsymbol{v}\right) &= \frac{1}{2\pi} \int e^{(Nm_{c_t}^{(t)} - \boldsymbol{w}_{c_t}^{(t)} \cdot \boldsymbol{v})\tilde{m}_{c_t}^{(t)}} d\tilde{m}_{c_t}^{(t)}, \\
\delta_{\mathrm{d}}\left(NR_{c_t}^{(t)} - \boldsymbol{w}_{1}^{(t-1)} \cdot \boldsymbol{w}_{c_t}^{(t)}\right) &= \frac{1}{2\pi} \int e^{(NR_{c_t}^{(t)} - \boldsymbol{w}_{1}^{(t-1)} \cdot \boldsymbol{w}_{c_t}^{(t)})\tilde{R}_{c_t}^{(t)}} d\tilde{R}_{c_t}^{(t)},
\end{aligned}$$

the integrals over $\{w_{a_t,j}^{(t)}\}$ can be independently performed for each j. As a result, the result no longer depends on the index i, hence, we can safely omit it. This is a natural consequence of the data distribution having rotation-invariant symmetry. For $t \ge 1$, let $\boldsymbol{\Theta}^{(t)}$ and $\hat{\boldsymbol{\Theta}}^{(t)}$ be the collection of the variables

$$Q^{(t)} = [Q_{c_t d_t}^{(t)}]_{\substack{1 \le c_t \le n_t \\ 1 \le d_t \le n_t}},\ \boldsymbol{m}^{(t)} = [m_{c_t}^{(t)}]_{1 \le c_t \le n_t},\ \boldsymbol{R}^{(t)} = [R_{c_t}^{(t)}]_{1 \le c_t \le n_t},\ \boldsymbol{B}^{(t)} = [b_{c_t}^{(t)}]_{1 \le c_t \le n_t},$$

and

$$\tilde{Q}^{(t)} = [\tilde{Q}_{c_t d_t}^{(t)}]_{\substack{1 \le c_t \le n_t \\ 1 \le d_t \le n_t}},\ \tilde{\boldsymbol{m}}^{(t)} = [\tilde{m}_{c_t}^{(t)}]_{1 \le c_t \le n_t},\ \tilde{\boldsymbol{R}}^{(t)} = [\tilde{R}_{c_t}^{(t)}]_{1 \le c_t \le n_t},$$

respectively. $\Theta^{(0)}$ and $\hat{\Theta}^{(0)}$ are defined similarly. Also, let Θ and $\hat{\Theta}$ be $\{\Theta\}_{t=0}^{T}, \{\hat{\Theta}\}_{t=0}^{T}$, respectively. Then, we find that $\phi_{n_0,\ldots,n_T}^{(T)}$ can be written as

$$\phi_{n_0,\ldots,n_T}^{(T)} = \lim_{N,\beta \to \infty} \int e^{N\mathcal{S}(\Theta,\hat{\Theta})} \mathbb{E}_{\{\boldsymbol{w}^{(t)}\} \sim \tilde{p}_{\mathrm{eff},w}} \left[e^{\epsilon_w g_w(\{w_1^{(t)}\}_{t=0}^{T})} \right] e^{\epsilon_B g_B(\{B_1^{(t)}\}_{t=0}^{T})} d\Theta d\hat{\Theta}, \quad (13)$$

where

$$\tilde{p}_{\mathrm{eff},w}(\{\boldsymbol{w}^{(t)}\}_{t=0}^{T}) \propto e^{-\frac{1}{2}(\boldsymbol{w}^{(0)})^{\top}(\tilde{Q}^{(0)}+\beta^{(t)}I_{n_t})\boldsymbol{w}^{(0)}+(\tilde{\boldsymbol{m}}^{(t)})\cdot\boldsymbol{w}^{(0)}} \times \prod_{t=1}^{T} e^{-\frac{1}{2}(\boldsymbol{w}^{(t)})^{\top}(\tilde{Q}^{(t)}+\beta^{(t)}I_{n_t})\boldsymbol{w}^{(t)}+(\tilde{\boldsymbol{m}}^{(t)}+w_1^{(t-1)}\boldsymbol{R}^{(t)})\cdot\boldsymbol{w}^{(t)}}.$$

Thus, Θ and $\hat{\Theta}$ are evaluated by the extremum condition $\mathrm{extr}_{\Theta,\hat{\Theta}}\mathcal{S}(\Theta,\hat{\Theta})$ when $N \gg 1$, where $\mathrm{extr}_{\Theta,\hat{\Theta}}\dots$ represents an operation taking extremum of a function over $\Theta,\hat{\Theta}$. The concrete form of $\mathcal{S}$ is given as (46) in Appendix A.2. $\tilde{p}_{\mathrm{eff},w}$ is a density function over $\{\boldsymbol{w}^{(t)}\}_{t=0}^{T}$. Here $\boldsymbol{w}^{(t)} = (w_1^{(t)},\dots w_{n_t}^{(t)})\cdot \in \mathbb{R}^{n_t}$. Recall that the index 1 in (13) is the replica index, and the result no longer depends on the index $i \in [N]$. It should also be noted that essentially only Gaussian integrals and the saddle-point method are required for the above calculations. This computational simplicity is a major advantage of the replica method, which introduces the replicated system (12) that allows us to first consider the average over the noise.

The key non-trivial point is that, in order to extrapolate as $n_0,\dots,n_T \to 0$, we need to impose some symmetry on the saddle point. Otherwise, the discreteness of n_t prevents us from extrapolating as $n_t \to 0$. The simplest choice is the replica symmetric (RS) one:

$$\begin{aligned}
Q^{(t)} &= \frac{\chi^{(t)}}{\beta^{(t)}}I_{n_t} + q^{(t)}\mathbf{1}_{n_t}\mathbf{1}_{n_t}\cdot, && (14)\\
\boldsymbol{R}^{(t)} &= R^{(t)}\mathbf{1}_{n_t},\\
\boldsymbol{m}^{(t)} &= m^{(t)}\mathbf{1}_{n_t},\\
\boldsymbol{B}^{(t)} &= B^{(t)}\mathbf{1}_{n_t},\\
\tilde{Q}^{(t)} &= \beta^{(t)}\hat{Q}^{(t)}I_{n_t} - (\beta^{(t)})^2\hat{\chi}^{(t)}\mathbf{1}_{n_t}\mathbf{1}_{n_t}\cdot,\\
\tilde{\boldsymbol{R}}^{(t)} &= \beta^{(t)}\hat{R}^{(t)}\mathbf{1}_{n_t},\\
\tilde{\boldsymbol{m}}^{(t)} &= \beta^{(t)}\hat{m}^{(t)}\mathbf{1}_{n_t}, && (15)
\end{aligned}$$

that reflect the symmetry of the replicated system (12). Under this ansatz, for $t \geq 1$, $\Theta^{(t)}$ and $\hat{\Theta}^{(t)}$ represents $(q^{(t)},\chi^{(t)},R^{(t)},m^{(t)},B^{(t)})$ and $(\hat{Q}^{(t)},\hat{\chi}^{(t)},\hat{R}^{(t)},\hat{m}^{(t)})$, respectively. Imposing this symmetry, we can obtain an analytical expression that can be formally extrapolated as $n_t \in \mathbb{R}$ from $n_t \in \mathbb{N}$. Specifically, let $\mathcal{O}(n) \equiv \sum_{t=0}^{T}\mathcal{O}(n_t)$ be terms that vanish at the limit $n_0,n_1,\dots,n_T \to 0$. Then, under the RS assumption, it can be shown that $\mathcal{S}(\Theta,\hat{\Theta}) = \mathcal{O}(n)$. Hence $e^{N\mathcal{S}(\Theta,\hat{\Theta})}$ yields a factor of unity at $n \to 0$. Thus, the generating functional can be written as

$$\phi_{n_0,\dots,n_T}^{(T)} = \lim_{\beta\to\infty}\mathbb{E}_{\xi_w^{(t)}\sim_{\mathrm{iid}}\mathcal{N}(0,1)}\left[\mathbb{E}_{\{w^{(t)}\}\sim p_{\mathrm{eff},w}(\{w^{(t)}\}|\{\xi_w^{(t)}\})}\left[e^{\epsilon_w g_w(\{w^{(t)}\}_{t=0}^{T})}\right]\right] \times e^{\epsilon_B g_B(\{B^{(t)}\}_{t=0}^{T})} + \mathcal{O}(n),$$

where

$$p_{\mathrm{eff},w}(\{w^{(t)}\}_{t=0}^{T}|\{\xi_w^{(t)}\}) = \mathcal{N}\left(w^{(0)} \mid \frac{\hat{m}^{(0)} + \sqrt{\hat{\chi}^{(0)}}\xi_w^{(0)}}{\hat{Q}^{(0)} + \lambda^{(0)}}, \frac{\hat{Q}^{(0)} + \lambda^{(0)}}{\beta^{(0)}}\right) \times \prod_{t=1}^{T} \mathcal{N}\left(w^{(t)} \mid \frac{\hat{m}^{(t)} + \hat{R}^{(t)} w^{(t-1)} + \sqrt{\hat{\chi}^{(t)}}\xi_w^{(t)}}{\hat{Q}^{(t)} + \lambda^{(t)}}, \frac{\hat{Q}^{(t)} + \lambda^{(t)}}{\beta^{(t)}}\right),$$

This implies that, at the limit $\beta \to \infty, n_0, \ldots, n_T \to 0$, it is governed by a one-dimensional Gaussian process:

$$\begin{aligned}
\Xi_{\mathrm{ST}}(\epsilon_w, \epsilon_B) &= \mathbb{E}_{\xi_w^{(t)} \sim_{\mathrm{iid}} \mathcal{N}(0,1)} \left[e^{\epsilon_w g_w(\{\hat{\mathsf{w}}^{(t)}\}_{t=0}^{T})} \right] e^{\epsilon_B g_B(\{B^{(t)}\}_{t=0}^{T})}, \\
\hat{\mathsf{w}}^{(0)} &= \frac{\hat{m}^{(0)} + \sqrt{\hat{\chi}^{(0)}}\xi_w^{(0)}}{\hat{Q}^{(0)} + \lambda^{(0)}}, \\
\hat{\mathsf{w}}^{(t)} &= \frac{\hat{m}^{(t)} + \hat{R}^{(t)} w^{(t-1)} + \sqrt{\hat{\chi}^{(t)}}\xi_w^{(t)}}{\hat{Q}^{(t)} + \lambda^{(t)}},
\end{aligned}$$

where, $\Theta, \hat{\Theta}$ are determined by the saddle point condition.

Similarly, it can be possible to consider another generating functional regarding $y_\nu^{(t)}$, $\tilde{u}_\nu = \boldsymbol{x}_\nu^{(t)} \cdot \boldsymbol{w}^{(t-1)}/\sqrt{N} + B^{(t-1)}$ and $u_\nu = \boldsymbol{x}_\nu^{(t)} \cdot \boldsymbol{w}^{(t)}/\sqrt{N} + B^{(t)}$:

$$\Xi_{\mathrm{ST}}(\epsilon_u) = \lim_{N,\beta \to \infty} \mathbb{E}_{\{\boldsymbol{\theta}^{(t)}\}_{t=0}^{T} \sim p_{\mathrm{ST}}, D} \left[e^{\epsilon_u g_u(\{(y_\nu^{(t)}, \tilde{u}_\nu^{(t)}, u_\nu^{(t)})\}_{t=1}^{T})} \right], \tag{16}$$

where $\nu \in [M_U]$ and $t \geq 1$. The calculation procedure is completely analogous to that of $\Xi_{\mathrm{ST}}(\epsilon_w, \epsilon_B)$. Analyzing this generating functional yields the statistical properties of the logits. See Appendix A.3 for more detail.

The generating functional obtained by imposing this symmetry is called the *RS solution*. As already commented in the beginning of subsection 3.2, in general, the expression for the replicated system (12) with integer $\{n_t\}_{t=0}^{T}$ alone cannot uniquely determine the expression for the replicated system at real $\{n_t\}_{t=0}^{T}$. However, for log-convex Boltzmann distributions, it has been empirically known that the replica symmetric choice of the saddle point, which should be valid for $n_t \in \mathbb{N}$, yields the correct extrapolation (Gabrié et al., 2018; Barbier and Macris, 2019; Barbier et al., 2019; Mignacco et al., 2020a; Gerbelot et al., 2020, 2023; Montanari and Sen, 2024) in the sense that the same result by the replica method with the RS assumption have been obtained through a different mathematically rigorous approach. Since the Boltzmann distributions in our setup are log-convex functions once conditioned on the data and the parameter of the previous iteration step, we can expect the RS assumption to yield the correct result even in the current iterative optimization.

3.3 RS Solution

In this subsection, we present the asymptotic properties of the regressors obtained by ST under the RS ansatz on the saddle point (14)-(15). It is described by a small finite set of scalar quantities determined as a solution of nonlinear equations, which we refer to as *self-consistent equations*. See Appendix A for the derivations.

The statements in this subsection are non-rigorous predictions obtained from the replica calculation under the RS ansatz. Hence, we refer to them as Predictions, rather than Theorems, to make their mathematical status explicit. They characterize the predicted large-system limit of the estimator and the logits. Their accuracy is assessed against finite-size numerical experiments in Section 3.4.

3.3.1 RS Saddle Point

First, we define the self-consistent equations that determine the scalar order parameters $\Theta^{(t)}$ and $\hat{\Theta}^{(t)}$. At a high level, these parameters provide a finite-dimensional effective description of the high-dimensional estimator at step t. The parameters in $\Theta^{(t)}$ describe macroscopic properties of the learned classifier, such as the norm of the weight vector, its alignment with the cluster direction. These quantities determine the effective distribution of the logits. The conjugate parameters in $\hat{\Theta}^{(t)}$ govern the associated scalar effective process for individual coordinates of the learned weight vector. The two sets of parameters are jointly determined by the self-consistent equations below. Let us define $l_L(y,x)$ and $l_U(y,x;\tilde{\Gamma})$ as

$$\begin{aligned} l_L(y,x) &= l(y,\sigma(x)),\\ l_U(y,x;\tilde{\Gamma}) &= \mathbb{1}(|y|>\tilde{\Gamma}) l_{\mathrm{pl}}(\sigma_{\mathrm{pl}}(y),\sigma(x)). \end{aligned}$$

Then, the self-consistent equations are summarized as follows.

Definition 1 (Self-consistent equations) *The following set of quantities $\Theta = \{\Theta^{(t)}\}_{t=0}^{T}, \hat{\Theta} = \{\hat{\Theta}^{(t)}\}_{t=0}^{T}$ are determined as the solution of the following set of non-linear equations that are referred to as self-consistent equations:*

$$\begin{aligned} \Theta^{(0)} &= (q^{(0)}, \chi^{(0)}, m^{(0)}, B^{(0)}),\\ \Theta^{(t)} &= (q^{(t)}, \chi^{(t)}, R^{(t)}, m^{(t)}, B^{(t)}), \quad t \in [T],\\ \hat{\Theta}^{(0)} &= (\hat{Q}^{(0)}, \hat{\chi}^{(0)}, \hat{m}^{(0)}),\\ \hat{\Theta}^{(t)} &= (\hat{Q}^{(t)}, \hat{\chi}^{(t)}, \hat{R}^{(t)}, \hat{m}^{(t)}), \quad t \in [T]. \end{aligned}$$

Let $\hat{\mathfrak{u}}^{(0)}$ be the solution of the following one-dimensional randomized optimization problem:

$$\begin{aligned} \hat{\mathfrak{u}}^{(0)} &= \arg\min_{u^{(0)}\in\mathbb{R}} \left[\frac{(u^{(0)})^2}{2\Delta_L \chi^{(0)}} + l\left(y^{(0)}, \sigma\left(h_u^{(0)} + u^{(0)}\right)\right)\right],\\ h_u^{(0)} &= (2y^{(0)}-1)m^{(0)} + B^{(0)} + \sqrt{q^{(0)}}\xi_u^{(0)}, \quad \xi_u^{(0)} \sim \mathcal{N}(0,1). \end{aligned}$$

Then, the self-consistent equation for $\Theta^{(0)}$ and $\hat{\Theta}^{(0)}$ are given as follows:

$$\begin{aligned} q^{(0)} &= \frac{(\hat{m}^{(0)})^2 + \hat{\chi}^{(0)}}{(\hat{Q}^{(0)} + \lambda^{(0)})^2}, \\ \chi^{(0)} &= \frac{1}{\hat{Q}^{(0)} + \lambda^{(0)}}, \\ m^{(0)} &= \frac{\hat{m}^{(0)}}{\hat{Q}^{(0)} + \lambda^{(0)}}, \end{aligned} \tag{17}$$

$$
\begin{aligned}
\hat{Q}^{(0)} &= \alpha_L \Delta_L \mathbb{E}_{\xi_u^{(0)} \sim \mathcal{N}(0,\Delta_L), y^{(0)} \sim p_{y,L}} \left[\frac{d}{dh_u^{(0)}} \partial_2 l_L(y^{(0)}, h_u^{(0)} + \hat{\mathfrak{u}}^{(0)}) \right], \\
\hat{\chi}^{(0)} &= \alpha_L \Delta_L \mathbb{E}_{\xi_u^{(0)} \sim \mathcal{N}(0,\Delta_L), y^{(0)} \sim p_{y,L}} \left[\left(\partial_2 l_L(y^{(0)}, h_u^{(0)} + \hat{\mathfrak{u}}^{(0)}) \right)^2 \right], \\
\hat{m}^{(0)} &= \alpha_L \mathbb{E}_{\xi_u^{(0)} \sim \mathcal{N}(0,\Delta_L), y^{(0)} \sim p_{y,L}} \left[(2y-1) \partial_2 l_L(y^{(0)}, h_u^{(0)} + \hat{\mathfrak{u}}^{(0)}) \right], \\
0 &= \mathbb{E}_{\xi_u^{(0)} \sim \mathcal{N}(0,\Delta_L), y^{(0)} \sim p_{y,L}} \left[\partial_2 l_L(y^{(0)}, h_u^{(0)} + \hat{\mathfrak{u}}^{(0)}) \right],
\end{aligned}
$$

Similarly, let $\hat{\mathfrak{u}}^{(t)}$ *be the solution of the following randomized optimization problem:*

$$
\hat{\mathfrak{u}}^{(t)} = \arg\min_{u^{(t)} \in \mathbb{R}} \left[\frac{(u^{(t)})^2}{2\Delta_U \chi^{(t)}} + \mathbb{1}(|\tilde{h}_u^{(t)}| > \Gamma \sqrt{q^{(t-1)}}) l_{\mathrm{pl}} \left(\sigma_{\mathrm{pl}} \left(\tilde{h}_u^{(t)} \right), \sigma \left(h_u^{(t)} + u^{(t)} \right) \right) \right], \tag{18}
$$

$$
\tilde{h}_u^{(t)} = (2y^{(t)} - 1) m^{(t-1)} + B^{(t-1)} + \sqrt{q^{(t-1)}} \xi_{u,1}^{(t)}, \tag{19}
$$

$$
h_u^{(t)} = (2y^{(t)} - 1) m^{(t)} + B^{(t)} + \frac{R^{(t)}}{\sqrt{q^{(t-1)}}} \xi_{u,1}^{(t)} + \sqrt{q^{(t)} - \frac{(R^{(t)})^2}{q^{(t-1)}}} \xi_{u,2}^{(t)}, \tag{20}
$$

$$
\xi_{u,1}^{(t)}, \xi_{u,2}^{(t)} \sim \mathcal{N}(0,1).
$$

Then, the self-consistent equations for $\Theta^{(t)}, \hat{\Theta}^{(t)}, t \in [T]$ *are given as follows:*

$$
q^{(t)} = \frac{(\hat{m}^{(t)})^2 + \hat{\chi}^{(t)} + (\hat{R}^{(t)})^2 q^{(t-1)} + 2\hat{m}^{(t)} \hat{R}^{(t)} m^{(t-1)}}{(\hat{Q}^{(t)} + \lambda^{(t)})^2}, \tag{21}
$$

$$
\chi^{(t)} = \frac{1}{\hat{Q}^{(t)} + \lambda^{(t)}},
$$

$$
m^{(t)} = \frac{\hat{m}^{(t)} + \hat{R}^{(t)} m^{(t-1)}}{\hat{Q}^{(t)} + \lambda^{(t)}},
$$

$$
R^{(t)} = \frac{\hat{m}^{(t)} m^{(t-1)} + \hat{R}^{(t)} q^{(t-1)}}{\hat{Q}^{(t)} + \lambda^{(t)}}, \tag{22}
$$

$$
\hat{Q}^{(t)} = \alpha_U^{(t)} \Delta_U^{(t)} \mathbb{E}_{\xi_{u,1}^{(t)}, \xi_{u,2}^{(t)} \sim \mathcal{N}(0,\Delta_U^{(t)}), y^{(t)} \sim p_y^{(t)}} \left[\frac{d}{dh_u^{(t)}} \partial_2 l_U(\tilde{h}_u^{(t)}, h_u^{(t)} + \hat{\mathfrak{u}}^{(t)}; \Gamma \sqrt{q^{(t-1)}}) \right],
$$

$$
\hat{\chi}^{(t)} = \alpha_U^{(t)} \Delta_U^{(t)} \mathbb{E}_{\xi_{u,1}^{(t)}, \xi_{u,2}^{(t)} \sim \mathcal{N}(0,\Delta_U^{(t)}), y^{(t)} \sim p_y^{(t)}} \left[\left(\partial_2 l_U(\tilde{h}_u^{(t)}, h_u^{(t)} + \hat{\mathfrak{u}}^{(t)}; \Gamma \sqrt{q^{(t-1)}}) \right)^2 \right], \tag{23}
$$

$$
\hat{m}^{(t)} = \alpha_U^{(t)} \mathbb{E}_{\xi_{u,1}^{(t)}, \xi_{u,2}^{(t)} \sim \mathcal{N}(0,\Delta_U^{(t)}), y^{(t)} \sim p_y^{(t)}} \left[(2y^{(t)} - 1) \partial_2 l_U(\tilde{h}_u^{(t)}, h_u^{(t)} + \hat{\mathfrak{u}}^{(t)}; \Gamma \sqrt{q^{(t-1)}}) \right], \tag{24}
$$

$$
\hat{R}^{(t)} = -\alpha_U^{(t)} \Delta_U^{(t)} \mathbb{E}_{\xi_{u,1}^{(t)}, \xi_{u,2}^{(t)} \sim \mathcal{N}(0,\Delta_U^{(t)}), y^{(t)} \sim p_y^{(t)}} \left[\frac{d}{d\tilde{h}_u^{(t)}} \partial_2 l_U(\tilde{h}_u^{(t)}, h_u^{(t)} + \hat{\mathfrak{u}}^{(t)}; \Gamma \sqrt{q^{(t-1)}}) \right] \tag{25}
$$

$$
0 = \mathbb{E}_{\xi_{u,1}^{(t)}, \xi_{u,2}^{(t)} \sim \mathcal{N}(0,\Delta_U^{(t)}), y^{(t)} \sim p_y^{(t)}} \left[\partial_2 l_U(\tilde{h}_u^{(t)}, h_u^{(t)} + \hat{\mathfrak{u}}^{(t)}; \Gamma \sqrt{q^{(t-1)}}) \right], \tag{26}
$$

3.3.2 Generating Functional for the Model Parameter

The solution of the above self-consistent equations gives the RS solution of the generating functional (8).

Prediction 1 *Let $\hat{\Theta}^{(t)}, t \geq 0$ be the solution of the self-consistent equations in Definition 1. Then, the generating functional* (8) *is given as follows:*

$$\Xi_{\mathrm{ST}}(\epsilon_w, \epsilon_B) = \mathbb{E}_{\xi_w^{(t)} \sim_{\mathrm{iid}} \mathcal{N}(0,1)} \left[e^{\epsilon_w g_w(\{\hat{\mathsf{w}}^{(t)}\}_{t=0}^T)} \right] e^{\epsilon_B g_B(\{B^{(t)}\}_{t=0}^T)},$$

where $\{\hat{\mathsf{w}}^{(t)}\}_{t=0}^T$ follows the following Gaussian process:

$$\hat{\mathsf{w}}^{(0)} = \frac{\hat{m}^{(0)} + \sqrt{\hat{\chi}^{(0)}}\xi_w^{(0)}}{\hat{Q}^{(0)} + \lambda^{(0)}}, \quad \hat{\mathsf{w}}^{(t)} = \frac{\hat{m}^{(t)} + \hat{R}^{(t)}\hat{\mathsf{w}}^{(t-1)} + \sqrt{\hat{\chi}^{(t)}}\xi_w^{(t)}}{\hat{Q}^{(t)} + \lambda^{(t)}}. \tag{27}$$

In addition, using the formulae of the generating functional (9) and (10), we can see that the averaged quantities regarding the weight and the bias are determined as follows.

Prediction 2 *Let $\{\hat{\mathsf{w}}^{(t)}\}_{t=0}^T$ be the trajectory of the Gaussian process* (27) *in Prediction 1, and $\{B^{(t)}\}_{t=0}^T$ be the solution of self-consistent equation in Definition 1. Then, the averaged quantities $\mathbb{E}_D[g_w(\{\hat{w}_i^{(t)}\})], i \in [N]$ and $\mathbb{E}_D[g_B(\{\hat{B}^{(t)}\})]$ are obtained as follows:*

$$\begin{aligned}
&\mathbb{E}_D[g_w(\{\hat{w}_i^{(t)}\})] = \mathbb{E}_{\xi_w^{(t)} \sim_{\mathrm{iid}} \mathcal{N}(0,1)} \left[g_w(\{\hat{\mathsf{w}}^{(t)}\}_{t=0}^T) \right], \\
&\mathbb{E}_D[g_B(\{\hat{B}^{(t)}\})] = g_B(\{B^{(t)}\}_{t=0}^T).
\end{aligned}$$

Hence, the next result also follows:

$$\mathbb{E}_D[\frac{1}{N}\sum_{i=1}^N g_w(\{\hat{w}_i^{(t)}\})] = \mathbb{E}_{\xi_w^{(t)} \sim_{\mathrm{iid}} \mathcal{N}(0,1)} \left[g_w(\{\hat{\mathsf{w}}^{(t)}\}_{t=0}^T) \right].$$

This prediction indicates that each element of the weight vectors $\{\hat{w}_i^{(t)}\}_{t=0}^T, i \in [N]$ is statistically equivalent to the random variables $\{\hat{\mathsf{w}}^{(t)}\}_{t=0}^T$, which behaves as an effective surrogate for $\{\hat{w}_i^{(t)}\}_{t=0}^T$, i.e., $\{\hat{w}_i^{(t)}\} \overset{\mathrm{d}}{=} \hat{\mathsf{w}}^{(t)}$. Here, it is understood that $\xi_w^{(t)}$ effectively plays the role of randomness coming from data D. In this effective description, the parameter $\hat{\Theta}^{(t)}$ has clear meanings; (i) $\hat{m}^{(t)}$ describes the correlation with the direction of $\boldsymbol{v} = (1, \ldots, 1)$, (ii) $\hat{R}^{(t)}$ describes the amount of the information propagation from the previous step, (iii) $\hat{\chi}^{(t)}$ is the strength of the noise, and (iv) $\hat{Q}^{(t)}$ is the amount of confidence. Schematically, it can be represented as follows:

$$\hat{\mathsf{w}}^{(t)} = \underbrace{\frac{1}{\hat{Q}^{(t)} + \lambda_U}}_{\text{confidence}} \left(\underbrace{\hat{m}^{(t)}}_{\text{signal}} + \underbrace{\sqrt{\hat{\chi}^{(t)}}\xi_w^{(t)}}_{\text{noise}} + \underbrace{\hat{R}^{(t)}\hat{\mathsf{w}}^{(t-1)}}_{\substack{\text{information} \\ \text{propagation}}} \right). \tag{28}$$

Hence, the statistical properties of the components of the high-dimensional weight vector are actually described by a Gaussian process that is governed by a finite number of parameters. These parameters are determined by the self-consistent equations in Definition 1. This finite-dimensional description enables us to evaluate the generalization error in the asymptotic limit efficiently, and to search for the optimal hyperparameters. Such a hyperparameter optimization is computationally difficult in finite-size simulations, because it requires a black-box optimization, such as the Nelder-Mead method, applied to results that are averaged over many realizations of the dataset.

Remark: Although the above formula is for averaged quantities, we expect that the values of macroscopic quantities of the form $\frac{1}{N}\sum_{i=1}^{N} g(\hat{w}_i^{(T)})$, such as the quantities in (7), and the biases will concentrate at LSL; in other words, their variances vanish in LSL. Although we do not prove this concentration property, such concentration properties have been rigorously proven in analyses of convex optimization or Bayes-optimal inferences, such as the logistic regression (Mignacco et al., 2020a), Bayes-optimal the generalized linear estimation (Barbier et al., 2019). Since the optimization at each step of ST is convex, we expect the concentration property in our setting, too.

Under the above assumption of concentration, the parameter Θ that is the solution of the self-consistent equations can be interpreted as follows. Let $\hat{\boldsymbol{\theta}}_{\epsilon_i}^{(t)}$ be the estimator obtained when a small perturbation is added to the original cost function:

$$\hat{\boldsymbol{\theta}}_{\epsilon_i}^{(t)} = \arg\min_{\boldsymbol{\theta}^{(t)}} \mathcal{L}^{(t)}(\boldsymbol{\theta}^{(t)} | D_U^{(t)}, \hat{\boldsymbol{\theta}}^{(t-1)}) + \epsilon_i w_i^{(t)},$$

where the additional term $\epsilon_i w_i^{(t)}$ is included to measure the stability of the solution $\hat{\boldsymbol{\theta}}^{(t)}$ at $\epsilon_i = 0$. Then $q^{(t)}, \chi^{(t)}, m^{(t)}$, and $R^{(t)}$ can be interpreted as follows:

$$q^{(t)} = \lim_{N\to\infty} \frac{1}{N} \|\hat{\boldsymbol{w}}^{(t)}\|_2^2, \tag{29}$$

$$\begin{aligned}\chi^{(t)} &= \lim_{N,\beta\to\infty} \frac{1}{N}\sum_{i=1}^{N} \beta^{(t)} \mathbb{V}\mathrm{ar}_{p_{\mathrm{ST}}}\left[w_i^{(t)}\right] \\ &= \lim_{N\to\infty,\epsilon_i\to 0} \frac{1}{N}\sum_{i=1}^{N} \frac{\partial \hat{w}_{\epsilon_i,i}^{(t)}}{\partial \epsilon_i},\end{aligned}$$

$$m^{(t)} = \lim_{N\to\infty} \frac{1}{N} \boldsymbol{v}\cdot\hat{\boldsymbol{w}}^{(t)}, \tag{30}$$

$$R^{(t)} = \lim_{N\to\infty} \frac{1}{N} \hat{\boldsymbol{w}}^{(t-1)}\cdot\hat{\boldsymbol{w}}^{(t)}, \tag{31}$$

The interpretation of $q^{(t)}, m^{(t)}$ and $R^{(t)}$ is clear from the right-hand-sides of these equations. The interpretation of $\chi^{(t)}$ may be slightly non-trivial, but it can be understood as representing how the learning result of the weight vector is stable against a small perturbation to the loss function. We shall refer to $\chi^{(t)}$ as *linear susceptibility* following the custom of statistical mechanics.

3.3.3 Generalization Error

Using the solution of the self-consistent equations, the generalization error (5) is obtained as follows.

Prediction 3 *Let $q^{(t)}, m^{(t)}, B^{(t)}, t = 0, \ldots, T$ be the solution of the self-consistent equations in Definition 1. Then, at LSL, the macroscopic quantities* (7) *and the generalization error* (5) *are obtained as follows:*

$$\hat{B}^{(t)} = B^{(t)}, \tag{32}$$

$$\bar{\epsilon}_{\mathrm{g}}^{(t)} = \epsilon_{\mathrm{g}}^{(t)} \equiv \rho_U H\left(\frac{m^{(t)} + B^{(t)}}{\sqrt{\Delta_U q^{(t)}}}\right) + (1-\rho_U) H\left(\frac{m^{(t)} - B^{(t)}}{\sqrt{\Delta_U q^{(t)}}}\right). \tag{33}$$

Thus, by numerically solving the self-consistent equations (17)-(18), we can precisely evaluate the generalization error (5) in LSL with a limited number of variables. Notice that now the problem is finite-dimensional. In Appendix G, we sketch how to obtain numerical solutions of the self-consistent equations. Also, the cosine similarity between the direction of the classification plane $\hat{\boldsymbol{w}}^{(t)}$ and the direction of the cluster center $\boldsymbol{v}$ can be evaluated by using the above formula as

$$\frac{\hat{\boldsymbol{w}}^{(t)} \cdot \boldsymbol{v}}{\|\hat{\boldsymbol{w}}^{(t)}\|_2 \|\boldsymbol{v}\|_2} \to \frac{m^{(t)}}{\sqrt{q^{(t)}}}. \tag{34}$$

3.3.4 Generating Functional for the Logits

As in the previous discussion of weight vectors, we can characterize the statistical properties regarding $y_\nu^{(t)}$, $\tilde{u}_\nu = \boldsymbol{x}_\nu^{(t)} \cdot \boldsymbol{w}^{(t-1)}/\sqrt{N} + B^{(t-1)}$ and $u_\nu = \boldsymbol{x}_\nu^{(t)} \cdot \boldsymbol{w}^{(t)}/\sqrt{N} + B^{(t)}$ by computing another generating functional (16).

Prediction 4 *Let $\hat{\mathsf{u}}^{(t)}$ be the solution of the one-dimensional optimization problem in* (18) *otherwise. Also, $\tilde{h}_u^{(t)}$ and $h_u^{(t)}$ be the quantity defined in* (19) *and* (20)*, respectively. Then, the averaged quantities $\mathbb{E}[g_u(\{(y_\nu^{(t)}, \tilde{u}_\nu^{(t)}, u_\nu^{(t)})\}_{t=1}^T)]$ can be evaluated as follows:*

$$\mathbb{E}_D\left[g_u(\{(y_\nu^{(t)}, \tilde{u}_\nu^{(t)}, u_\nu^{(t)})\}_{t=1}^T)\right] = \mathbb{E}_{\{y^{(t)}, \xi_{u,1}^{(t)}, \xi_{u,2}^{(t)}\}_{t=1}^T}\left[g_u\left(\{(y^{(t)}, \tilde{h}_u^{(t)}, \hat{\mathsf{u}}^{(t)} + h_u^{(t)})\}_{t=1}^T\right)\right].$$

Hence, the next result also follows:

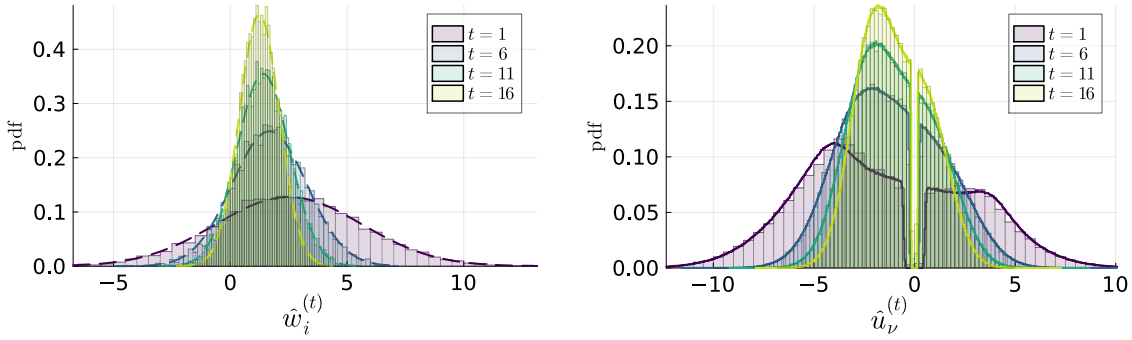
$$\mathbb{E}_D[\frac{1}{M_U}\sum_{\nu=1}^{M_U} g_u(\{(y_\nu^{(t)}, \tilde{u}_\nu^{(t)}, u_\nu^{(t)})\}_{t=1}^T)] = \mathbb{E}_{\{y^{(t)}, \xi_{u,1}^{(t)}, \xi_{u,2}^{(t)}\}_{t=1}^T}\left[g_u\left(\{(y^{(t)}, \tilde{h}_u^{(t)}, \hat{\mathsf{u}}^{(t)} + h_u^{(t)})\}_{t=1}^T\right)\right].$$

Here g_u is arbitrary as long as the above integral is convergent.

This prediction indicates that $\hat{\mathsf{u}}^{(t)}, \tilde{h}_u^{(t)}$ and $h_u^{(t)}$ effectively describe the statistical properties of the logits, i.e., $\{\hat{\boldsymbol{w}}^{(t)} \cdot \boldsymbol{x}_\nu^{(t)}/\sqrt{N} + \hat{B}^{(t)}\}_{\nu=1}^{M_U} \overset{\mathrm{d}}{=} \hat{\mathsf{u}}^{(t)} + h_u^{(t)}$, similarly to the finite-dimensional description of the weight vector in Predictions 1 and 2. The minimization problem (18), which determines the value of $\hat{\mathsf{u}}^{(t)}$, can be interpreted as an effective surrogate to determine the distribution of the logits. Schematically, it can be represented as follows:

$$\begin{aligned}
\tilde{h}_u^{(t)} &= \underbrace{B^{(t-1)}}_{\text{bias}} + \underbrace{(2y^{(t)} - 1)m^{(t-1)}}_{\substack{\text{correlation with} \\ \text{the cluster center}}} + \underbrace{\sqrt{q^{(t-1)}}\xi_{u,1}^{(t)}}_{\substack{\text{fluctuation in} \\ \text{input points}}}, \\
h_u^{(t)} &= \underbrace{B^{(t)}}_{\text{bias}} + \underbrace{(2y^{(t)} - 1)m^{(t)}}_{\substack{\text{correlation with} \\ \text{the cluster center}}} + \underbrace{\frac{R^{(t)}}{\sqrt{q^{(t-1)}}}\xi_{u,1}^{(t)}}_{\substack{\text{correlation with} \\ \text{the prediction}}} + \underbrace{\sqrt{q^{(t)} - \frac{(R^{(t)})^2}{q^{(t-1)}}}\xi_{u,2}^{(t)}}_{\substack{\text{fluctuation in} \\ \text{input points}} + \substack{\text{noise induced at a} \\ \text{parameter update}}}.
\end{aligned} \tag{35}$$

Here, $\Theta^{(t)}$, $\xi_{u,1}^{(t)}$ and $\xi_{u,2}^{(t)}$ have clear interpretations. First, $B^{(t-1)}$ and $B^{(t)}$ represent the bias terms at step $t-1$ and t. Second, $(2y^{(t)} - 1)m^{(t-1)}$ and $(2y^{(t)} - 1)m^{(t)}$ represent correlation with the cluster center at step t. Thirdly, the term $\sqrt{q^{(t-1)}}\xi_{u,1}^{(t)}$ represents the fluctuation of



(a) $\{\hat{w}_i^{(t)}\} : (N, \rho) = (2^{13}, 0.4)$ (b) $\{\hat{u}_\nu^{(t)}\} : (N, \rho) = (2^{13}, 0.4)$

Figure 1: Representative comparison of the empirical distributions and the RS predictions for the learned weights and logits. The panels show the label-imbalanced case $\rho = 0.4$ with $N = 2^{13}$ and $T = 16$. The histograms are obtained from finite-size numerical experiments, and the solid lines show the RS predictions.

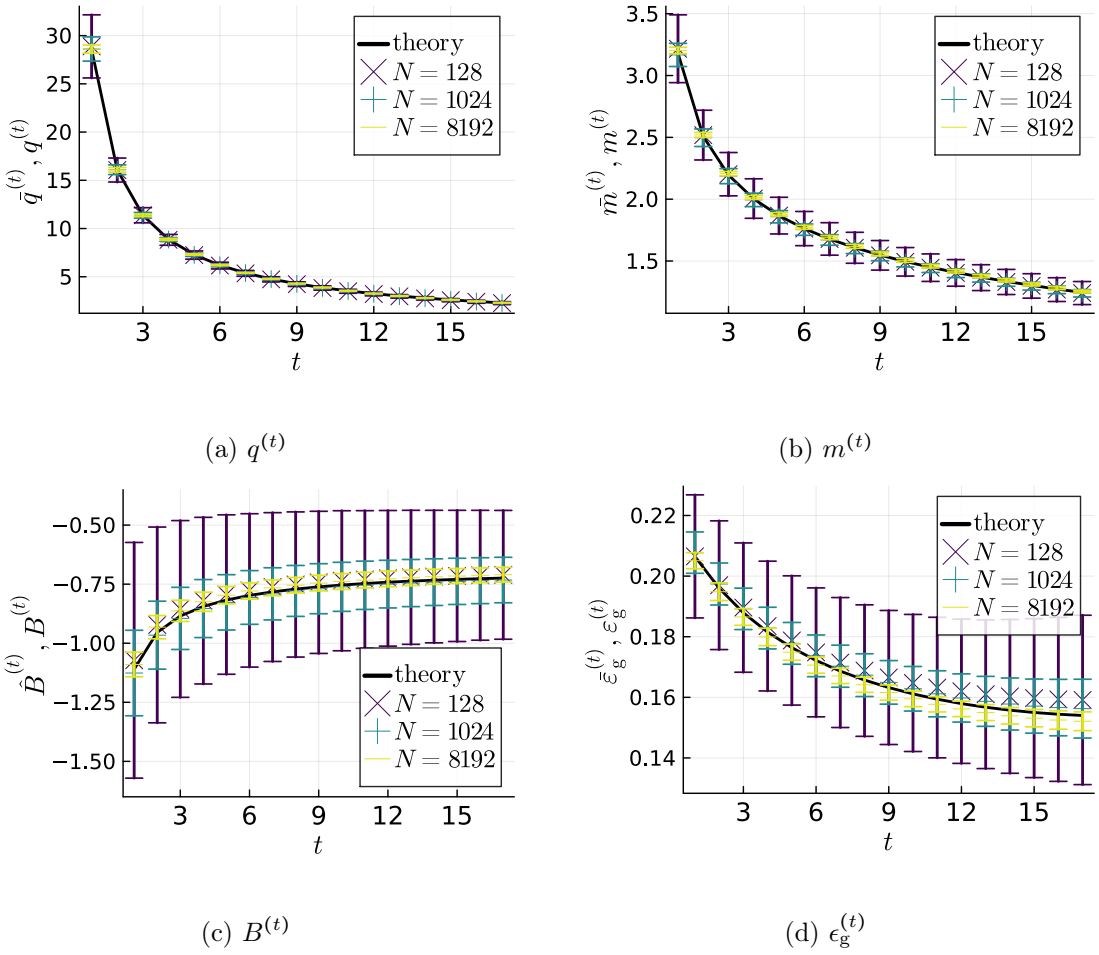
the prediction. This indicates that the Gaussian random variable $\xi_{u,1}^{(t)}$ comes from the statistical variation of the input points of the training data used at step t. Also $R^{(t)}/\sqrt{q^{(t-1)}}\xi_{u,1}^{(t)}$ in $h_u^{(t)}$ represents the correlation with the prediction. Finally, $\sqrt{q^{(t)} - (R^{(t)})^2/q^{(t-1)}}\xi_{u,2}^{(t)}$ contains two kinds of fluctuations; (i) the fluctuation of the inputs and (ii) the noise that arises when updating the parameters $\boldsymbol{\theta}$ using a finite amount of training data in the sense that the ratio to the input dimension N is finite. Note that the equations (29) and (31) imply $q^{(t)} = R^{(t)} = q^{(t-1)}$ when $\hat{\boldsymbol{\theta}}^{(t)} = \hat{\boldsymbol{\theta}}^{(t-1)}$. Hence the factor $\sqrt{q^{(t)} - (R^{(t)})^2/q^{(t-1)}}\xi_{u,2}^{(t)}$ makes no contribution when there is no update. By using these quantities, the logit $\hat{\mathrm{u}}^{(t)} + h_u^{(t)}$ is determined by the one dimensional optimization problem (18).

The predictions 1-4 are the first main results of this paper.

3.4 Cross-Checking with Numerical Experiments

To check the validity of the predictions presented in the previous section, we briefly compare the result of numerical solution of the self-consistent equations with numerical experiments of finite-size systems. The details of the numerical treatment of the self-consistent equations are described in Appendix G.

For simplicity, we investigate the case without domain shift, i.e., $\rho_U = \rho_L = \rho$ and $\Delta_L = \Delta_U = \Delta$. Also, we focus on the case of $\lambda^{(0)} = \lambda_L, \lambda^{(t)} = \lambda_U = \text{Const.}$ for $t \geq 1$ because tuning the regularization parameters in $\mathbb{R}^{T+1}$ is computationally demanding. The nonlinear function of the model, the pseudo-labeler, and the loss function are $\sigma(x) = \sigma_{\mathrm{pl}}(x) = 1/(1 + e^{-x})$ (sigmoid) and $l(p, q) = l_{\mathrm{pl}}(p, q) = -p \log q - (1 - p) \log(1 - q)$ (logistic loss), respectively. We optimize the regularization parameter λ_L, λ_U and Γ so that the generalization error $\epsilon_{\mathrm{g}}^{(T)}$ (the generalization error at the last step) is minimized by using the Nelder-Mead method in Optim.jl library (Mogensen and Riseth, 2018). Here, the total number of iterations is fixed as $T = 16$, and the comparison at steps $t = 1, 6, 11, 16$ are shown. We remark that for



(a) $q^{(t)}$ (b) $m^{(t)}$

(c) $B^{(t)}$ (d) $\epsilon_g^{(t)}$

Figure 2: Representative comparison of macroscopic quantities obtained from finite-size experiments and the RS predictions in the label-imbalanced case $\rho = 0.4$. The markers with error bars show finite-size experiments, and the solid lines show the RS predictions.

small iteration cases with almost balanced clusters, such as $T = 1$ and $\rho \simeq 1/2$, the optimal regularization parameter often shows a diverging tendency as in the logistic regression (Dobriban and Wager, 2018; Mignacco et al., 2020a) where infinitely large regularization yields the Bayes-optimal classifier. However, it is known that such a large regularization parameter induces a pathologically large finite-size effect as reported in (Mignacco et al., 2020a). Therefore, we restrict the range of the regularization parameter as $\lambda_L, \lambda_U \in (0, 0.1)$. See Appendix H for more detail.

Figures 1 and 2 show representative comparisons between the RS predictions and finite-size numerical experiments. We use the label-imbalanced setting $\rho = 0.4$ as a representative nontrivial case. When showing the histogram for logits, the contribution from these points are excluded from the figures, since the logits on the data points excluded by PLS are not determined. Additional comparisons for other values of ρ and system sizes are provided in Appendix B.

Overall, the RS solution agrees well with the results of numerical experiments on finite size systems. Therefore, we expect that the RS solution exactly describes the statistical properties of the regressors obtained by ST. In the following section, we will use the RS solution to perform a more comprehensive analysis of the behavior of ST.

4. Analysis of RS Solution

In this section, using the RS solution obtained in the previous section, we study what specific properties of classifiers can be obtained.

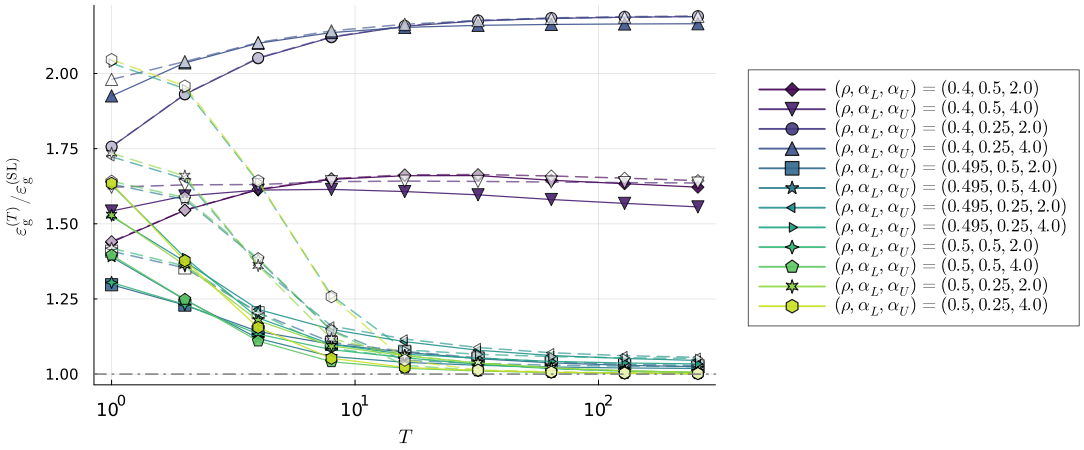
First, in subsection 4.1, to check the quantitative tendency, we numerically investigate how the generalization error depends on the label bias, the size of the unlabeled data used at each iteration, and the regularization parameter. Next, in subsection 4.2, based on the finding of the numerical inspection, we analyze the behavior of ST at weak regularization limit $\lambda_U \ll 1$ in the under-parameterized setting. There, we find that the classification plane finds the best direction as long as the initial classifier is not completely uninformative, but the bias term may not be optimal in label imbalanced cases.

4.1 Numerical Inspection

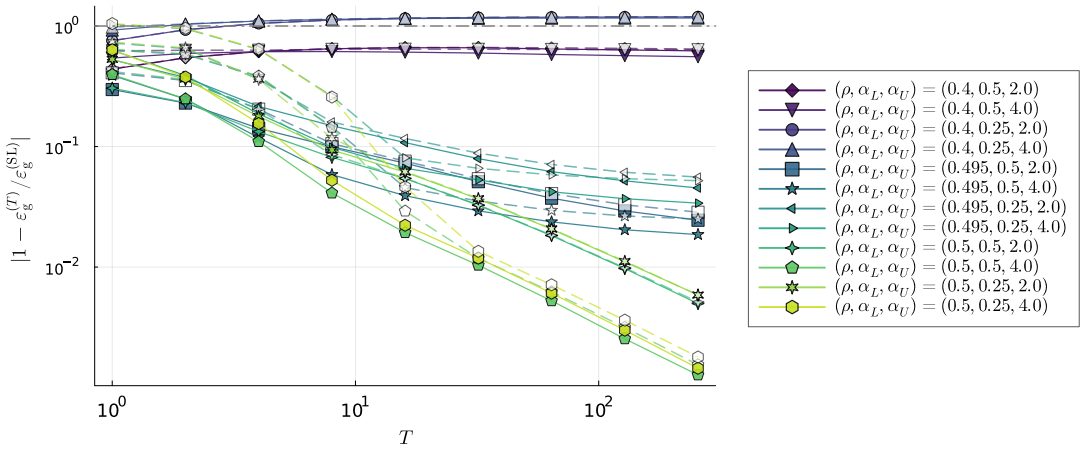
Setup: To check quantitative tendency, we investigate the generalization error and macroscopic quantities in more detail. As in the previous subsection, we investigate the case $\rho_U = \rho_L = \rho$ and $\Delta_L = \Delta_U = \Delta$ (no domain shift). Also, we focus on the case of $\lambda^{(0)} = \lambda_L, \lambda^{(t)} = \lambda_U = \text{Const.}$ for $t \geq 1$ because tuning the regularization parameters in $\mathbb{R}^{T+1}$ is computationally demanding. The nonlinear function of the model, the pseudo-labeler, and the loss function are $\sigma(x) = \sigma_{\text{pl}}(x) = 1/(1+e^{-x})$ (sigmoid) and $l(p,q) = l_{\text{pl}}(p,q) = -p\log q - (1-p)\log(1-q)$ (logistic loss), respectively. Also, the hyper parameters are optimized so that the generalization at the last step T is minimized. In the following, we mainly focus on how the various quantities vary as the total number of iterations T changes. Recall that $t \in [T]$ represents the iteration step of ST with the total number of iterations T, while T represents the total number of iterations in ST.

Dependence on the label bias: Figure 3 shows the ratio of the generalization error obtained by ST (33) to the SL (ridge-regularized logistic regression) with a labeled dataset of size $N(\alpha_L + \alpha_U \times T)$. The generalization error of SL is obtained by the formula in (Mignacco et al., 2020a). When the label imbalance is absent ($\rho = 1/2$) and the total number of iterations T is large, the performance of ST is almost compatible with the SL with the same data size but with the true label. This observation is compatible with previous studies (Oymak and Gulcu, 2020, 2021; Frei et al., 2022). On the other hand, when there is a label imbalance, the performance of ST does not approach that of SL, even if we optimize the regularization parameter and the threshold of PLS. This point is more apparent in Figure 3b. Although PLS improves the generalization error, especially when the total number of iterations are small (see Figure 5a), ST's performance still significantly degrades as the label bias grows. Similarly, Figure 4 shows the cosine similarity (34) between the direction of the classification plane $\hat{\boldsymbol{w}}^{(t)}$ and the cluster center $\boldsymbol{v}$. The cosine similarity monotonically increases as the number of the total iteration grows. However, the growth is slow in label-biased cases.

Effect of PLS: Figure 5 shows the comparison of the generalization error between the case with and without PLS. The upper panel shows the result of both label balanced and imbal-



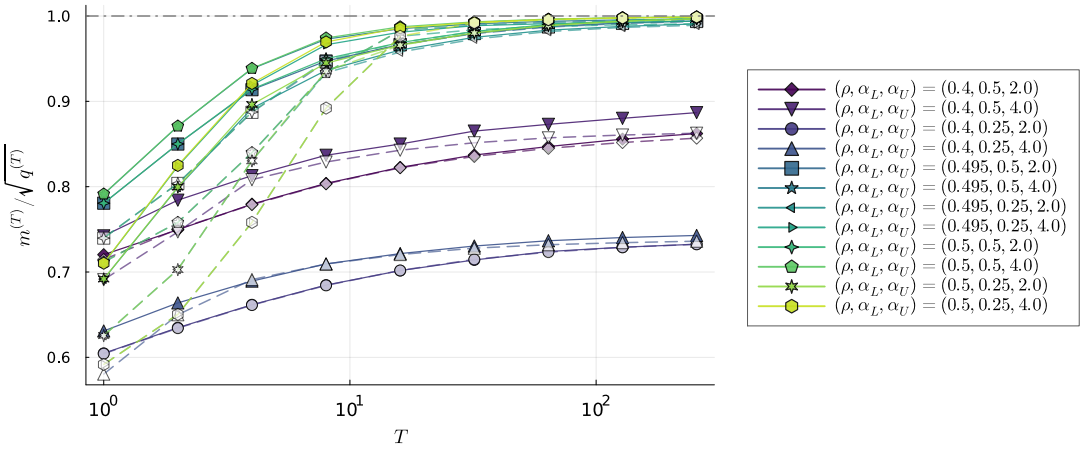
(a) $\epsilon_{\mathrm{g}}^{(T)}/\epsilon_{\mathrm{g}}^{(\mathrm{SL})}$ (relative generalization error)



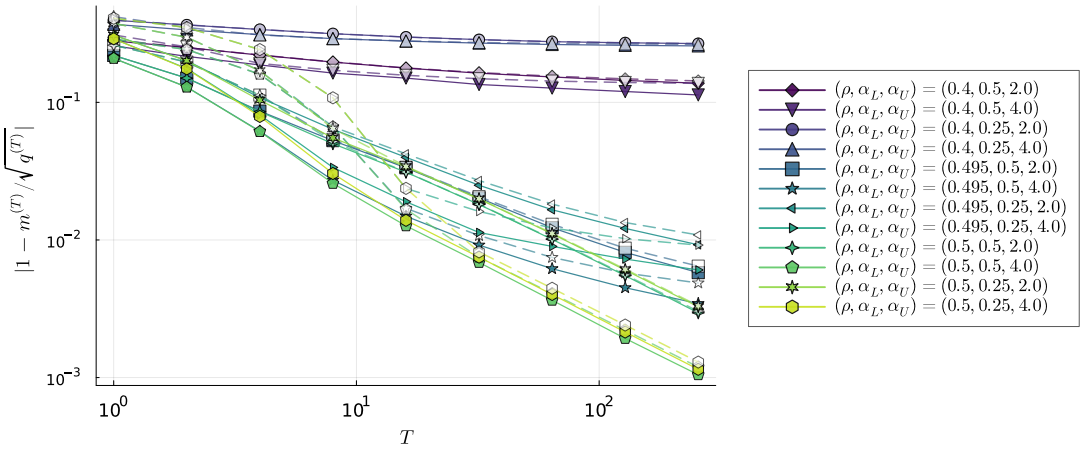
(b) $|1 - \epsilon_{\mathrm{g}}^{(T)}/\epsilon_{\mathrm{g}}^{(\mathrm{SL})}|$ (absolute value of deviation from unity)

Figure 3: The ratio of the generalization error obtained at the end of ST (33) ($t = T$) to the SL with a labeled dataset of size $N(\alpha_L + \alpha_U \times T)$. When the ratio is unity, the ST with an unlabeled dataset has the same performance as the SL with labeled data of the same size. Different colors and lines represent different pairs of (ρ, Δ). The filled markers with solid lines represent the result with PLS, where Γ is optimized so that $\epsilon_{\mathrm{g}}^{(T)}$ is minimized. The white markers with dashed lines represent the result without PLS ($\Gamma = 0$). The upper panel shows the raw values. The lower panel shows their absolute values of the deviation from unity in the log scale.

anced cases. When the total number of iterations T is small, ST's performance is largely improved, which suggests that minimizing the cost function (4) has the intuitive effect of fitting to the pseudo-correct-*label*, as originally proposed in (Lee et al., 2013). Hence, just fitting the model to relatively reliable labels is appropriate. Recall that the pseudo-labels selected by PLS have relatively large input and closer to hard labels. This tendency is evident especially when labels are balanced ($\rho = 1/2$). Detailed data for the label balanced



(a) $m^{(t)}/\sqrt{q^{(T)}}$ (raw value)

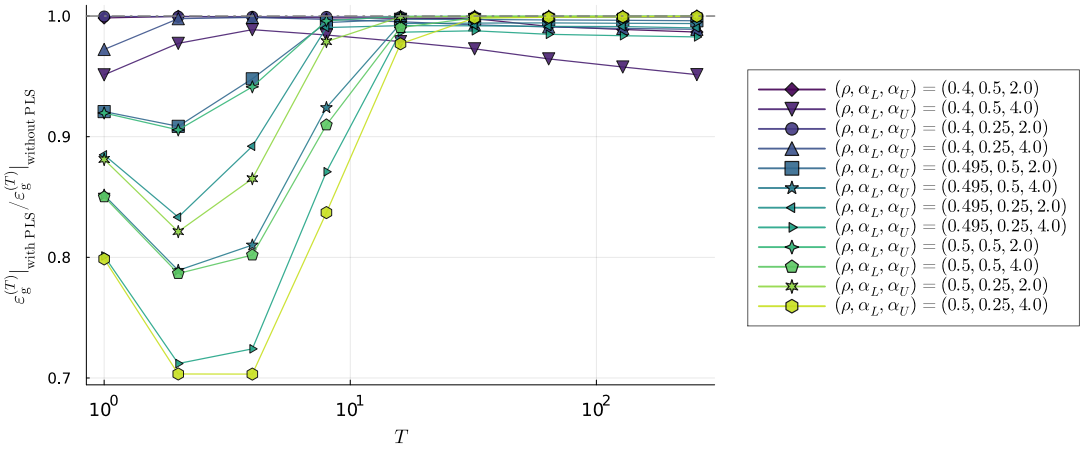


(b) $|1 - m^{(t)}/\sqrt{q^{(T)}}|$ (absolute value of deviation from unity)

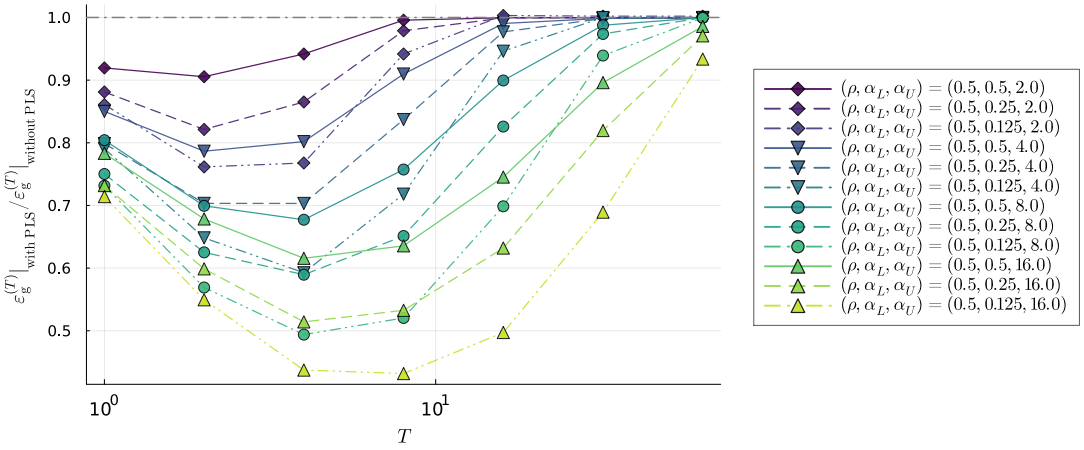
Figure 4: The cosine similarity (34) obtained at the end of ST (33) ($t = T$). When the value is unity, the classification plane is oriented in the optimal direction. Different colors and lines represent different pairs of $(\rho, \alpha_L, \alpha_U)$. The filled markers with solid lines represent the result with PLS, where Γ is optimized so that $\epsilon_{\rm g}^{(T)}$ is minimized. The white markers with dashed lines represent the result without PLS ($\Gamma = 0$). The upper panel shows the raw values. The lower panel shows their absolute values of the deviation from unity in the log scale.

cases is shown in Figure 5b. As shown in this figure, PLS is effective especially when the initial classifier is poor (α_L is small) and the size of the unlabeled data used at each iteration is large (α_U is large).

Dependence on the size of unlabeled data: Figure 6 shows α_U dependence of relative generalization error when $(\rho, \Delta) = (1/2, 0.75^2)$. As already observed, PLS drastically reduces the generalization error when the total number of iterations is small. However, as T grows, the effect of PLS becomes minor and the decrease of the generalization error gets slower,



(a) Including label imbalanced cases



(b) Only label balanced cases

Figure 5: Comparison of generalization error with and without the implementation of PLS. Upper panel shows the result including label imbalanced cases. Lower panel shows the result of label balanced cases only.

indicating that the nature of self-training gradually changes at large T. Also, as the size of the unlabeled data used at each iteration α_U decreases, the approach to the supervised loss becomes slower, and the performance is catastrophically poor when α_U is smaller than some threshold. This instability at small α_U can also be seen from the linear susceptibility. See Appendix C for more detail.

Optimal hyper parameters: To obtain further insight on ST when α_U is moderately large so that long iteration of ST yields a better generalization, we show the optimal regularization parameter λ_U^* in Figure 7. The figure indicates that the optimal regularization strength decreases in a power law at $T \gg 1$. Combined with the observations in Figure 6, it is suggested that ST can find a good classifier by accumulating small parameter updates when a large number of iterations can be used. Recall that, in our setting, ST returns the same

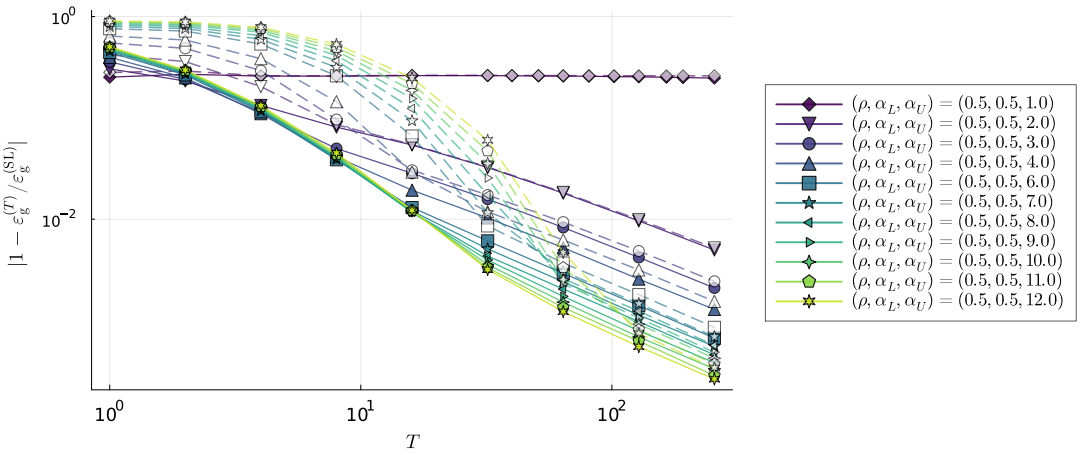


Figure 6: Relative generalization error $\epsilon_g^{(T)}/\epsilon_g^{(ST)}$ when $(\rho, \Delta) = (1/2, 0.75^2)$. Different colors represent different α_U.

parameters obtained in the previous step, i.e., $\hat{\boldsymbol{\theta}}^{(t)} = \hat{\boldsymbol{\theta}}^{(t-1)}$ when the training at each step of ST is in an underparametrized setup and the regularization is removed ($\lambda_U = 0$). We also remark that this tendency does not depend on the label bias ρ.

Summary of numerical observations: The main numerical observations are summarized as follows.

1. **Small number of iterations.** When the total number of iterations T is small, using pseudo-labels with high confidence, which also means that such pseudo-labels are close to hard labels, is effective. These observations imply that the optimal ST, whose hyperparameters are optimized, fits the model to relatively reliable pseudo-labels. In this sense, the pseudo-label plays the intuitive role of pseudo-*labels*, as originally termed in (Lee et al., 2013).

2. **Large number of iterations.** When T is large and α_U is moderately large, the optimal regularization parameter λ_U^* becomes small. Thus, the difference in the model parameters between successive time steps is small due to the pseudo-label loss, and ST improves the classifier by accumulating small parameter updates. In this regime, the pseudo-label loss is better interpreted as maintaining continuity with the previous iterate, rather than simply fitting noisy labels.

3. **Effect of label imbalance.** The large-T behavior is qualitatively similar for different values of the label bias ρ, but the final performance relative to supervised learning depends strongly on the imbalance. ST is almost comparable to supervised learning when the label bias is small, whereas it remains suboptimal when the label imbalance is large.

To deepen our understanding of the above observation, we will consider a perturbative analysis in the next subsection.

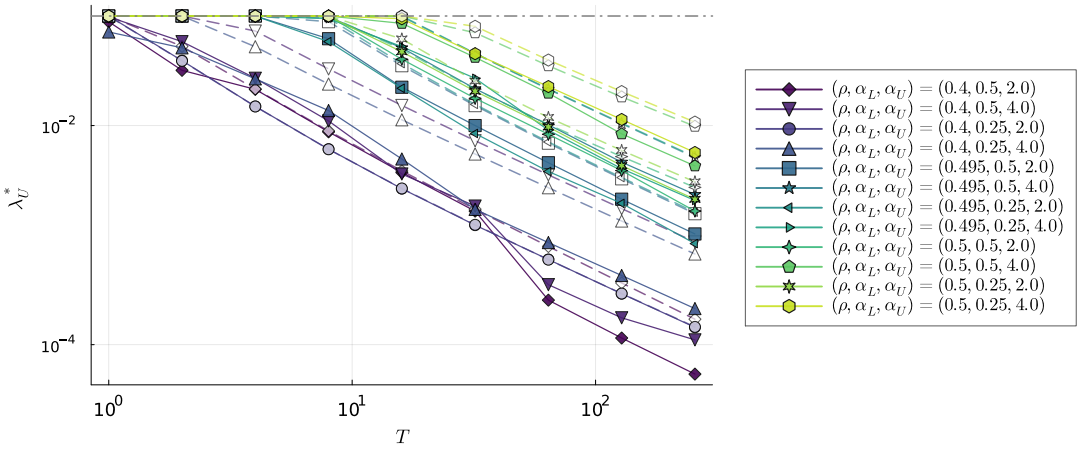


Figure 7: Optimal regularization parameter λ_U^*. Its dependence on the total number of iterations T is shown. Since λ_U is upper bounded during the optimization, there is a plateau at $\lambda_U = 0.1$.

4.2 Small Regularization Limit: Perturbative Analysis

According to the observations made in the previous subsection, to obtain a good classifier with long iterations of ST, the best strategy is to accumulate small parameter updates by using small regularizations and underparametrized setups at each iteration. To gain a deeper understanding of the behavior of ST when $T \gg 1$ and $\lambda_U \ll 1$, we consider a perturbative analysis. This is a formal perturbative analysis of the self-consistent equations predicted in Section 3.3. Accordingly, Predictions 5, 6, and 8 are conditional on Predictions 1–4 and on the assumptions stated below. In particular, the existence and uniform validity of the expansion in λ_U and the continuum limit are not established in a rigorous sense. Proposition 7 is a direct consequence of Prediction 6.

Specifically, the variables $\Theta^{(t)}$ and $\hat{\Theta}^{(t)}$ in Definition 1 are expanded in terms of λ_U like $q^{(t)} = q_0^{(t)} + q_1^{(t)}\lambda_U + \dots$. These expansion coefficients are then determined at each order of λ_U in a self-consistent manner. Under appropriate assumptions, the variables at step t are equal to those of the previous steps at $\lambda_U = 0$: $(\Theta^{(t)}, \hat{\Theta}^{(t)})|_{\lambda_U=0} = (\Theta^{(t-1)}, \hat{\Theta}^{(t-1)})$. Hence, one can expect that the expansion coefficients at the first order determine the evolution of these parameters at $\lambda_U \to 0$ like $(q^{(t)} - q^{(t-1)})/\lambda_U \to q_1^{(t)}$. In other words, by taking λ_U as the time unit and $\tilde{t} = \lambda_U \times t$ as the continuous time, it can determine the continuous-time dynamics at $\lambda_U \to 0$ as $\frac{dq_{\tilde{t}}}{d\tilde{t}} = q_{\tilde{t}}$, where $q_{\tilde{t}}$ is the value of $q^{(t)}$ in this continuous time scale. In this context, we also refer to the limit $\lambda_U \to 0$ as the *continuum limit*.

We use the following two assumptions for the perturbative analysis. The first assumption is about the behavior of the loss function.

Assumption 1 *Let $\tilde{l}_{\tilde{\Gamma}}(y, x) = \mathbb{1}(|y| > \tilde{\Gamma}) l_{\mathrm{pl}}(\sigma_{\mathrm{pl}}(y), \sigma(x))$. Also, let $\hat{u}$ be the solution of the following equation with a positive constant $\tilde{c} > 0$:*

$$\hat{u}(y, x) = \arg\min_u \frac{u^2}{2\tilde{c}} + \tilde{l}_{\tilde{\Gamma}}(y, x + u).$$

Then the following holds for $\hat{u}(y,x)$ and $\tilde{l}$.

1. *$\hat{u}(y,x)$ is unique.*
2. *The derivative of the loss function is zero when $y = x$:*

$$\frac{\partial}{\partial x}\tilde{l}_{\tilde{\Gamma}}(y,x)\Big|_{y=x} = 0, \tag{36}$$

This is a natural condition for moderately regular loss function such as the squared loss $l_{\mathrm{pl}}(\sigma_{\mathrm{pl}}(y),\sigma(x)) = (y-x)^2/2$ or the cross-entropy loss with the soft-sigmoid pseudo-label $l_{\mathrm{pl}}(\sigma_{\mathrm{pl}}(y),\sigma(x)) = -\sigma(y)\log\sigma(x) - (1-\sigma(y))\log(1-\sigma(x))$.

The other assumption is the following:

Assumption 2 *At $\lambda_U = 0$, ST returns the same parameter with the previous step:*

$$\hat{\boldsymbol{\theta}}^{(t)}|_{\lambda_U=0} = \hat{\boldsymbol{\theta}}^{(t-1)}. \tag{37}$$

Also the solution of the self-consistent equation can be series expanded in λ_U:

$$\Theta^{(t)} = \Theta_0^{(t)} + \Theta_1^{(t)}\lambda_U + \dots, \tag{38}$$

$$\hat{\Theta}^{(t)} = \hat{\Theta}_0^{(t)} + \hat{\Theta}_1^{(t)}\lambda_U + \dots, \tag{39}$$

which should be interpreted as component-wise expansion. For example, $q^{(t)} = q_0^{(t)} + q_1^{(t)}\lambda_U + \dots, m^{(t)} = m_0^{(t)} + m_1^{(t)}\lambda_U + \dots,$ and so on.

The equation (37) is expected to hold naturally when the following conditions are used; (i) moderately large α_U, i.e., underparametrized setup, (ii) monotonically increasing nonlinear functions $\sigma_{\mathrm{pl}} = \sigma$, and (iii) moderately regular loss functions, such as the squared loss and the cross-entropy loss. The latter equations (38)-(39) are the technical assumption.

Under the Assumptions 1 and 2, we first obtain the following result by analyzing the saddle point conditions except the condition (26) that determines the update of $\hat{B}^{(t)}$:

Prediction 5 *The lowest order of the expansion of $\hat{\chi}^{(t)}$ is at the second order:*

$$\hat{\chi}^{(t)} = \mathcal{O}(\lambda_U^2).$$

Also, the lowest order of the expansion of $q^{(t)} - (R^{(t)})^2/q^{(t-1)}$ is at the second order:

$$q^{(t)} - \frac{(R^{(t)})^2}{q^{(t-1)}} = \mathcal{O}(\lambda_U^2).$$

See Appendix D for the derivation. The consequence of this result is that the contributions from the noise terms $\sqrt{\hat{\chi}^{(t)}}\xi_w^{(t)}$ in (28) and $\sqrt{q^{(t)} - \frac{(R^{(t)})^2}{q^{(t-1)}}}\xi_{u,2}^{(t)}$ in (35) are higher order in $\lambda_U^{(t)}$ because these terms appear only in the form of a square (second order in $\lambda_U^{(t)}$) or the raw average (equals to zero). This implies that the contributions of the noise terms are negligible at the leading order in the weight update of ST in this small regularization setup. Hence, ST can extract information from the data in an almost noiseless way at least when updating the direction of the classification plane. This intuition is further supported by the following predictions.

For the square of the cosine similarity (34), we obtain the following.

Prediction 6 *Let $M^{(t)}$ be the square of the cosine similarity (34): $M^{(t)} = (m^{(t)})^2/q^{(t)}$. Then, at the first order of expansion, it obeys the following equation:*

$$M^{(t)} = M^{(t-1)} + C^{(t)} M^{(t-1)}(1 - M^{(t-1)})\lambda_U + \mathcal{O}(\lambda_U^2),$$
$$C^{(t)} = \frac{2}{\hat{Q}_0^{(t)}} \frac{\hat{m}_1^{(t)}}{m^{(t-1)}},$$

Also,

$$\hat{m}_0^{(t)} = 0,$$

which shows that the zeroth-order of the series expansion of $\hat{m}^{(t)}$ is zero. Moreover, $\hat{Q}_0^{(t)} > 0$.

Equivalently, it can be rephrased as follows. Let $(M_{\tilde{t}}, C_{\tilde{t}})$ be the value of $(M^{(t)}, C^{(t)})$ in the continuous time scale. Then, $M_{\tilde{t}}$ follows the following differential equation at the continuum limit:

$$\frac{dM_{\tilde{t}}}{d\tilde{t}} = C_{\tilde{t}} M_{\tilde{t}}(1 - M_{\tilde{t}}), \quad C_{\tilde{t}} > 0,$$

See Appendix E for the derivation of this prediction. Remarkably, the saddle point condition (26) that determines the update of $\hat{B}^{(t)}$ is not used to derive Prediction 6, which indicates that formally same expression for $M_1^{(t)}$ is obtained even if $B^{(t)}$ is kept constant.

Since $\hat{m}_1^{(t)}$ is the leading order of $\hat{m}^{(t)}$ that represents the signal component of $\hat{\mathsf{w}}^{(t)}$ accumulated at step t as depicted in (28), $C^{(t)}$ is positive if the training yields a positive accumulation of the signal component to $\hat{\mathsf{w}}^{(t)}$ when $m^{(t-1)} = \boldsymbol{v} \cdot \hat{\boldsymbol{w}}^{(t-1)}/N > 0$, i.e., the classification plane at step $t-1$ is positively correlated with the cluster center, which seems to naturally hold when using a legitimate loss function. We remark that, in general, due to the presence of the noise term $\sqrt{\hat{\chi}^{(t)}}\xi_w^{(t)}$, an increase in cosine similarity is not guaranteed even if $\hat{m}^{(t)} > 0$.

In the following, we assume that $\hat{m}^{(t)}$ is positive at the leading order in λ_U.

Assumption 3 *The quantity $\hat{m}^{(t)}$ is positive at the leading order of λ_U, i.e.,*

$$\hat{m}_1^{(t)} > 0.$$

Under this assumption, $C^{(t)}$ or $C_{\tilde{t}}$ is positive.

When the time-dependent variable $C_{\tilde{t}}$ is positive, the fixed point is only $M_{\tilde{t}} = 0$ or $M_{\tilde{t}} = 1$. Furthermore, as long as the initial classifier is informative, i.e., $M_0 > 0$, which is naturally expected by the supervised learning at the initial stage, the classification plane is oriented to the optimal direction at the long time limit. Hence, we can expect that the classification plane will find the best direction by accumulating small updates of ST. This is summarized as follows.

Proposition 7 *Under Assumption 3, at the long time limit $\tilde{t} = \lambda_U \times t \to \infty$, the classification plane is oriented to the best direction if the initial classifier is informative $M^{(0)} > 0$:*

$$\lim_{\tilde{t}\to\infty} M_{\tilde{t}} = 1.$$

The Prediction 5 and Proposition 7 are the second main results of this paper. We remark that the above result is valid even if $\rho^{(t)}$ or $\Delta^{(t)}$ are time-dependent, although such time-dependence affects the value of $C_{\tilde{t}}$ and $C^{(t)}$. See the derivations in Appendix D-E.

Proposition 7 is a consequence of the noiseless accumulation of information shown in Predictions 5 and 6. It corresponds to the message in the introduction that, when the number of iterations is large, ST improves the model by accumulating small parameter updates with a small regularization parameter. This supports the numerical observation that says weak regularization is better in large T regime.

We note, however, that Proposition 7 concerns the direction of the classification plane only. As we discuss below, the generalization error is not necessarily optimal in the label-imbalanced case, even when the direction is aligned, because it also depends on the balance between the norm of the weight vector and the magnitude of the bias.

Notably, we can obtain a closed form solution when considering the squared loss without PLS: $\tilde{l}_{\tilde{\Gamma}}(y,x) = \frac{1}{2}(y-x)^2$, which can be a legitimate loss function in this Gaussian mixture setting (Mignacco et al., 2020a; Oymak and Gulcu, 2020, 2021). In this case, we can obtain the following result:

Prediction 8 *Let $V_U = 4\rho_U(1-\rho_U)$ be the variance of the true label. When $\tilde{l}_{\tilde{\Gamma}}(y,x) = \frac{1}{2}(y-x)^2$, the following hold at the continuum limit $\lambda_U \to 0$ under Assumption 2:*

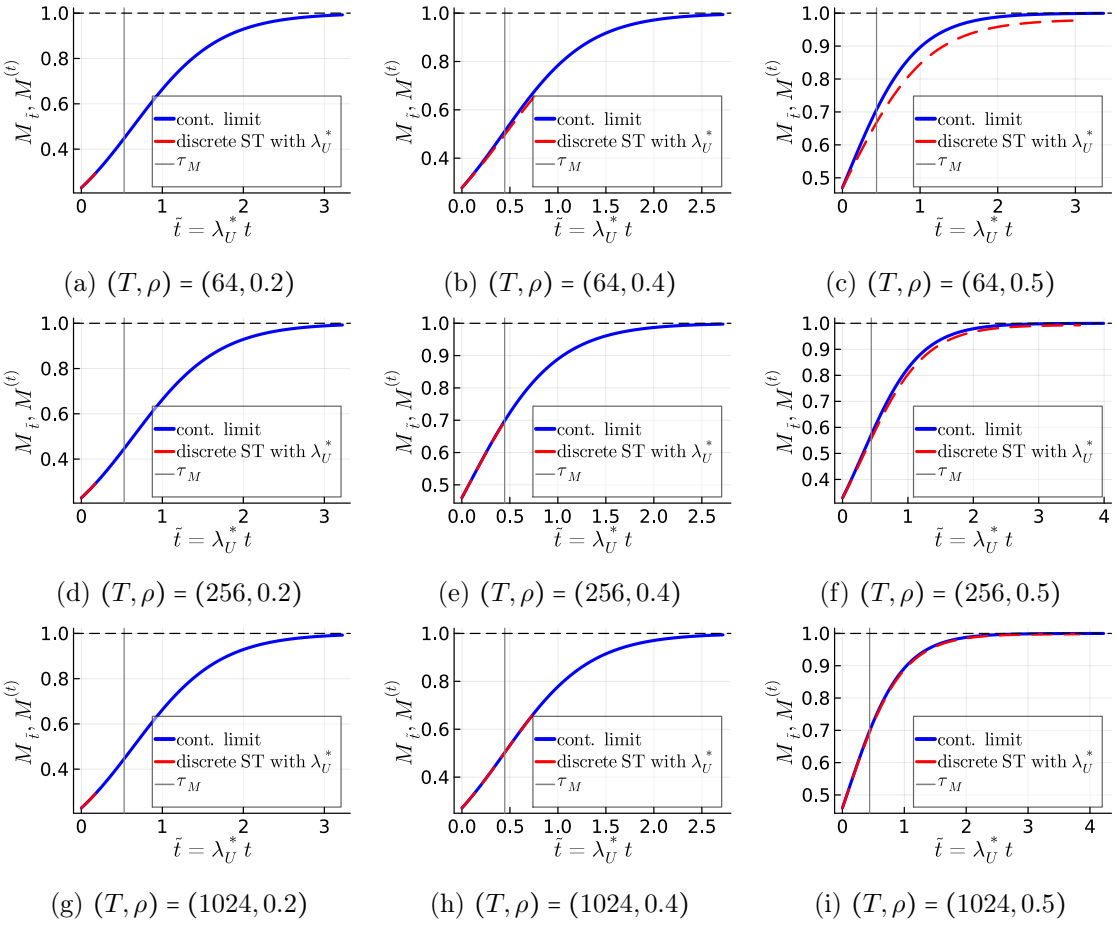
$$
\begin{aligned}
M_{\tilde{t}} &= \frac{1}{1+(1/M_0-1)e^{-\tilde{t}/\tau_M}}, \\
m_{\tilde{t}} &= m_0 e^{-\frac{\tilde{t}}{\tau_m}} \\
B_{\tilde{t}} &= B_0 + (2\rho_U - 1)m_0(1-e^{-\frac{\tilde{t}}{\tau_m}}), \\
\tau_M &= \frac{1}{2}\frac{\Delta}{V_U}(\alpha_U - 1)(\Delta + V_U), \\
\tau_m &= (\alpha_U - 1)(\Delta_U + V_U),
\end{aligned}
\tag{40}
$$

where M_0, m_0, B_0 are the initial conditions.

See Appendix F for the derivation of this prediction. As expected, the cosine similarity monotonically increases. In particular, the deviation of $M_{\tilde{t}}$ from unity is decreasing in an exponential manner. Moreover, this result is valid as long as $\alpha_U > 1$. Otherwise the time scales τ_m, τ_M are negative, causing the divergence of M_t, m_t. This indicates that the theory is not self-consistent. Recall that the value $\alpha_U = 1$ is exactly the point that separates the under-/over-parameterized setup in this squared loss case. This perturbative analysis supports the numerical observation that a batch size α_U that is too small may not improve performance, implying that other special care is needed in overparameterized cases.

4.2.1 Implications for the Label-Imbalanced Cases

From Predictions 5, 6, 8 and Proposition 7, we can expect that the classification plane is pointing in the best direction as the number of iterations grows with moderately small regularization parameter λ_U. However, as we saw in Figure 3 and 4, it is not the case in label-biased cases even after hundreds of iterations, with practical values of λ_U^*.



(a) $(T, \rho) = (64, 0.2)$ (b) $(T, \rho) = (64, 0.4)$ (c) $(T, \rho) = (64, 0.5)$

(d) $(T, \rho) = (256, 0.2)$ (e) $(T, \rho) = (256, 0.4)$ (f) $(T, \rho) = (256, 0.5)$

(g) $(T, \rho) = (1024, 0.2)$ (h) $(T, \rho) = (1024, 0.4)$ (i) $(T, \rho) = (1024, 0.5)$

Figure 8: Comparison of the continuous-time dynamics (40) and the actual evolution of the discrete iteration of ST. $M^{(t)}$ and $M_{\tilde{t}}$ are the square of the cosine similarity (34) in discrete and continuous time scale, respectively. The blue solid curve represents the continuous dynamics (40). The red dashed line represents the result of the actual discrete update of ST. τ_M is the time scale required for the classification plane to sufficiently align in the same direction as $\boldsymbol{v}$. $(\alpha_L, \alpha_U, \Delta) = (0.5, 2.0, 0.75^2)$.

We argue that this is due to a potential imbalance between the magnitude of the bias $\hat{B}^{(t)}$ and the norm of the weight vector $\sqrt{q^{(t)}} = \sqrt{\|\hat{\boldsymbol{w}}^{(t)}\|_2^2/N}$. As the generalization error (33) depends on the ratio between these quantities $\hat{B}^{(t)}/\sqrt{q^{(t)}}$, the generalization error may be at a random level if the ratio is significantly large. Indeed, this ratio can become large when the labels are imbalanced for the following reasons. Since the pseudo-label just ensures the continuity between the successive ST steps in our setup under Assumption 2, the norm of the weight vector gradually decreases as the iteration grows due to the shrinkage effect of the regularizer $\lambda_U \|\boldsymbol{w}^{(t)}\|_2^2$. On the other hand, the magnitude of the bias term does not necessarily decrease because it is not directly regularized.

This point is evident when $\tilde{l}_\Gamma(y,x) = (y-x)^2/2$. From the closed-form solution in Prediction 8, we see that $B_{\tilde{t}} \to B_\infty \neq 0$ while $M_{\tilde{t}} \to 1$ and $m_{\tilde{t}} \to 0$, at $\tilde{t} \to \infty$. This indicates that the norm of the weight vanishes. Hence $|B_{\tilde{t}}/\sqrt{q_{\tilde{t}}}| \to \infty$. To avoid this pathological behavior, the best regularization parameter λ_U^* tends to be very small so that $\tilde{t}$ does not diverge. In figure 8, we compare the continuous-time dynamics (40) and the actual evolution of the discrete iteration of ST. As T increases, λ_U^* decreases to justify the description by continuous-time dynamics (40). However, the elapsed time $\lambda_U^* \times t$ of the actual optimal regularization parameter λ_U^* is not sufficiently larger than τ_M when $\rho \neq 1/2$, although $M_{\tilde{t}} \simeq 1$ at $\tilde{t} \gg 1$ as expected. Note that the above continuous time argument is valid as long as the regularization parameter λ_U is small so that the second-order term of the perturbative expansion can be ignored, but it says nothing about the regularization parameter λ_U^* that is actually obtained by minimizing the generalization error in the last step of ST.

Therefore, we may need some heuristics to compensate for the shrinkage of the norm of the weight vector due to the regularization, at least in our online setup.

5. Heuristics for Label Imbalanced Cases

In the previous section, we saw that ST can find a classification plane with the optimal direction with a large number of iterations and a small regularization. However, the imbalance of the norm of the weight and the magnitude of the bias can be problematic in a label imbalanced case. To overcome the problems in label imbalanced cases, we introduce the following two heuristics.

Heuristic 1 (Pseudo-label annealing) *In the ST steps, we use the model's nonlinear function as the pseudo-labeler, but gradually amplify its input:*

$$\begin{aligned} \sigma_{\rm pl}(x) &= \sigma(\gamma^{(t)} x), \\ \gamma^{(t)} &= 1 + at \quad a > 0. \end{aligned} \tag{41}$$

We expect that the classification plane will gradually align in the optimal direction at small t because $\gamma^{(t)} \simeq 1$ at such time, which is close to the condition assumed in the Prediction 6. Later, by gradually making the pseudo-labels closer to hard labels, the norm shrinkage of the weight vector will be compensated. However, even with this heuristic, it may still be difficult to learn the bias and the weight at the same time. Therefore, we propose to fix the bias term to that of the initial classifier:

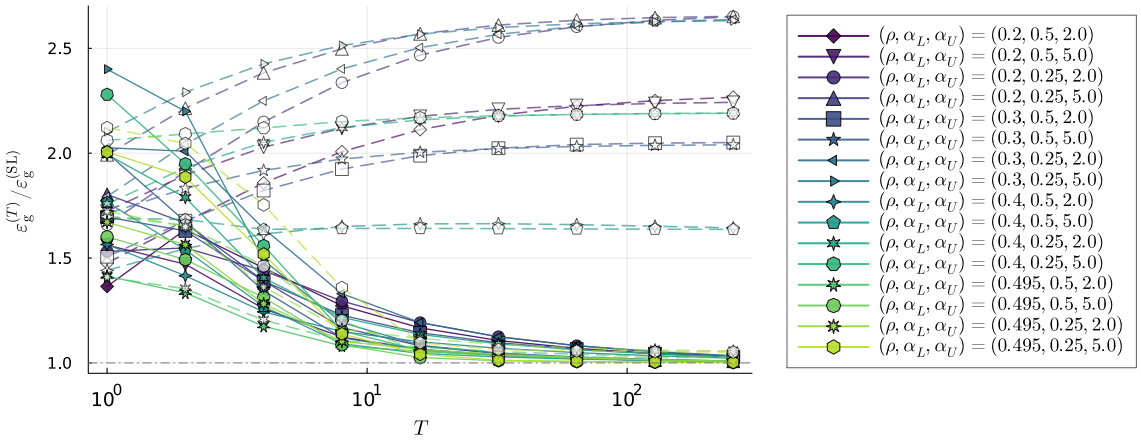
Heuristic 2 (Bias-fixing) *Fix the bias term at that obtained at the supervised learning at* $t = 0$*:*

$$\hat{B}^{(t)} = \hat{B}^{(0)}, \quad t = 1, 2, \ldots.$$

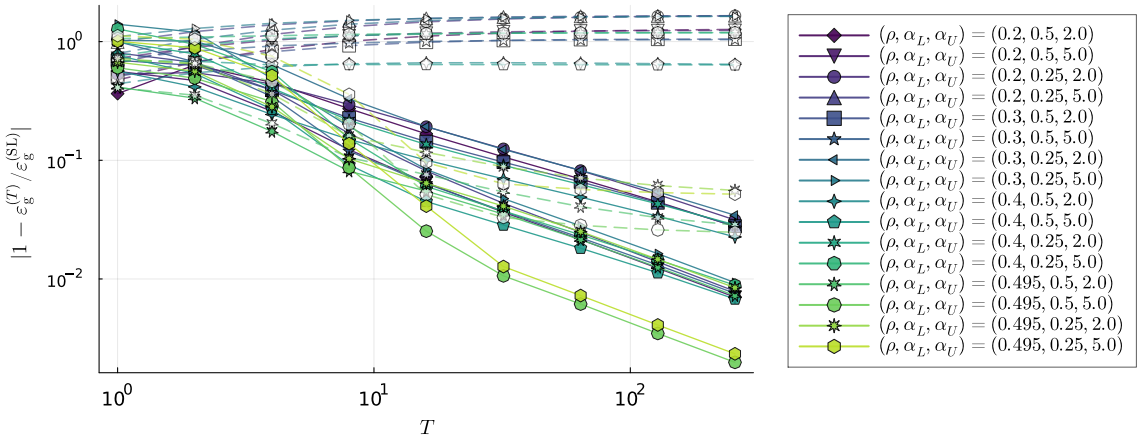
We expect the bias term of the initial classifier to be moderately good unless the labeled data is severely limited.

5.1 Demonstration

To check the effectiveness of the above heuristics, we numerically solve the self-consistent equations and investigate the generalization error. To include Heuristics 2, we remove



(a) $\epsilon_g^{(T)}/\epsilon_g^{(SL)}$ (relative generalization error)



(b) $|1 - \epsilon_g^{(T)}/\epsilon_g^{(SL)}|$ (absolute value of deviation from unity)

Figure 9: The ratio of the generalization error obtained at the end of ST ($t = T$) to the SL with a labeled dataset of size $N(\alpha_L + \alpha_U \times T)$. The cases with the Heuristics 1 and 2 are included. The white markers with dash lines represent the "naive" ST with $l(y, x) = -y \log x - (1 - y) \log(1 - x), \sigma_{\rm pl}(x) = \sigma(x) = 1/(1 + e^{-x})$. The filled markers with solid lines represent the result with the heuristics. Here, Γ is set to be zero (without PLS). The upper panel shows the raw values. The lower panel shows their absolute values of the deviation from unity in the log scale.

the equation (26) from the self-consistent equation, which corresponds to the saddle point condition with respect to the bias. Here, the loss function is the logistic loss $l(y, x) = -y \log x - (1 - y) \log(1 - x)$ and the nonlinear function of the model is the sigmoid function $\sigma(x) = 1/(1+e^{-x})$. Also, Γ is set to be zero (without PLS) since PLS did not make significant difference when $T \gg 1$ and $\rho_U = 1/2$ (see Figure 3). The annealing parameter a is optimized to minimize the generalization error in the last step of ST.

Figure 9 shows the relative generalization error as in Figure 3. It is clearly shown that the performance is significantly improved by using the heuristics. Indeed, it is shown

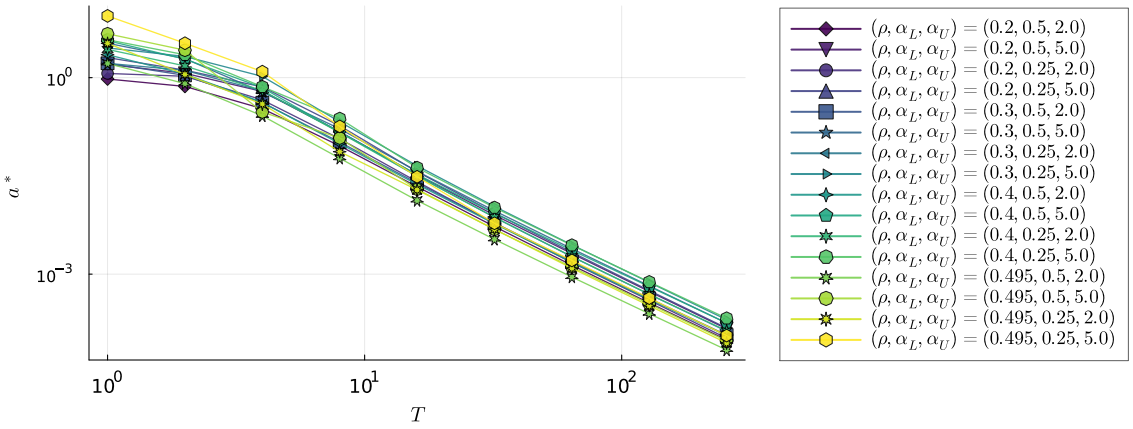


Figure 10: Optimal values of the annealing parameter a in (41). The settings are the same as in Figure 9.

that the generalization error obtained by ST with the heuristics is almost compatible with the supervised learning with true labels, even in the highly label imbalanced case $\rho = 0.2$, demonstrating the effectiveness of the proposed heuristics. Without the heuristics, the $|1 - \epsilon_g^{(T)}/\epsilon_g^{(\mathrm{ST})}|$ converges to a non-zero value even when almost label balanced case $\rho = 0.495$ as shown in Figure 9b.

Figure 10 shows the optimal values of the annealing parameter a^* in (41). As the total number of iterations T increases, this parameter gradually decreases, indicating that the pseudo-label is slowly approaching hard labels when the total number of iterations T is large. In such long iterations, the cost function (4) involving the pseudo-labels can be interpreted as playing the role of maintaining continuity with the previous step, rather than the intuitive role of fitting to a pseudo-*label* as already observed in subsection 4.1. On the other hand, when the total number of iterations T is small, its value is rather large, indicating ST fits the model to the almost hard labels. In such cases, the cost function (4) can be interpreted as playing the intuitive role of fitting the pseudo-*label*. These observations suggest that there is a crossover regarding the role of the pseudo-label as the total number of iterations changes.

Finally, we remark that Heuristic 1 and Heuristic 2 are motivated by the understanding of ST obtained in Section 4.2, which indicates the usefulness of our theoretical analysis.

6. Summary and Conclusion

In this work, we developed a theoretical framework for analyzing the iterative ST, when a new batch of unlabeled data is fed into the algorithm at each training step. The main idea was to analyze the generating functional (8) by using the replica method. We applied the developed theoretical framework to the iterative ST that uses the ridge-regularized convex loss at each training step. In particular, we considered the data generation model in which each data point is generated from a binary Gaussian mixture with a general variance of each cluster Δ, the label bias ρ, and the magnitude of the PLS Γ. Prediction 1-4 itself, which captures the sharp asymptotics of ST by the effective single body problem, is valid even when the cluster size and the label bias are step-dependent.

Based on the derived formula, we studied the behavior of ST quantitatively. First, we explored the behavior of ST by numerically solving the self-consistent equations defined in Definition 1. Our findings include that (i) ST can find a model whose performance is nearly compatible with supervised learning with true labels when the label bias is small, (ii) when large number of iterations T can be used, the best achievable model results from accumulating small parameter updates over long iterations by using a small regularization parameter and moderately large batches of unlabeled data, and (iii) when available number of iterations T is small, PLS drastically reduce the generalization error. Second, to gain a deeper understanding of the behavior of ST when $1 \ll T$ and $\lambda_U \ll 1$, we performed a perturbative analysis by expanding the solution of the self-consistent equation in terms of λ_U. As a result, under some assumptions on the loss functions, it was shown that ST can potentially find a classification plane with the optimal direction regardless of the label bias by considering a perturbative analysis in λ_U. Intuitively speaking, this is due to the fact that the noise terms in effective description of the weight (28) and logits (35) becomes higher order in λ_U, hence noiseless accumulation of the information is possible. However, the imbalance between the norm of the weight and the magnitude of the bias might be problematic in label imbalanced cases. Notably, when using the squared loss without PLS, we could derive a closed-form solution for the evolution of cosine similarity, which clearly confirms the above result. Finally, to overcome the issues in label imbalanced cases, we proposed two heuristics: namely the pseudo-label annealing (Heuristics 1) and the bias-fixing (Heuristics 2). These heuristics drastically improved the performance of ST in label imbalanced cases. It was demonstrated that ST's performance is nearly compatible with supervised learning using true labels, even in cases of severe label imbalance. In label balanced case $\rho = 1/2$, we do not need to include the bias term, i.e., $\hat{B}^{(t)} = 0$ is optimal and the imbalance between the bias and the weight's norm is irrelevant. Thus, our findings are compatible with the previous studies (Oymak and Gulcu, 2020, 2021; Frei et al., 2022) in the label balanced case. However, our results suggest that ST is effective even in label imbalanced cases.

Overall, the numerical results suggest a crossover in the role of pseudo-labels depending on the number of ST iterations. When T is small, PLS and nearly hard pseudo-labels are effective, indicating that ST mainly benefits from fitting relatively reliable high-confidence pseudo-labels. This picture is consistent with the name of pseudo-labels. When T is large and α_U is moderately large, the optimal regularization parameter $\lambda_U^\star$ becomes small, so consecutive models remain close to each other and the pseudo-label loss acts mainly as a continuity constraint between successive iterates. In this regime, ST improves the classifier by accumulating many small updates, rather than by directly fitting noisy pseudo-labels.

Promising future directions include extending the present analysis to more advanced models such as the random feature model (Gerace et al., 2020), kernel method (Canatar et al., 2021b; Dietrich et al., 1999), and multi-layer neural networks (Schwarze and Hertz, 1992; Yoshino, 2020). Also, quantitatively comparing with other semi-supervised learning methods such as the entropy regularization (Grandvalet and Bengio, 2006) is a natural future research direction. Although the literature (Chen et al., 2020) shows that ST is the same with the entropy regularization if the pseudo-labels are updated after each SGD step, the connection is vague when the pseudo-label is fixed until the student model completely minimizes the loss as in our setup. Another natural but challenging direction is to extend our analysis to the case in which the same labeled/unlabeled data are used in each iteration,

which requires careful handling of the correlations of the estimators at different iteration steps as was done in the analysis of alternating minimization (Okajima and Takahashi, 2025). In such case, as reported in the analysis of knowledge distillation, one may need another heuristics such as early stopping (Takanami et al., 2025). The mixture of hard/soft-label is also an issue, requiring thorough analysis.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers JP21K21310, JP22H05117, JP23K16960, and JP26K02981, and by JST ACT-X Grant Number JPMJAX24CG.

Appendix A. Replica Method Calculation

In this appendix, we outline the replica calculation leading to the predictions 1-4. Also, let $\lambda_U^{(t)}, \alpha_U^{(t)}, \rho_U^{(t)}$ and $\Delta_U^{(t)}$ be the values of the regularization parameter, the relative size of the unlabeled data, and the label imbalance and the size of the cluster at step $t \geq 1$. Although we mainly focus on the case of $\lambda_U^{(t)} = \lambda_U, \alpha_U^{(t)} = \alpha_U, \rho_U^{(t)} = \rho_U$ and $\Delta_U^{(t)} = \Delta_U$ in the main text, the following derivation is valid even in these step-dependent cases.

A.1 Replica Method for Generating Functional of ST

We begin with recalling the basic strategy for evaluating the generating functional (8) using the replica method. The key observation is that the evaluation of the generating functional (8) requires averaging the inverse of the normalization constants $(Z^{(0)}(D_L))^{-1}$ and $\prod_{t=1}^{T-1}(Z^{(t)}(D_U^{(t)}, \boldsymbol{\theta}^{(t-1)}))^{-1}$ in the Boltzmann distributions:

$$\Xi_{\mathrm{ST}}(\epsilon_w, \epsilon_B) = \lim_{N,\beta\to\infty} \mathbb{E}_D \Bigg[\int e^{\epsilon_w g_w(\{w_i^{(t)}\}_{t=0}^T) + \epsilon_B g_B(\{B^{(t)}\}_{t=0}^T)} \times \frac{1}{Z^{(0)}(D_L)} e^{-\beta^{(0)}\mathcal{L}^{(0)}(\boldsymbol{\theta}^{(0)}; D_L)} \prod_{t=1}^{T} \frac{1}{Z^{(t)}(D_U^{(t)}, \boldsymbol{\theta}^{(t-1)})} e^{-\beta^{(t)}\mathcal{L}^{(t)}(\boldsymbol{\theta}^{(t)}; D_U^{(t)}, \boldsymbol{\theta}^{(t-1)})} d\boldsymbol{\theta}^{(0)} \dots d\boldsymbol{\theta}^{(T)} \Bigg].$$

Since the dependence of the normalization constants on D is non-trivial, taking average over D is technically difficult.

To resolve this difficulty, we use the replica method as follows. This method rewrites Ξ_{ST} using the identity $(Z^{(t)})^{-1} = \lim_{n_t\to 0}(Z^{(t)})^{n_t-1}$ as

$$\Xi_{\mathrm{ST}} = \lim_{n_0,\dots,n_T\to 0} \phi^{(T)}_{n_0,\dots,n_T}, \tag{42}$$

$$\begin{aligned}\phi^{(T)}_{n_0,\dots,n_T} = \lim_{N,\beta\to\infty} \mathbb{E}_D \Bigg[\int & e^{\epsilon_w g_w(\{w_i^{(t)}\}_{t=0}^T) + \epsilon_B g_B(\{B^{(t)}\}_{t=0}^T)} \\ & \times \left(Z^{(0)}(D_L)\right)^{n_0-1} e^{-\beta^{(0)}\mathcal{L}^{(0)}(\boldsymbol{\theta}^{(0)}; D_L)} \\ & \times \prod_{t=1}^{T} \left(Z^{(t)}(D_U^{(t)}, \boldsymbol{\theta}^{(t-1)})\right)^{n_t-1} e^{-\beta^{(t)}\mathcal{L}^{(t)}(\boldsymbol{\theta}^{(t)}; D_U^{(t)}, \boldsymbol{\theta}^{(t-1)})} d\boldsymbol{\theta}^{(0)} \dots d\boldsymbol{\theta}^{(T)} \Bigg].\end{aligned}$$

Although the equation (42) itself is just an identity, the evaluation of $\phi^{(T)}_{n_0,\dots,n_T}$ for $n_t \in \mathbb{R}$ is difficult. Still, for positive integers $n_0, \dots, n_T = 1, 2, \dots$, it has an appealing expression:

$$\phi^{(T)}_{n_0,\dots,n_T} = \lim_{N,\beta\to\infty} \mathbb{E}_D \Bigg[\int e^{\epsilon_w g_w(\{w^{(t)}_{1,i}\}^T_{t=0}) + \epsilon_B g_B(\{B^{(t)}_1\}^T_{t=0})} \prod_{a_0=1}^{n_0} e^{-\beta^{(0)}\mathcal{L}^{(0)}(\boldsymbol{\theta}^{(0)}_{a_t}; D_L)} \\ \times \prod_{t=1}^{T} \prod_{a_t=1}^{n_t} e^{-\beta^{(t)}\mathcal{L}^{(t)}(\boldsymbol{\theta}^{(t)}_{a_t}; D^{(t)}_U, \boldsymbol{\theta}^{(t-1)}_1)} d^{n_0}\boldsymbol{\theta}^{(0)} \dots d^{n_T}\boldsymbol{\theta}^{(T)} \Bigg],$$

where $d^{n_t}\boldsymbol{\theta}^{(t)}, t = 0, 1, \dots, T$ are the shorthand notations for $d\boldsymbol{\theta}^{(t)}_1 \dots d\boldsymbol{\theta}^{(t)}_{n_t}$, and $\{a_t\}^T_{t=0}$ are indices to distinguish the additional variables introduced by the replica method. Hereafter we will refer to this type of index as the *replica index*. Note that the index 1 that appears in $\epsilon_w g_w(\{w^{(t)}_{1,i}\}^T_{t=0}) + \epsilon_B g_B(\{B^{(t)}_1\}^T_{t=0})$ is also the replica index. The augmented system (43), which we refer to as the *replicated system*, is much easier to handle than the original generating functional because all of the factors to be evaluated are now explicit. Indeed, we can consider the average over D *before* completing the integral with respect to $\{\boldsymbol{\theta}^{(t)}_{c_t}\}$:

$$\phi^{(T)}_{n_0,\dots,n_T} = \lim_{N,\beta\to\infty} \int e^{\epsilon_w g_w(\{w^{(t)}_{1,i}\}^T_{t=0}) + \epsilon_B g_B(\{B^{(t)}_1\}^T_{t=0})} \mathbb{E}_D \Bigg[\prod_{a_0=1}^{n_0} e^{-\beta^{(0)}\mathcal{L}^{(0)}(\boldsymbol{\theta}^{(0)}_{a_t}; D_L)} \\ \times \prod_{t=1}^{T} \prod_{a_t=1}^{n_t} e^{-\beta^{(t)}\mathcal{L}^{(t)}(\boldsymbol{\theta}^{(t)}_{a_t}; D^{(t)}_U, \boldsymbol{\theta}^{(t-1)}_1)} \Bigg] d^{n_0}\boldsymbol{\theta}^{(0)} \dots d^{n_T}\boldsymbol{\theta}^{(T)}. \tag{43}$$

In the following, utilizing this formula, we obtain an analytical expression for $n_t \in \mathbb{N}$. Subsequently, under appropriate symmetry assumption of the resultant expression, we extrapolate it to the limit as $n_t \to 0$. This analytical continuation $n_t \to 0$ from $n_t \in \mathbb{N}$ is the procedure that has not yet been developed into a mathematically rigorous formulation. This aspect is what renders the replica method a non-rigorous technique. In principle, the evaluation of the replicated system for $n_t \in \mathbb{N}$ is a mathematically well-defined problem, and this parts could be made rigorous by an appropriate probabilistic analysis. See (Montanari and Sen, 2024) for an example in the analysis of the spiked tensor model. However, the expression for the replicated system (43) with integer $\{n_t\}^T_{t=0}$ alone may not uniquely determine the expression for the replicated system at real $\{n_t\}^T_{t=0}$. Hence the analytical continuation of $\phi^{(T)}_{n_0,\dots,n_T}$ from $n_t \in \mathbb{N}$ to $n_t \in \mathbb{R}$ is a mathematically ill-defined problem. Currently, what we can do is just *guess* the appropriate form from $n_t \in \mathbb{N}$ and extrapolate it to $n_t \in \mathbb{R}$. Nevertheless, for linear models, a widely applicable extrapolation procedure is known, which has succeeded in obtaining exact results for a variety of non-trivial problems as we already commented in subsection 1.1. In the following, we outline the treatment of the replicated system (43).

A.2 Handling of the Replicated System

We next consider how to evaluate the replicated system (43) using the RS assumption and saddle point method.

Inserting the explicit form of the loss functions (3)-(4) and the feature vectors (1)-(2) and using the independence of each data point, we obtain the following:

$$\begin{aligned}
\phi^{(T)}_{n_0,\ldots,n_T} &= \lim_{N,\beta\to\infty}\int e^{\epsilon_w g_w(\{w^{(t)}_{1,i}\}^T_{t=0})+\epsilon_B g_B(\{B^{(t)}_1\}^T_{t=0})} \\
&\times \prod_{\mu=1}^{M_L}\mathbb{E}_{\boldsymbol{x}^{(0)}_\mu,y^{(0)}_\mu}\left[\prod_{c_0=1}^{n_0} e^{-\beta^{(0)} l\left(y^{(0)}_\mu,\sigma(f^{(0)}_{c_0,\mu})\right)}\right] \\
&\times \prod_{t=1}^{T}\prod_{\nu=1}^{M_U}\mathbb{E}_{\boldsymbol{x}^{(t)}_\nu,y^{(t)}_\nu}\left[\prod_{c_t=1}^{n_t} e^{-\beta^{(t)}\mathbb{1}(\left|\tilde{f}^{(t)}_{c_t,\nu}\right|>\Gamma\sqrt{\|\boldsymbol{w}^{(t)}_{c_t}\|_2^2/N})l_{\mathrm{pl}}\left(\sigma_{\mathrm{pl}}(\tilde{f}^{(t)}_\nu,),\sigma(f^{(t)}_{c_t,\nu})\right)}\right] \\
&\times \prod_{t=0}^{T}\prod_{j=1}^{N}\prod_{c_t=1}^{n_t} e^{-\frac{\beta^{(t)}\lambda^{(t)}}{2}(w^{(t)}_{c_t,i})^2} d^{n_0}\boldsymbol{\theta}^{(0)}\ldots d^{n_T}\boldsymbol{\theta}^{(T)},
\end{aligned} \tag{44}$$

with the notations

$$\begin{aligned}
f^{(0)}_{c_0,\mu} &= B^{(0)}_{c_0} + (2y^{(0)}_\mu - 1)\frac{1}{N}\boldsymbol{w}^{(0)}_{c_0}\cdot\boldsymbol{v} + \frac{1}{\sqrt{N}}\boldsymbol{w}^{(0)}_{c_0}\cdot\boldsymbol{z}^{(0)}_\mu, \\
\tilde{f}^{(t)}_\nu &= B^{(t-1)}_1 + (2y^{(t)}_\nu - 1)\frac{1}{N}\boldsymbol{w}^{(t-1)}_1\cdot\boldsymbol{v} + \frac{1}{\sqrt{N}}\boldsymbol{w}^{(t-1)}_1\cdot\boldsymbol{z}^{(t)}_\nu, \\
f^{(t)}_{c_t,\nu} &= B^{(t)}_{c_t} + (2y^{(t)}_\nu - 1)\frac{1}{N}\boldsymbol{w}^{(t)}_1\cdot\boldsymbol{v} + \frac{1}{\sqrt{N}}\boldsymbol{w}^{(t)}_1\cdot\boldsymbol{z}^{(t)}_\nu.
\end{aligned}$$

A.2.1 Introduction of Order Parameters and the Saddle Point Method

The important observation is that the Gaussian noise terms in (44) appear only through the following vectors:

$$\begin{aligned}
\boldsymbol{u}^{(0)}_\mu &= \left(\frac{1}{\sqrt{N}}\boldsymbol{w}^{(0)}_1\cdot\boldsymbol{z}^{(0)}_\mu, \frac{1}{\sqrt{N}}\boldsymbol{w}^{(0)}_2\cdot\boldsymbol{z}^{(0)}_\mu, \ldots, \frac{1}{\sqrt{N}}\boldsymbol{w}^{(0)}_{n_0}\cdot\boldsymbol{z}^{(0)}_\mu\right)^\top \in \mathbb{R}^{n_0}, \\
\boldsymbol{u}^{(t)}_\nu &= \left(\frac{1}{\sqrt{N}}\boldsymbol{w}^{(t-1)}_1\cdot\boldsymbol{z}^{(t)}_\nu, \frac{1}{\sqrt{N}}\boldsymbol{w}^{(t)}_1\cdot\boldsymbol{z}^{(t)}_\nu, \frac{1}{\sqrt{N}}\boldsymbol{w}^{(t)}_2\cdot\boldsymbol{z}^{(t)}_\nu, \ldots, \frac{1}{\sqrt{N}}\boldsymbol{w}^{(t)}_{n_t}\cdot\boldsymbol{z}^{(t)}_\nu\right)^\top \in \mathbb{R}^{n_t+1}, \\
&\quad t = 1,2,\ldots,T,
\end{aligned}$$

which follows independently the multivariate Gaussians as

$$\begin{aligned}
\boldsymbol{u}^{(0)}_\mu &\sim_{\mathrm{iid}} \mathcal{N}(0,\Sigma^{(0)}), \quad \Sigma^{(0)} = \Delta_L \times \begin{bmatrix} Q^{(0)}_{11} & \cdots & Q^{(0)}_{1n_0} \\ \vdots & \ddots & \vdots \\ Q^{(0)}_{n_0 1} & \cdots & Q^{(0)}_{n_0 n_0} \end{bmatrix}, \\
Q^{(0)}_{c_0 d_0} &\equiv \frac{1}{N}\boldsymbol{w}^{(0)}_{c_0}\cdot\boldsymbol{w}^{(0)}_{d_0}, \quad c_0, d_0 = 1,2,\ldots,n_0,
\end{aligned}$$

and

$$\begin{aligned}
\boldsymbol{u}^{(t)}_\nu &\sim_{\mathrm{iid}} \mathcal{N}(0,\Sigma^{(t)}), \quad \Sigma^{(t)} = \Delta^{(t)}_U \times \left[\begin{array}{c|ccc} Q^{(t-1)}_{11} & R^{(t)}_1 & \cdots & R^{(t)}_{n_t} \\ \hline R^{(t)}_1 & Q^{(t)}_{11} & \cdots & Q^{(t)}_{1n_t} \\ \vdots & \vdots & \ddots & \vdots \\ R^{(t)}_{n_t} & Q^{(t)}_{n_t 1} & \cdots & Q^{(t)}_{n_t n_t} \end{array}\right] \\
Q^{(t)}_{c_t d_t} &= \frac{1}{N}\boldsymbol{w}^{(t)}_{c_t}\cdot\boldsymbol{w}^{(t)}_{d_t}, \quad R^{(t)}_{c_t} = \frac{1}{N}\boldsymbol{w}^{(t-1)}_1\cdot\boldsymbol{w}^{(t)}_{c_t}, \quad c_t, d_t = 1,2,\ldots,n_t,
\end{aligned}$$

for a fixed set of $\{\boldsymbol{w}_{c_t}^{(t)}\}_{c_t=1}^{n_t}, t=0,\dots,T$. Also, the center of the cluster $\boldsymbol{v}$ appears only in the form of the inner product $m_{c_t}^{(t)} = \frac{1}{N}\boldsymbol{v}\cdot\boldsymbol{w}_{c_t}^{(t)}$. Thus, the replicated system (12) depends on $\{\boldsymbol{w}_{c_t}^{(t)}\}_{c_t=1}^{n_t}, t=0,\dots,T$ only through their inner products, such as

$$\begin{aligned}
&\frac{1}{N}\boldsymbol{w}_{c_t}^{(t)}\cdot\boldsymbol{w}_{d_t}^{(t)}, \quad c_t, d_t = 1,2,\dots,n_t,\ t=0,1,\dots T,\\
&\frac{1}{N}\boldsymbol{w}_{c_t}^{(t)}\cdot\boldsymbol{v}, \quad c_t = 1,\dots,n_t,\ t=0,1,\dots,T,\\
&\frac{1}{N}\boldsymbol{w}_{1}^{(t-1)}\cdot\boldsymbol{w}_{c_t}^{(t)}, \quad c_t = 1,2,\dots,n_t,\ t=1,\dots,T,
\end{aligned}$$

which capture the geometric relations between the estimators and the centroid of clusters $\boldsymbol{v}$. We refer to them as *order parameters*. Furthermore, since each data point is independent, the integral over $(\boldsymbol{u}_\mu^{(0)}, y_\mu^{(0)})$ and $(\boldsymbol{u}_\nu^{(t)}, y_\nu^{(t)})$ can be taken independently. As a result, the generating functional does not depend on the index μ and ν.

The above observation indicates that the factors regarding the loss function in (44) can be evaluated as Gaussian integrals once the order parameters are fixed. To implement this idea, we insert the trivial identities of the delta functions

$$\begin{aligned}
1 &= \prod_{t=0}^{T}\prod_{1\le c_t\le d_t\le n_t} N\int \delta_{\mathrm{d}}\left(NQ_{c_td_t}^{(t)} - \boldsymbol{w}_{c_t}^{(t)}\cdot\boldsymbol{w}_{d_t}\right)dQ_{c_td_t}^{(t)},\\
1 &= \prod_{t=0}^{T}\prod_{c_t=1}^{n_t} N\int \delta_{\mathrm{d}}\left(Nm_{c_t}^{(t)} - \boldsymbol{w}_{c_t}^{(t)}\cdot\boldsymbol{v}\right)dm_{c_t}^{(t)},\\
1 &= \prod_{t=1}^{T}\prod_{c_t=1}^{n_t} N\int \delta_{\mathrm{d}}\left(NR_{c_t}^{(t)} - \boldsymbol{w}_{1}^{(t-1)}\cdot\boldsymbol{w}_{c_t}^{(t)}\right)dR_{c_t}^{(t)},
\end{aligned}$$

into (44). Then the replicated system can be written as follows:

$$\begin{aligned}
\phi_{n_0,\dots,n_T}^{(T)} = \lim_{N,\beta\to\infty}\int\ & e^{\epsilon_w g_w(\{w_{1,i}^{(t)}\}_{t=0}^{T})+\epsilon_B g_B(\{B_1^{(t)}\}_{t=0}^{T})} e^{N\alpha_L\varphi_u^{(0)}+\sum_{t=1}^{T}N\alpha_U\varphi_u^{(t)}}\\
&\times\prod_{t=0}^{T}\prod_{1\le c_t\le d_t\le n_t}\delta_{\mathrm{d}}\left(NQ_{c_td_t}^{(t)} - \boldsymbol{w}_{c_t}^{(t)}\cdot\boldsymbol{w}_{d_t}\right)\prod_{c_t=1}^{n_t}\delta_{\mathrm{d}}\left(Nm_{c_t}^{(t)} - \boldsymbol{w}_{c_t}^{(t)}\cdot\boldsymbol{v}\right)\\
&\times\prod_{t=1}^{T}\prod_{c_t=1}^{n_t}\delta_{\mathrm{d}}\left(NR_{c_t}^{(t)} - \boldsymbol{w}_{1}^{(t-1)}\cdot\boldsymbol{w}_{c_t}^{(t)}\right)\prod_{t=0}^{T}\prod_{j=1}^{N}\prod_{c_t=1}^{n_t} e^{-\frac{\beta^{(t)}\lambda^{(t)}}{2}(w_{c_t,i}^{(t)})^2} d^{n_0}\boldsymbol{\theta}^{(0)}\dots d^{n_T}\boldsymbol{\theta}^{(T)}d\Theta,
\end{aligned}$$

$$\begin{aligned}
e^{\varphi_u^{(0)}} &= \mathbb{E}_{\boldsymbol{u}^{(0)}\sim\mathcal{N}(\boldsymbol{0},\Sigma^{(0)}),\ y^{(0)}\sim p_y^{(0)}}\left[\prod_{c_0=1}^{n_0} e^{-\beta^{(0)}l(y^{(0)},\sigma(f_{c_0}^{(0)}))}\right],\\
e^{\varphi_u^{(t)}} &= \mathbb{E}_{\boldsymbol{u}^{(t)}\sim\mathcal{N}(\boldsymbol{0},\Sigma^{(t)}),\ y^{(t)}\sim p_y^{(t)}}\left[\prod_{c_t=1}^{n_t} e^{-\beta^{(t)}\mathbb{1}(|\tilde{f}^{(t)}|>\Gamma\sqrt{Q_{11}^{(t-1)}})l_{\mathrm{pl}}(\sigma_{\mathrm{pl}}(\tilde{f}^{(t)}),\sigma(f_{c_t}^{(t)}))}\right]\\
f_{c_0}^{(0)} &= B_{c_0}^{(0)} + (2y^{(0)}-1)m_{c_0}^{(0)} + u_{c_0}^{(0)},\\
\tilde{f}^{(t)} &= B_1^{(t-1)} + (2y^{(t)}-1)m_1^{(t-1)} + u_1^{(t)},\\
f_{c_t}^{(t)} &= B_{c_t}^{(t)} + (2y^{(t)}-1)m_{c_t}^{(t)} + u_{c_t+1}^{(t)},
\end{aligned}$$

where Θ is the collection of the following variables

$$Q^{(t)} = [Q^{(t)}_{c_t d_t}]_{\substack{1\le c_t\le n_t\\ 1\le d_t\le n_t}},\ \boldsymbol{m}^{(t)} = [m^{(t)}_{c_t}]_{1\le c_t\le n_t},\ \boldsymbol{R}^{(t)} = [R^{(t)}_{c_t}]_{1\le c_t\le n_t},\ \boldsymbol{B}^{(t)} = [b^{(t)}_{c_t}]_{1\le c_t\le n_t}.$$

A.2.2 Decoupling the Integrals and the Saddle Point Method

To complete the remaining integrals over $\{\boldsymbol{\theta}^{(t)}_{c_t}\}$, we use the Fourier representations of the delta functions:

$$\begin{aligned}
\delta_{\mathrm{d}}\left(NQ^{(t)}_{c_t d_t} - \boldsymbol{w}^{(t)}_{c_t}\cdot\boldsymbol{w}_{d_t}\right) &= \frac{1}{4\pi}\int e^{-(NQ^{(t)}_{c_t d_t} - \boldsymbol{w}^{(t)}_{c_t}\cdot\boldsymbol{w}_{d_t})\frac{\tilde{Q}^{(t)}_{c_t d_t}}{2}} d\tilde{Q}^{(t)}_{c_t d_t},\\
\delta_{\mathrm{d}}\left(Nm^{(t)}_{c_t} - \boldsymbol{w}^{(t)}_{c_t}\cdot\boldsymbol{v}\right) &= \frac{1}{2\pi}\int e^{(Nm^{(t)}_{c_t} - \boldsymbol{w}^{(t)}_{c_t}\cdot\boldsymbol{v})\tilde{m}^{(t)}_{c_t}} d\tilde{m}^{(t)}_{c_t},\\
\delta_{\mathrm{d}}\left(NR^{(t)}_{c_t} - \boldsymbol{w}^{(t-1)}_{1}\cdot\boldsymbol{w}^{(t)}_{c_t}\right) &= \frac{1}{2\pi}\int e^{(NR^{(t)}_{c_t} - \boldsymbol{w}^{(t-1)}_{1}\cdot\boldsymbol{w}^{(t)}_{c_t})\tilde{R}^{(t)}_{c_t}} d\tilde{R}^{(t)}_{c_t}.
\end{aligned}$$

After using these expressions, the integrals over $\{w^{(t)}_{a_t,j}\}$ can be independently performed for each j. As a result, the result no longer depends on the index i, hence, we can safely omit it. This is a natural consequence of the data distribution having rotation-invariant symmetry.

Specifically, we find that $\phi^{(T)}_{n_0,\dots,n_T}$ can be written as

$$\phi^{(T)}_{n_0,\dots,n_T} = \lim_{N,\beta\to\infty}\int e^{N\mathcal{S}(\Theta,\hat{\Theta})}\mathbb{E}_{\{\boldsymbol{w}^{(t)}\}\sim\tilde{p}_{\mathrm{eff},w}}\left[e^{\epsilon_w g_w(\{w^{(t)}_1\}^T_{t=0})}\right]e^{\epsilon_B g_B(\{B^{(t)}_1\}^T_{t=0})}d\Theta d\hat{\Theta}, \tag{45}$$

where $\hat{\Theta}$ is the collection of variables $\tilde{Q}^{(t)} = [\tilde{Q}^{(t)}_{c_t d_t}]_{\substack{1\le c_t\le n_t\\ 1\le d_t\le n_t}}$, $\tilde{\boldsymbol{m}}^{(t)} = [\tilde{m}^{(t)}_{c_t}]_{1\le c_t\le n_t}$, $\tilde{\boldsymbol{R}}^{(t)} = [\tilde{R}^{(t)}_{c_t}]_{1\le c_t\le n_t}$, and

$$\begin{aligned}
\tilde{p}_{\mathrm{eff},w}(\{\boldsymbol{w}^{(t)}\}^T_{t=0}) &= \frac{1}{\tilde{Z}_{\mathrm{eff},w}}e^{-\frac{1}{2}(\boldsymbol{w}^{(0)})^\top(\tilde{Q}^{(0)}+\beta^{(t)}I_{n_t})\boldsymbol{w}^{(0)}+(\tilde{\boldsymbol{m}}^{(t)})\cdot\boldsymbol{w}^{(0)}}\\
&\qquad\times\prod_{t=1}^{T}e^{-\frac{1}{2}(\boldsymbol{w}^{(t)})^\top(\tilde{Q}^{(t)}+\beta^{(t)}I_{n_t})\boldsymbol{w}^{(t)}+(\tilde{\boldsymbol{m}}^{(t)}+w^{(t-1)}_1\boldsymbol{R}^{(t)})\cdot\boldsymbol{w}^{(t)}},\\
\mathcal{S}(\Theta,\hat{\Theta}) &= (1-1/N)\alpha_L\varphi^{(0)}_u + \sum_{t=1}^{T}(1-1/N)\alpha^{(t)}_U\varphi^{(t)}_u + (1-1/N)\varphi_w, \qquad (46)\\
\varphi_w &= \log\tilde{Z}_{\mathrm{eff},w},
\end{aligned}$$

where the factors $1/N$ that appears in the coefficients of $\varphi^{(t)}_u$ and φ_w are due to the presence of g_w and g_B. Also $\boldsymbol{w}^{(t)} = (w^{(t)}_1,\dots w^{(t)}_{n_t})\cdot\in\mathbb{R}^{n_t}$. Thus, Θ and $\hat{\Theta}$ are evaluated by the extremum condition $\mathrm{extr}_{\Theta,\hat{\Theta}}\mathcal{S}(\Theta,\hat{\Theta})$, in which the above factor $1/N$ is negligible. Recall that the index 1 in (13) is the replica index, and the result no longer depends on the index $i\in[N]$. It should also be noted that, essentially, only Gaussian integrals and the saddle-point method are required for the above calculations once the order parameters are determined. This computational simplicity is a major advantage of the replica method that introduces the replicated system (43), which enables us to consider the average over the noise before completing the integral (or the optimization at $\beta\to\infty$) over $\{\boldsymbol{\theta}_{c_t}\}^T_{t=0}$.

To obtain the saddle point condition regarding $\hat{\Theta}$, it is useful to use the following identity for the Gaussian integral with mean $\bar{\boldsymbol{\mu}}$ and covariance S and twice-differentiable function $\mathcal{F}$:

$$\mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(\bar{\boldsymbol{\mu}},S)}[\mathcal{F}(\boldsymbol{z})] = \exp\left(\frac{1}{2}\sum_{i,j} S_{ij}\frac{\partial^2}{\partial h_i \partial h_j}\right)\mathcal{F}(\boldsymbol{h})\Big|_{\boldsymbol{h}=\bar{\boldsymbol{\mu}}}, \tag{47}$$

which is commonly used in the mean-field theory of glassy systems (Parisi et al., 2020). The right-hand side of (47) is defined by a formal Taylor expansion of the exponential function:

$$\exp\left(\frac{1}{2}\sum_{i,j} S_{ij}\frac{\partial^2}{\partial h_i \partial h_j}\right)\mathcal{F}(\boldsymbol{h}) \equiv \left(1 + \frac{1}{2}\sum_{i,j} S_{ij}\frac{\partial^2}{\partial h_i \partial h_j} + \dots\right)\mathcal{F}(\boldsymbol{h}).$$

Let $\alpha^{(0)}$ and $\Delta^{(0)}$ be α_L and Δ_L, respectively. Let $\alpha^{(t)}$ be α_L if $t = 0$ and α_U otherwise. Also, define $\tilde{p}^{(t)}_{\mathrm{eff},u|y}, \tilde{p}^{(t)}_{\mathrm{eff},uy}$ and $l^{(t)}_{c_t}$ as follows:

$$\begin{aligned}
\tilde{p}^{(t)}_{\mathrm{eff},u|y}(\boldsymbol{u}^{(t)}|y^{(t)}) &\propto \prod_{c_t=1}^{n_t} e^{-\beta^{(t)} l^{(t)}_{c_t}} \times \mathcal{N}(\boldsymbol{u}^{(t)}; \boldsymbol{0}, \Sigma^{(t)}),\\
\tilde{p}^{(t)}_{\mathrm{eff},uy}(\boldsymbol{u}^{(t)}, y^{(t)}) &= \tilde{p}_{\mathrm{eff},u|y}(\boldsymbol{u}^{(t)}|y^{(t)}) p^{(t)}_y(y^{(t)}),\\
l^{(0)}_{c_0} &= l^{(0)}_{c_0}(y^{(0)}, f^{(0)}_{c_0}) = l\left(y^{(0)}, \sigma(f^{(0)}_{c_0})\right),\\
l^{(t)}_{c_t} &= l^{(t)}_{c_t}(\tilde{f}^{(t)}, f^{(t)}_{c_t}) = \mathbb{1}\left(|\tilde{f}^{(t)}| > \Gamma\sqrt{Q^{(t-1)}_{11}}\right) l_{\mathrm{pl}}\left(\sigma_{\mathrm{pl}}(\tilde{f}^{(t)}), \sigma(f^{(t)}_{c_t})\right).
\end{aligned}$$

Then, by using the identity (47) for expressing the Gaussian average in $\varphi^{(t)}_u$, the saddle point condition over $(\Theta, \hat{\Theta})$ can be written as follows:

$$\begin{aligned}
Q^{(t)}_{c_t d_t} &= \mathbb{E}_{\{\boldsymbol{w}^{(t)}\}\sim\tilde{p}_{\mathrm{eff},w}}\left[w^{(t)}_{c_t} w^{(t)}_{d_t}\right],\\
R^{(t)}_{c_t} &= \mathbb{E}_{\{\boldsymbol{w}^{(t)}\}\sim\tilde{p}_{\mathrm{eff},w}}\left[w^{(t)}_{c_t} w^{(t-1)}_{1}\right],\\
m^{(t)}_{c_t} &= \mathbb{E}_{\{\boldsymbol{w}^{(t)}\}\sim\tilde{p}_{\mathrm{eff},w}}\left[w^{(t)}_{c_t}\right],\\
\tilde{Q}^{(t)}_{c_t d_t} &= -\alpha^{(t)}\Delta^{(t)}\mathbb{E}_{\boldsymbol{u}^{(t)}, y^{(t)}\sim p^{(t)}_{\mathrm{eff},uy}}\left[(\beta^{(t)})^2 \partial_2 l^{(t)}_{c_t}\partial_2 l^{(t)}_{d_t} - \beta^{(t)}\partial_2^2 l^{(t)}_{c_t}\delta_{c_t,d_t}\right],\\
\tilde{R}^{(t)}_{c_t d_t} &= \alpha^{(t)}\Delta^{(t)}\beta^{(t)}\mathbb{E}_{\boldsymbol{u}^{(t)}, y^{(t)}\sim p^{(t)}_{\mathrm{eff},uy}}\left[\partial_1\partial_2 l^{(t)}_{c_t}\right],\\
\tilde{m}^{(t)}_{c_t} &= -\alpha^{(t)}\beta^{(t)}\mathbb{E}_{\boldsymbol{u}^{(t)}, y^{(t)}\sim p^{(t)}_{\mathrm{eff},uy}}\left[(2y^{(t)} - 1)\partial_2 l^{(t)}_{c_t}\right],\\
0 &= \mathbb{E}_{\boldsymbol{u}^{(t)}\sim p^{(t)}_{\mathrm{eff},u}\, y\sim p^{(t)}_y}\left[\partial_2 l^{(t)}_{c_t}\right].
\end{aligned}$$

Recall that, for a bivariate function $\mathcal{F}(y, x)$, let $\partial_i\mathcal{F}$ be the partial derivative of $\mathcal{F}$ with respect to the i-th argument as defined in subsection 1.2. For example, $\partial_1\mathcal{F}(Y, X) = \frac{\partial\mathcal{F}}{\partial y}\Big|_{y=Y,x=X}$. Higher-order derivatives are defined similarly. In fact, the saddle point condition regarding $Q^{(t)}_{11}, m^{(t)}_1$ and $B^{(t)}_1, t \in [T-1]$ yields the derivative regarding $l^{(t+1)}_{c_{t+1}}$. However, such terms are an effect from step $t+1$ to t, which is causally inconsistent. Hence we omit them. In fact, a rigorous calculation reveals that these terms vanish in the limit as $n_{t+1} \to 0$ under the RS assumption that will be introduced below.

A.2.3 RS Saddle Point Condition

Since equation (45) depends on the discrete nature of n_t, the extrapolation of $\lim_{n\to 0}\phi^{(T)}_{n_0,\ldots,n_T}$ is still difficult. The key issue here is to identify the correct form of the saddle point. The simplest form is the following RS saddle point:

$$
\begin{aligned}
Q^{(t)} &= \frac{\chi^{(t)}}{\beta^{(t)}} I_{n_t} + q^{(t)} \mathbf{1}_{n_t}\mathbf{1}_{n_t}{}^{\cdot}, \\
\boldsymbol{R}^{(t)} &= R^{(t)}\mathbf{1}_{n_t}, \\
\boldsymbol{m}^{(t)} &= m^{(t)}\mathbf{1}_{n_t}, \\
\boldsymbol{B}^{(t)} &= B^{(t)}\mathbf{1}_{n_t}, \\
\tilde{Q}^{(t)} &= \beta^{(t)}\hat{Q}^{(t)} I_{n_t} - (\beta^{(t)})^2 \hat{\chi}^{(t)} \mathbf{1}_{n_t}\mathbf{1}_{n_t}{}^{\cdot}, \\
\tilde{\boldsymbol{R}}^{(t)} &= \beta^{(t)}\hat{R}^{(t)}\mathbf{1}_{n_t}, \\
\tilde{\boldsymbol{m}}^{(t)} &= \beta^{(t)}\hat{m}^{(t)}\mathbf{1}_{n_t},
\end{aligned}
$$

which is the simplest form of the saddle point reflecting the symmetry of the variational function. This choice is motivated by the fact that $\mathcal{S}^{(T)}(\Theta, \hat{\Theta})$ is invariant under the permutation of the indices $c_t \in \{1, 2, \ldots, n_t\}$ for any $t = 0, 1, \ldots$. In a mathematically rigorous sense, the expression for the replicated system (12) with integer $\{n_t\}_{t=0}^T$ alone cannot uniquely determine the expression for the replicated system at real $\{n_t\}_{t=0}^T$. However, for log-convex Boltzmann distributions, it is empirically known that the replica symmetric choice of the saddle point yields the correct extrapolation (Barbier and Macris, 2019; Barbier et al., 2019; Mignacco et al., 2020a; Gerbelot et al., 2020, 2023; Montanari and Sen, 2024) in the sense that the same result by the replica method with the RS assumption have been obtained through a different mathematically rigorous approach. Since the Boltzmann distributions are log-convex functions once conditioned on the data and the parameter of the previous iteration step, we can expect the RS assumption to yield the correct result even in the current iterative optimization.

Under the RS assumption, after simple algebra, $\tilde{p}_{\mathrm{eff},w}$ and $\tilde{p}_{\mathrm{eff},u|y}$ can be rewritten as follows:

$$
\begin{aligned}
\tilde{p}_{\mathrm{eff},w}(\{\boldsymbol{w}^{(t)}\}) &= \mathbb{E}_{\xi_w^{(t)}\sim_{\mathrm{iid}}\mathcal{N}(0,1)}\left[\prod_{c_0=1}^{n_0}\mathcal{N}\left(w_{c_0}^{(0)} \mid \frac{\hat{m}^{(0)} + \sqrt{\hat{\chi}^{(0)}}\xi_w^{(0)}}{\hat{Q}^{(0)} + \lambda^{(0)}}, \frac{\hat{Q}^{(0)} + \lambda^{(0)}}{\beta^{(0)}}\right)\right. \\
&\qquad \left. \times \prod_{t=1}^{T}\prod_{c_t=1}^{n_t}\mathcal{N}\left(w_{c_t}^{(t)} \mid \frac{\hat{m}^{(t)} + \hat{R}^{(t)} w_1^{(t-1)} + \sqrt{\hat{\chi}^{(t)}}\xi_w^{(t)}}{\hat{Q}^{(t)} + \lambda^{(t)}}, \frac{\hat{Q}^{(t)} + \lambda^{(t)}}{\beta^{(t)}}\right)\right], \\
\tilde{p}^{(0)}_{\mathrm{eff},u|y}(\boldsymbol{u}^{(0)}|y^{(0)}) &= \mathbb{E}_{\xi_u^{(0)}\sim\mathcal{N}(0,\Delta_L)}\left[\prod_{c_0=1}^{n_0} e^{-\beta^{(0)} l\left(y^{(0)}, \sigma(h_u^{(0)} + u_{c_0}^{(0)})\right)}\mathcal{N}(u_{c_0}^{(0)} \mid 0, \frac{\chi^{(0)}\Delta_L}{\beta^{(0)}})\right], \\
\tilde{p}^{(t)}_{\mathrm{eff},u|y}(\boldsymbol{u}^{(t)}|y^{(t)}) &= \mathbb{E}_{\xi_{u,1}^{(t)}, \xi_{u,2}^{(t)}\sim_{\mathrm{iid}}\mathcal{N}(0,\Delta_U^{(t)})}\left[\prod_{c_t=1}^{n_t} e^{-\beta^{(t)} l_{\mathrm{pl}}\left(\sigma_{\mathrm{pl}}(\tilde{h}_u^{(t)} + \tilde{u}^{(t)}), \sigma(h_u^{(t)} + u_{c_t}^{(t)})\right)}\right. \\
&\qquad \left. \times \mathcal{N}(u_{c_t}^{(t)} \mid 0, \frac{\chi^{(t)}\Delta_U^{(t)}}{\beta^{(t)}})\mathcal{N}(\tilde{u}^{(t)} \mid 0, \frac{\chi^{(t-1)}\Delta_U^{(t)}}{\beta^{(t-1)}})\right], \quad (48)
\end{aligned}
$$

$$
\begin{aligned}
h_u^{(0)} &= B^{(0)} + (2y^{(t)} - 1)m^{(0)} + \sqrt{q^{(0)}}\xi_u^{(0)}, \\
\tilde{h}_u^{(t)} &= B^{(t-1)} + (2y^{(t)} - 1)m^{(t-1)} + \sqrt{q^{(t-1)}}\xi_{u,1}^{(t)}, \\
h_u^{(t)} &= B^{(t)} + (2y^{(t)} - 1)m^{(t)} + \frac{R^{(t)}}{\sqrt{q^{(t-1)}}}\xi_{u,1}^{(t)} + \sqrt{q^{(t)} - \frac{(R^{(t)})^2}{q^{(t-1)}}}\xi_{u,2}^{(t)}.
\end{aligned}
$$

Importantly, the factors inside the average over $\xi_w^{(t)}, \xi_{u,1}^{(t)}, \xi_{u,2}^{(t)}$ are factorized over the replica indexes. This indicates that the integral over $w_{c_t}^{(t)}, u_{c_t}^{(t)}$ are taken independently before the taking the averaging over $\xi_w^{(t)}, \xi_{u,1}^{(t)}, \xi_{u,2}^{(t)}$, which makes the computation considerably simple. Also, due to this symmetry, the resultant expressions can often be formally extrapolated from $n_t \in \mathbb{N}$ to $n_t \in \mathbb{R}$. To see this, first, let us consider $\varphi_u^{(t)}$. By conducting straightforward integrals, we obtain the following

$$
\begin{aligned}
e^{\varphi_u^{(0)}} &= \mathbb{E}_{y^{(0)}, \xi_u^{(0)}} \left[\left(\int e^{-\beta^{(0)} l\left(y^{(0)}, \sigma(h_u^{(0)} + u^{(0)})\right)} \mathcal{N}(u^{(0)} \mid 0, \frac{\chi^{(0)} \Delta_L}{\beta^{(0)}}) du^{(0)} \right)^{n_0} \right], \\
e^{\varphi_u^{(t)}} &= \mathbb{E}_{y^{(t)}, \xi_{u,1}^{(t)}, \xi_{u,2}^{(t)}, \tilde{u}^{(t)}} \left[\left(\int e^{-\beta^{(t)} l_{\mathrm{pl}}\left(\sigma_{\mathrm{pl}}(\tilde{h}_u^{(t)} + \tilde{u}^{(t)}), \sigma(h_u^{(t)} + u^{(t)})\right)} \mathcal{N}(u^{(t)} \mid 0, \frac{\chi^{(t)} \Delta_U^{(t)}}{\beta^{(t)}}) du^{(t)} \right)^{n_t} \right].
\end{aligned}
$$

Next, let us consider φ_w. For this, let us define $p_{\mathrm{eff},w}^{(t)}$ as follows.

$$
\begin{aligned}
p_{\mathrm{eff},w}^{(0)}(w^{(0)}) &= \frac{1}{Z_{\mathrm{eff},w}^{(0)}} e^{-\beta^{(0)}(\frac{\hat{Q}^{(0)} + \lambda^{(0)}}{2}(w^{(0)})^2 - (\hat{m}^{(0)} + \sqrt{\hat{\chi}^{(0)}}\xi_w^{(t)})w^{(0)})}, \\
p_{\mathrm{eff},w}^{(t)}(w^{(t)}) &= \frac{1}{Z_{\mathrm{eff},w}^{(t)}(w^{(t-1)})} e^{-\beta^{(t)}(\frac{\hat{Q}^{(t)} + \lambda^{(t)}}{2}(w^{(t)})^2 - (\hat{m}^{(t)} + \hat{R}^{(t)} w^{(t-1)} + \sqrt{\hat{\chi}^{(t)}}\xi_w^{(t)})w^{(t)})}, \\
w^{(t-1)} &\sim p_{\mathrm{eff},w}^{(t-1)}.
\end{aligned}
$$

Then, by performing the integrals successively on $\{w_{c_T}^{(T)}\}, \{w_{c_{T-1}}^{(T-1)}\}, \ldots$, we obtain the following:

$$
\begin{aligned}
\Phi^{(T)}(w^{(T-1)}) &= (Z_{\mathrm{eff}^{(T)}}(w^{(T-1)}))^{n_T}, \quad w^{(T-1)} \sim p_{\mathrm{eff},w}^{(T-1)}, \\
\Phi^{(t)}(w^{(t-1)}) &= \mathbb{E}_{w^{(t)} \sim p_{\mathrm{eff},w}^{(t)}} [\Phi^{(t+1)}(w^{(t)})] (Z_{\mathrm{eff},w}^{(t)}(w^{(t-1)}))^{n_t}, \\
&\qquad\qquad w^{(t-2)} \sim p_{\mathrm{eff},w}^{(t-2)}, \quad t = T-1, \ldots, 2, 1, \\
e^{\varphi_w} &= \mathbb{E}_{w^{(0)} \sim p_{\mathrm{eff},w}^{(0)}} [\Phi^{(1)}(w^{(0)})] (Z_{\mathrm{eff},w}^{(0)})^{n_0}.
\end{aligned}
$$

By formally considering $(\ldots)^{n_t}$ as an exponential function of n_t, we can extrapolate the above result to $n_t \in \mathbb{R}$. Then, it is clear that $\varphi_u^{(t)}$ and φ_w are expanded as $1 + \mathcal{O}(n)$, where $\mathcal{O}(n) \equiv \sum_{t=0}^{T} \mathcal{O}(n_t)$ be terms that vanish at the limit $n_0, n_1, \ldots, n_T \to 0$. From this, it can be straightforwardly shown that $\mathcal{S}$ evaluated as the saddle point is also expanded as $\mathcal{S} = \mathcal{O}(n)$.

Therefore, $\phi^{(T)}_{n_0,\ldots,n_T}$ can be evaluated as follows:

$$\phi^{(T)}_{n_0,\ldots,n_T} = \lim_{\beta\to\infty} \mathbb{E}_{\xi^{(t)}_w \sim_{\mathrm{iid}} \mathcal{N}(0,1)} \left[\mathbb{E}_{\{w^{(t)}\}\sim p_{\mathrm{eff},w}(\{w^{(t)}\}|\{\xi^{(t)}_w\})} \left[e^{\epsilon_w g_w(\{w^{(t)}\}^T_{t=0})} \right] \right] e^{\epsilon_B g_B(\{B^{(t)}\}^T_{t=0})} + \mathcal{O}(n),$$

where

$$\begin{aligned} p_{\mathrm{eff},w}(\{w^{(t)}\}^T_{t=0}|\{\xi^{(t)}_w\}) = \mathcal{N}\left(w^{(0)} \mid \frac{\hat{m}^{(0)} + \sqrt{\hat{\chi}^{(0)}}\xi^{(0)}_w}{\hat{Q}^{(0)} + \lambda^{(0)}}, \frac{\hat{Q}^{(0)} + \lambda^{(0)}}{\beta^{(0)}} \right) \\ \times \prod_{t=1}^{T} \mathcal{N}\left(w^{(t)} \mid \frac{\hat{m}^{(t)} + \hat{R}^{(t)} w^{(t-1)} + \sqrt{\hat{\chi}^{(t)}}\xi^{(t)}_w}{\hat{Q}^{(t)} + \lambda^{(t)}}, \frac{\hat{Q}^{(t)} + \lambda^{(t)}}{\beta^{(t)}} \right). \end{aligned}$$

The parameters such as $q^{(t)}, \chi^{(t)}, \ldots$ are determined by the saddle point conditions. Also, at the limit $\beta \to \infty, n_0, \ldots, n_T \to 0$, the measure defined by $p_{\mathrm{eff},w}$ concentrates, and it is governed by a one-dimensional Gaussian process:

$$\Xi_{\mathrm{ST}}(\epsilon_w, \epsilon_B) = \mathbb{E}_{\xi^{(t)}_w \sim_{\mathrm{iid}} \mathcal{N}(0,1)} \left[e^{\epsilon_w g_w(\{\hat{\mathsf{w}}^{(t)}\}^T_{t=0})} \right] e^{\epsilon_B g_B(\{B^{(t)}\}^T_{t=0})}, \tag{49}$$

$$\hat{\mathsf{w}}^{(0)} = \frac{\hat{m}^{(0)} + \sqrt{\hat{\chi}^{(0)}}\xi^{(0)}_w}{\hat{Q}^{(0)} + \lambda^{(0)}}, \tag{50}$$

$$\hat{\mathsf{w}}^{(t)} = \frac{\hat{m}^{(t)} + \hat{R}^{(t)} \hat{\mathsf{w}}^{(t-1)} + \sqrt{\hat{\chi}^{(t)}}\xi^{(t)}_w}{\hat{Q}^{(t)} + \lambda^{(t)}}, \tag{51}$$

where, $\Theta, \hat{\Theta}$ are determined by the saddle point condition.

Moreover, similar integrals over $w^{(t)}_{c_t}, \tilde{u}^{(t)}$ and $u^{(t)}_{c_t}$ considered above yields the expressions of the saddle point condition under RS assumption as follows:

$$\begin{aligned} q^{(t)} &= \mathbb{E}_{\xi^{(t)}_w \sim_{\mathrm{iid}} \mathcal{N}(0,1)} \left[\mathbb{E}_{\{w^{(t)}\}\sim p_{\mathrm{eff},w}(\{w^{(t)}\}|\{\xi^{(t)}_w\})} \left[w^{(t)} \right]^2 \right] + \mathcal{O}(n), \\ \chi^{(t)} &= \mathbb{E}_{\xi^{(t)}_w \sim_{\mathrm{iid}} \mathcal{N}(0,1)} \left[\beta^{(t)} \left(\mathbb{V}\mathrm{ar}_{\{w^{(t)}\}\sim p_{\mathrm{eff},w}(\{w^{(t)}\}|\{\xi^{(t)}_w\})} \left[w^{(t)} \right] \right) \right] + \mathcal{O}(n), \\ &= \mathbb{E}_{\xi^{(t)}_w \sim_{\mathrm{iid}} \mathcal{N}(0,1)} \left[\frac{\partial}{\partial(\sqrt{\hat{\chi}^{(t)}}\xi^{(t)}_w)} \mathbb{E}_{\{w^{(t)}\}\sim p_{\mathrm{eff},w}(\{w^{(t)}\}|\{\xi^{(t)}_w\})} \left[w^{(t)} \right] \right] + \mathcal{O}(n), \\ R^{(t)} &= \mathbb{E}_{\xi^{(t)}_w \sim_{\mathrm{iid}} \mathcal{N}(0,1)} \left[\mathbb{E}_{\{w^{(t)}\}\sim p_{\mathrm{eff},w}(\{w^{(t)}\}|\{\xi^{(t)}_w\})} \left[w^{(t)} w^{(t-1)} \right] \right] + \mathcal{O}(n), \\ m^{(t)} &= \mathbb{E}_{\xi^{(t)}_w \sim_{\mathrm{iid}} \mathcal{N}(0,1)} \left[\mathbb{E}_{\{w^{(t)}\}\sim p_{\mathrm{eff},w}(\{w^{(t)}\}|\{\xi^{(t)}_w\})} \left[w^{(t)} \right] \right] + \mathcal{O}(n). \end{aligned} \tag{52}$$

For $\hat{\Theta}$, we need to write separately for $t = 0$ and $t \geq 1$. For $t = 0$,

$$\begin{aligned} \hat{Q}^{(0)} &= \alpha_L \Delta_L \mathbb{E}_{\xi\sim\mathcal{N}(0,\Delta_L), y^{(0)}\sim p_{y,L}} \left[\frac{d}{dh^{(0)}_u} \mathbb{E}_{u^{(0)}\sim p^{(0)}_{\mathrm{eff},u|y}} \left[\partial_2 l^{(0)} \right] \right] + \mathcal{O}(n), \\ \hat{\chi}^{(0)} &= \alpha_L \Delta_L \mathbb{E}_{\xi\sim\mathcal{N}(0,\Delta_L), y^{(0)}\sim p_{y,L}} \left[\mathbb{E}_{u^{(0)}\sim p^{(0)}_{\mathrm{eff},u|y}} \left[\partial_2 l^{(0)} \right]^2 \right] + \mathcal{O}(n), \end{aligned}$$

$$\hat{m}^{(0)} = -\alpha_L \mathbb{E}_{\xi\sim\mathcal{N}(0,\Delta_L),y^{(0)}\sim p_{y,L}}\left[(2y^{(0)}-1)\mathbb{E}_{u^{(0)}\sim p^{(0)}_{\text{eff},u|y}}\left[\partial_2 l^{(0)}\right]\right] + \mathcal{O}(n)$$
$$0 = \mathbb{E}_{\xi\sim\mathcal{N}(0,\Delta_L),y^{(0)}\sim p_{y,L}}\left[\mathbb{E}_{u^{(0)}\sim p^{(0)}_{\text{eff},u|y}}\left[\partial_2 l^{(0)}\right]\right] + \mathcal{O}(n),$$

where

$$l^{(0)} = l^{(0)}(y, h_u^{(0)} + u^{(0)}) = l(y, \sigma(h_u^{(0)} + u^{(0)})),$$
$$p^{(0)}_{\text{eff},u|y}(u|y^{(0)}) = \mathcal{N}(u^{(0)} \mid 0, \frac{\chi^{(0)}\Delta_L}{\beta^{(0)}})e^{-\beta^{(0)}l^{(0)}},$$
$$h_u^{(0)} = B^{(0)} + (2y^{(0)}-1)m^{(t)} + \sqrt{\hat{\chi}^{(t)}}\xi_u^{(0)}.$$

For $t \geq 1$,

$$\hat{Q}^{(t)} = \alpha_U^{(t)}\Delta_U^{(t)}\mathbb{E}_{\xi^{(t)}_{u,1},\xi^{(t)}_{u,2}\sim\mathcal{N}(0,\Delta_U)^{(t)},y^{(t)}\sim p_y^{(t)}}\left[\frac{d}{dh_u^{(t)}}\mathbb{E}_{\tilde{u}^{(t)},u^{(t)}\sim p^{(t)}_{\text{eff},u|y}}\left[\partial_2 l^{(t)}\right]\right] + \mathcal{O}(n).$$
$$\hat{\chi}^{(t)} = \alpha_U^{(t)}\Delta_U^{(t)}\mathbb{E}_{\xi^{(t)}_{u,1},\xi^{(t)}_{u,2}\sim\mathcal{N}(0,\Delta_U^{(t)}),y^{(t)}\sim p_y^{(t)}}\left[\mathbb{E}_{\tilde{u}^{(t)},u^{(t)}\sim p^{(t)}_{\text{eff},u|y}}\left[\partial_2 l^{(t)}\right]^2\right] + \mathcal{O}(n).$$
$$\hat{R}^{(t)} = -\alpha_U^{(t)}\Delta_U^{(t)}\mathbb{E}_{\xi^{(t)}_{u,1},\xi^{(t)}_{u,2}\sim\mathcal{N}(0,\Delta_U^{(t)}),y^{(t)}\sim p_y^{(t)}}\left[\frac{d}{d\tilde{h}_u^{(t)}}\mathbb{E}_{\tilde{u}^{(t)},u^{(t)}\sim p^{(t)}_{\text{eff},u|y}}\left[\partial_2 l^{(t)}\right]\right] + \mathcal{O}(n).$$
$$\hat{m}^{(t)} = -\alpha_U^{(t)}\mathbb{E}_{\xi^{(t)}_{u,1},\xi^{(t)}_{u,2}\sim\mathcal{N}(0,\Delta_U^{(t)}),y^{(t)}\sim p_y^{(t)}}\left[(2y^{(t)}-1)\mathbb{E}_{\tilde{u}^{(t)},u^{(t)}\sim p^{(t)}_{\text{eff},u|y}}\left[\partial_2 l^{(t)}\right]\right] + \mathcal{O}(n).$$
$$0 = \mathbb{E}_{\xi^{(t)}_{u,1},\xi^{(t)}_{u,2}\sim\mathcal{N}(0,\Delta_U^{(t)}),y^{(t)}\sim p_y^{(t)}}\left[\mathbb{E}_{\tilde{u}^{(t)},u^{(t)}\sim p^{(t)}_{\text{eff},u|y}}\left[\partial_2 l^{(t)}\right]\right] + \mathcal{O}(n).$$

where

$$l^{(t)} = l^{(t)}(\tilde{h}_u^{(t)} + \tilde{u}^{(t)}, h_u^{(t)} + u^{(t)})$$
$$= \mathbb{1}\left(|\tilde{h}_u^{(t)} + \tilde{u}^{(t)}| > \Gamma\sqrt{q^{(t-1)}}\right) l_{\text{pl}}\left(\sigma_{\text{pl}}(\tilde{h}_u^{(t)} + \tilde{u}^{(t)}), \sigma(h_u^{(t)} + u^{(t)})\right),$$
$$p^{(t)}_{\text{eff},u|y}(u|y^{(t)}) = \mathcal{N}(u^{(0)} \mid 0, \frac{\chi^{(0)}\Delta_L}{\beta^{(0)}})\mathcal{N}(0, \frac{\chi^{(t-1)}\Delta_U^{(t)}}{\beta^{(t-1)}})e^{-\beta^{(t)}l^{(t)}},$$
$$\tilde{h}_u^{(t)} = B^{(t-1)} + (2y^{(t)}-1)m^{(t-1)}\sqrt{q^{(t-1)}}\xi_{u,1}^{(t)},$$
$$h_u^{(t)} = B^{(t)} + (2y^{(t)}-1)m^{(t)} + \frac{R^{(t)}}{\sqrt{q^{(t-1)}}}\xi_{u,1}^{(t)} + \sqrt{q^{(t)} - \frac{(R^{(t)})^2}{q^{(t-1)}}}\xi_{u,2}^{(t)}.$$

At the limit $\beta \to \infty$, in addition to $p_{\text{eff},w}$, $p^{(t)}_{\text{eff},u}$ also concentrates, and the average over $u^{(t)}, \tilde{u}^{(t)}$ are governed by one-dimensional effective problem with Gaussian disorder. Let us define $\hat{\mathfrak{u}}^{(t)}$ as follows:

$$\hat{\mathfrak{u}}^{(0)} = \arg\min_{u^{(0)}\in\mathbb{R}}\left[\frac{(u^{(0)})^2}{2\Delta_L\chi^{(0)}} + l\left(y^{(0)}, \sigma\left(h_u^{(0)} + u^{(0)}\right)\right)\right], \tag{53}$$
$$\hat{\mathfrak{u}}^{(t)} = \arg\min_{u^{(t)}\in\mathbb{R}}\left[\frac{(u^{(t)})^2}{2\Delta_U\chi^{(t)}} + \mathbb{1}(|\tilde{h}^{(t)}| > \Gamma\sqrt{q^{(t-1)}})l_{\text{pl}}\left(\sigma_{\text{pl}}\left(\tilde{h}_u^{(t)}\right), \sigma\left(h_u^{(t)} + u^{(t)}\right)\right)\right]. \tag{54}$$

Also, let us define $l_L(y,x)$ and $l_U(y,x;\tilde{\Gamma})$ as

$$\begin{aligned} l_L(y,x) &= l(y,\sigma(x)),\\ l_U(y,x;\tilde{\Gamma}) &= \mathbb{1}(|y|>\tilde{\Gamma}) l_{\mathrm{pl}}(\sigma_{\mathrm{pl}}(y),\sigma(x)). \end{aligned}$$

Then, the RS saddle point conditions are written at the limit $\beta\to\infty, n\to 0$ are summarized as follows:

$$\begin{aligned} q^{(0)} &= \mathbb{E}_{\xi_w^{(0)}\sim_{\mathrm{iid}}\mathcal{N}(0,1)}\left[(\hat{\mathsf{w}}^{(0)})^2\right],\\ \chi^{(0)} &= \mathbb{E}_{\xi_w^{(0)}\sim_{\mathrm{iid}}\mathcal{N}(0,1)}\left[\frac{\partial}{\partial(\sqrt{\hat{\chi}^{(0)}}\xi_w^{(0)})}\hat{\mathsf{w}}^{(0)}\right],\\ m^{(0)} &= \mathbb{E}_{\xi_w^{(0)}\sim_{\mathrm{iid}}\mathcal{N}(0,1)}\left[\hat{\mathsf{w}}^{(0)}\right],\\ \hat{Q}^{(0)} &= \alpha_L\Delta_L\mathbb{E}_{\xi_u^{(0)}\sim\mathcal{N}(0,\Delta_L),y^{(0)}\sim p_{y,L}}\left[\frac{d}{dh_u^{(0)}}\partial_2 l_L(y^{(0)},h_u^{(0)}+\hat{\mathsf{u}}^{(0)})\right],\\ \hat{\chi}^{(0)} &= \alpha_L\Delta_L\mathbb{E}_{\xi_u^{(0)}\sim\mathcal{N}(0,\Delta_L),y^{(0)}\sim p_{y,L}}\left[\left(\partial_2 l_L(y^{(0)},h_u^{(0)}+\hat{\mathsf{u}}^{(0)})\right)^2\right],\\ \hat{m}^{(0)} &= \alpha_L\mathbb{E}_{\xi_u^{(0)}\sim\mathcal{N}(0,\Delta_L),y^{(0)}\sim p_{y,L}}\left[(2y-1)\partial_2 l_L(y^{(0)},h_u^{(0)}+\hat{\mathsf{u}}^{(0)})\right],\\ 0 &= \mathbb{E}_{\xi_u^{(0)}\sim\mathcal{N}(0,\Delta_L),y^{(0)}\sim p_{y,L}}\left[\partial_2 l_L(y^{(0)},h_u^{(0)}+\hat{\mathsf{u}}^{(0)})\right], \end{aligned}$$

and

$$\begin{aligned} q^{(t)} &= \mathbb{E}_{\xi_w^{(t)}\sim_{\mathrm{iid}}\mathcal{N}(0,1)}\left[(\hat{\mathsf{w}}^{(t)})^2\right],\\ \chi^{(t)} &= \mathbb{E}_{\xi_w^{(t)}\sim_{\mathrm{iid}}\mathcal{N}(0,1)}\left[\frac{\partial}{\partial(\sqrt{\hat{\chi}^{(t)}}\xi_w^{(t)})}\hat{\mathsf{w}}^{(t)}\right],\\ R^{(t)} &= \mathbb{E}_{\xi_w^{(t)}\sim_{\mathrm{iid}}\mathcal{N}(0,1)}\left[\hat{\mathsf{w}}^{(t)}\hat{\mathsf{w}}^{(t-1)}\right],\\ m^{(t)} &= \mathbb{E}_{\xi_w^{(t)}\sim_{\mathrm{iid}}\mathcal{N}(0,1)}\left[\hat{\mathsf{w}}^{(t)}\right],\\ \hat{Q}^{(t)} &= \alpha_U^{(t)}\Delta_U^{(t)}\mathbb{E}_{\xi_{u,1}^{(t)},\xi_{u,2}^{(t)}\sim\mathcal{N}(0,\Delta_U^{(t)}),y^{(t)}\sim p_y^{(t)}}\left[\frac{d}{dh_u^{(t)}}\partial_2 l_U(\tilde{h}_u^{(t)},h_u^{(t)}+\hat{\mathsf{u}}^{(t)};\Gamma\sqrt{q^{(t-1)}})\right],\\ \hat{R}^{(t)} &= -\alpha_U^{(t)}\Delta_U^{(t)}\mathbb{E}_{\xi_{u,1}^{(t)},\xi_{u,2}^{(t)}\sim\mathcal{N}(0,\Delta_U^{(t)}),y^{(t)}\sim p_y^{(t)}}\left[\frac{d}{d\tilde{h}_u^{(t)}}\partial_2 l_U(\tilde{h}_u^{(t)},h_u^{(t)}+\hat{\mathsf{u}}^{(t)};\Gamma\sqrt{q^{(t-1)}})\right]\\ \hat{\chi}^{(t)} &= \alpha_U^{(t)}\Delta_U^{(t)}\mathbb{E}_{\xi_{u,1}^{(t)},\xi_{u,2}^{(t)}\sim\mathcal{N}(0,\Delta_U^{(t)}),y^{(t)}\sim p_y^{(t)}}\left[\left(\partial_2 l_U(\tilde{h}_u^{(t)},h_u^{(t)}+\hat{\mathsf{u}}^{(t)};\Gamma\sqrt{q^{(t-1)}})\right)^2\right],\\ \hat{m}^{(t)} &= \alpha_U^{(t)}\mathbb{E}_{\xi_{u,1}^{(t)},\xi_{u,2}^{(t)}\sim\mathcal{N}(0,\Delta_U^{(t)}),y^{(t)}\sim p_y^{(t)}}\left[(2y^{(t)}-1)\partial_2 l_U(\tilde{h}_u^{(t)},h_u^{(t)}+\hat{\mathsf{u}}^{(t)};\Gamma\sqrt{q^{(t-1)}})\right],\\ 0 &= \mathbb{E}_{\xi_{u,1}^{(t)},\xi_{u,2}^{(t)}\sim\mathcal{N}(0,\Delta_U^{(t)}),y^{(t)}\sim p_y^{(t)}}\left[\partial_2 l_U(\tilde{h}_u^{(t)},h_u^{(t)}+\hat{\mathsf{u}}^{(t)};\Gamma\sqrt{q^{(t-1)}})\right], \end{aligned}$$

for $t\geq 1$. Finally, taking the average over $\xi_w^{(t)}$ yields the self-consistent equation defined in 1. The parameters are determined so that they self-consistently satisfy the saddle-point condition. Hence the above saddle point equations are termed *self-consistent equations.*

A.3 RS Generating Functional

As we have already seen, the generating functional can be written as (49)-(51). Imposing that $\alpha_U^{(t)} = \alpha_U, \Delta_U^{(t)} = \Delta_U, \alpha_U^{(t)} = \alpha_U$ and $\rho_U^{(t)} = \rho_U$ yields the Prediction.

As discussed in the main text, $\{\hat{\mathsf{w}}^{(t)}\}$ in (50)-(51) effectively describes the statistical properties of the weight vectors. Although the weight vectors $\hat{\boldsymbol{w}}^{(t)}$ are the high-dimensional vectors in $\mathbb{R}^N$, the empirical distributions of the component of the vectors $\{\hat{w}_i^{(t)}\}$ are described by the one-dimensional Gaussian process. Similarly, we can consider another generating functional regarding $y_\nu^{(t)}$, $\tilde{u}_\nu = \boldsymbol{x}_\nu^{(t)} \cdot \boldsymbol{w}^{(t-1)}/\sqrt{N} + B^{(t-1)}$ and $u_\nu = \boldsymbol{x}_\nu^{(t)} \cdot \boldsymbol{w}^{(t)}/\sqrt{N} + B^{(t)}$:

$$\Xi_{\mathrm{ST}}(\epsilon_u) = \lim_{N,\beta\to\infty} \mathbb{E}_{\{\boldsymbol{\theta}^{(t)}\}_{t=0}^T \sim p_{\mathrm{ST}}, D}\left[e^{\epsilon_u g_u(\{(y_\nu^{(t)}, \tilde{u}_\nu^{(t)}, u_\nu^{(t)})\}_{t=1}^T)} \right],$$

where $\nu \in [M_U]$ and $t \geq 1$. The computation of this generating functional is completely analogous to those in the case of $\Xi_{\mathrm{ST}}(\epsilon_w, \epsilon_B)$. Specifically, the same form of the replica trick can be used by replacing the factor $e^{\epsilon_w g_w(\{w_{1,i}^{(t)}\}_{t=0}^T) + \epsilon_B g_B(\{B_1^{(t)}\}_{t=0}^T)}$ by $e^{\epsilon_u g_u(\{(y_\nu^{(t)}, \tilde{u}_{1,\nu}^{(t)}, u_{1,\nu}^{(t)})\}_{t=1}^T)}$. Again, the index 1 in the latter expression is the replica index. The following calculation procedures are also the same. Then, the counterpart to equation (45) is obtained as follows:

$$\begin{aligned}
\phi_{n_0,\ldots,n_T}^{(T)} &= \lim_{N,\beta\to\infty} \int e^{N\mathcal{S}_u(\Theta,\hat{\Theta})} \mathbb{E}_{\{y^{(t)}, \boldsymbol{u}^{(t)}\}_{t=1}^T} \left[e^{\epsilon_u g_u(\{(y^{(t)}, \tilde{u}_1^{(t)}, u_1^{(t)})\}_{t=1}^T)} \right] d\Theta d\hat{\Theta}, \\
y^{(t)} &\sim p_y^{(t)}, \boldsymbol{u}^{(t)} \sim \tilde{p}_{\mathrm{eff},u|y}, \\
\mathcal{S}_u(\Theta, \hat{\Theta}) &= \alpha_L \varphi_u^{(0)} + \sum_{t=1}^T (1 - 1/N) \alpha_U^{(t)} \varphi_u^{(t)} + \varphi_w.
\end{aligned}$$

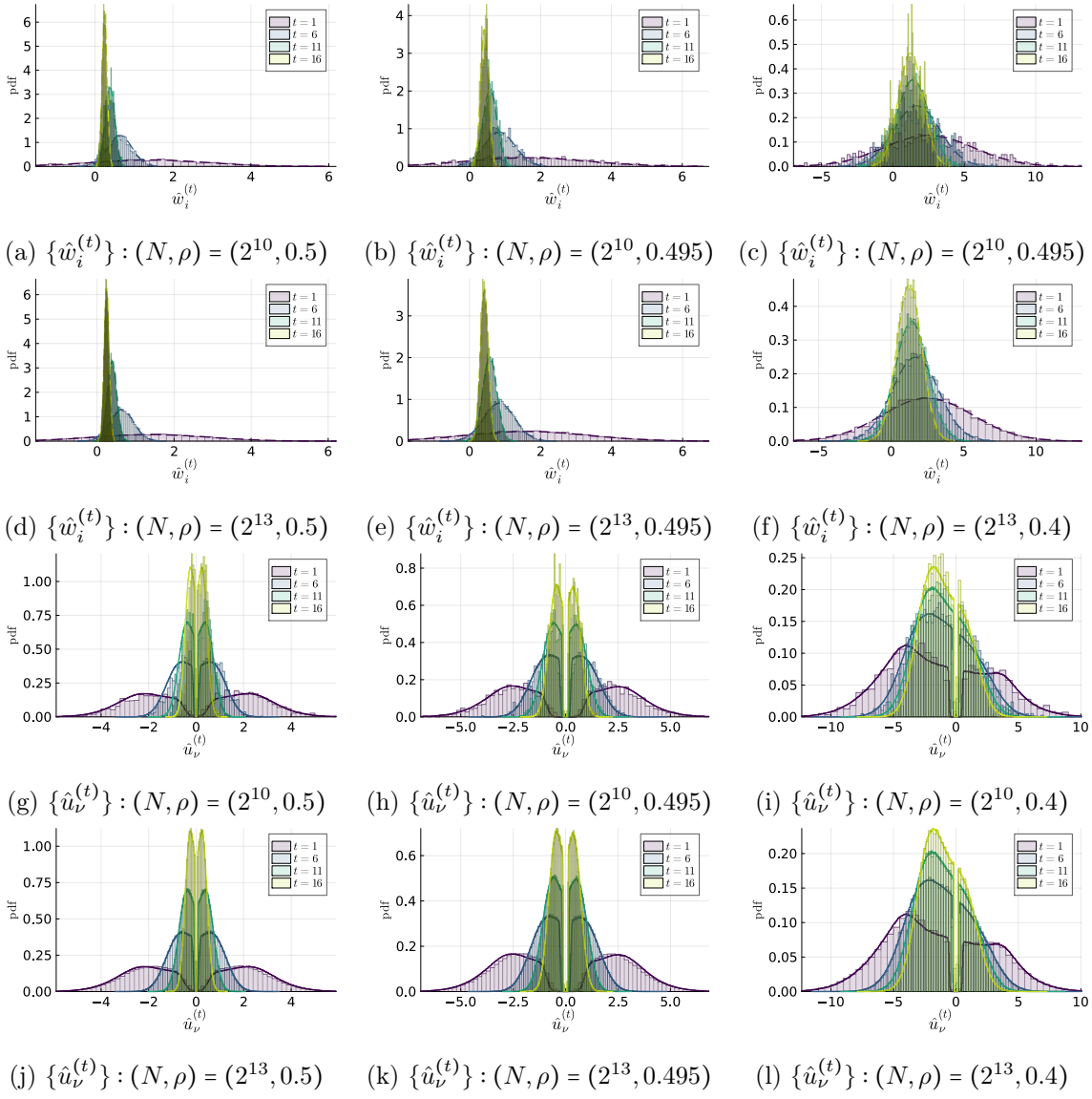
Recall that $\tilde{p}_{\mathrm{eff},u}$ is the density defined in (48). Also, $\mathcal{S}_u$ is equal to $\mathcal{S}$ when $N \gg 1$. The integral over $\Theta, \hat{\Theta}$ can be approximated by the saddle point at $N \gg 1$, which yields the same saddle point condition. After that, one can obtain the following result:

$$\Xi_{\mathrm{ST}}(\tilde{\epsilon}_u, \epsilon_u) = \mathbb{E}_{\xi_u^{(0)}, \xi_{u,1}^{(t)}, \xi_{u,2}^{(t)} \sim_{\mathrm{iid}} \mathcal{N}(0, \Delta_U^{(t)}), y^{(t)} \sim p_y^{(t)}} \left[e^{\tilde{\epsilon}_u g_{\tilde{u}}(\{\tilde{h}_u^{(t)}\}_{t=1}^T) + \epsilon_u g_u(\{h_u^{(t)} + \hat{\mathsf{u}}^{(t)}\}_{t=0}^T)} \right].$$

By a similar argument to that for $\Xi_{\mathrm{ST}}(\epsilon_w, \epsilon_B)$, it can be understood that $(\tilde{h}_u^{(t)}, \hat{\mathsf{u}}^{(t)} + h_u^{(t)})$ effectively describes the statistical properties of the logits. It effectively describes the statistical properties of the empirical distributions of the logits. The minimization problems (53)-(54) can be understood as effective one-dimensional problems to determine the logits. There, $\xi_{u,1}^{(t)}$ and $\xi_{u,2}^{(t)}$ effectively play the role of $D_U^{(t)}$.

Appendix B. Additional Finite Size Experiments

In this appendix, we provide additional comparisons between the RS predictions and finite-size numerical experiments. The main text shows representative examples in Figures 1 and 2. Here, we report the remaining plots for other values of the label bias ρ, system size N, and macroscopic quantities. These additional results support the same conclusion as in the main text: the RS predictions accurately capture both the empirical distributions and the macroscopic order parameters in the finite-size simulations.



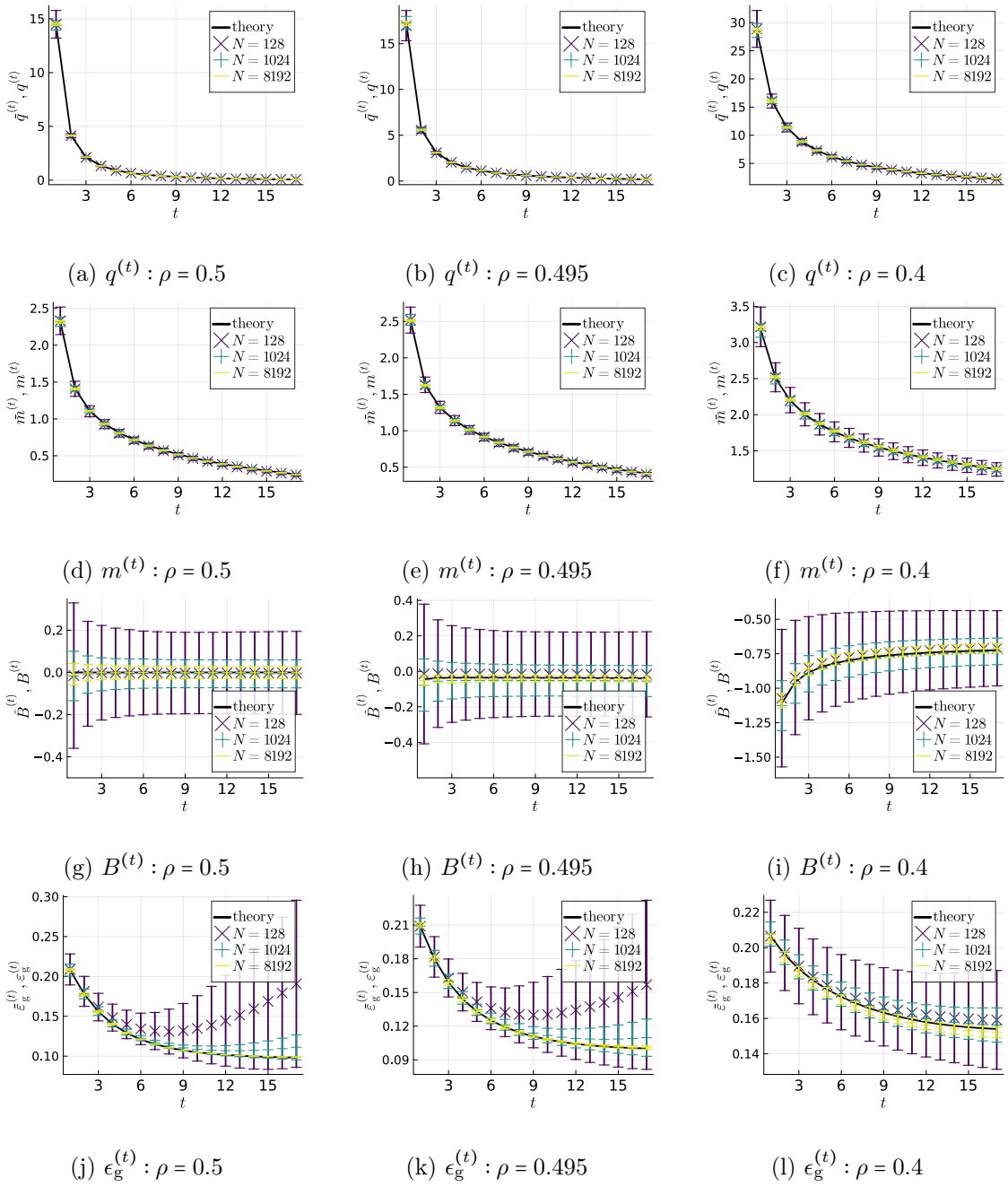
(a) $\{\hat{w}_i^{(t)}\} : (N, \rho) = (2^{10}, 0.5)$ (b) $\{\hat{w}_i^{(t)}\} : (N, \rho) = (2^{10}, 0.495)$ (c) $\{\hat{w}_i^{(t)}\} : (N, \rho) = (2^{10}, 0.495)$

(d) $\{\hat{w}_i^{(t)}\} : (N, \rho) = (2^{13}, 0.5)$ (e) $\{\hat{w}_i^{(t)}\} : (N, \rho) = (2^{13}, 0.495)$ (f) $\{\hat{w}_i^{(t)}\} : (N, \rho) = (2^{13}, 0.4)$

(g) $\{\hat{u}_\nu^{(t)}\} : (N, \rho) = (2^{10}, 0.5)$ (h) $\{\hat{u}_\nu^{(t)}\} : (N, \rho) = (2^{10}, 0.495)$ (i) $\{\hat{u}_\nu^{(t)}\} : (N, \rho) = (2^{10}, 0.4)$

(j) $\{\hat{u}_\nu^{(t)}\} : (N, \rho) = (2^{13}, 0.5)$ (k) $\{\hat{u}_\nu^{(t)}\} : (N, \rho) = (2^{13}, 0.495)$ (l) $\{\hat{u}_\nu^{(t)}\} : (N, \rho) = (2^{13}, 0.4)$

Figure 11: (a)-(f): Comparison between the empirical distribution of the elements of $\hat{\boldsymbol{w}}^{(t)}$ (histogram), which is obtained by single-shot numerical experiment of finite size, and the theoretical prediction given by the Gaussian process $\hat{\mathsf{w}}^{(t)}$ in (27) (solid line). (g)-(l): Comparison between the empirical distribution of the elements of $\{\hat{u}_\nu^{(t)} = \hat{\boldsymbol{w}}^{(t)} \cdot \boldsymbol{x}_\nu^{(t)} / \sqrt{N} + \hat{B}^{(t)}\}_{\nu=1}^{M_U}$ (histogram), which is obtained by single-shot numerical experiment of finite size, and the theoretical prediction given by $\hat{\mathsf{u}}^{(t)} + h_u^{(t)}$ in (18) and (35) (solid line). Different colors represent different iteration steps of ST with total number of iterations $T = 16$. For the details of the settings, refer to the main text.



(a) $q^{(t)} : \rho = 0.5$ (b) $q^{(t)} : \rho = 0.495$ (c) $q^{(t)} : \rho = 0.4$

(d) $m^{(t)} : \rho = 0.5$ (e) $m^{(t)} : \rho = 0.495$ (f) $m^{(t)} : \rho = 0.4$

(g) $B^{(t)} : \rho = 0.5$ (h) $B^{(t)} : \rho = 0.495$ (i) $B^{(t)} : \rho = 0.4$

(j) $\epsilon_{\rm g}^{(t)} : \rho = 0.5$ (k) $\epsilon_{\rm g}^{(t)} : \rho = 0.495$ (l) $\epsilon_{\rm g}^{(t)} : \rho = 0.4$

Figure 12: Comparison of the macroscopic quantities obtained by experiments of finite-size systems (markers with error bars) and the theoretical prediction (black solid line). The error bars represent standard deviations. (a)-(c): The squared norm of the weight vector: $\|\hat{\boldsymbol{w}}^{(t)}\|_2^2/N$ and $q^{(t)}$ in (29). (d)-(f): The inner product between the cluster center and the weight vector: $\boldsymbol{v} \cdot \hat{\boldsymbol{w}}^{(t)}/N$ and $m^{(t)}$ in (30). (g)-(i): The bias: $\hat{B}^{(t)}$ and $B^{(t)}$ in (32). (j)-(l): Generalization error: $\epsilon_{\rm g}^{(t)}$ defined in (5) and that in (33).

Distribution of the weights and logits: The panels (a)-(f) in Figure 11 show the comparison between the empirical distribution of the elements of the weight vector $\{\hat{w}_i^{(t)}\}_{i=1}^N$, obtained by the numerical experiments of finite size systems, and the theoretical prediction given by the Gaussian process $\hat{\mathsf{w}}^{(t)}$ in (27). Each panel shows the result of different label bias ρ and the system size N. Also, the panels (g)-(l) in Figure 11 show the comparison between the empirical distribution of the logits $\{\hat{u}_\nu^{(t)} = \hat{\boldsymbol{w}}^{(t)} \cdot \boldsymbol{x}_\nu/\sqrt{N} + \hat{B}^{(t)}\}$ and the theoretical prediction given by $\hat{\mathsf{u}}^{(t)} + h_u^{(t)}$ in (18) and (35). Since the logits on the data points excluded by PLS are not determined, the contribution from these points is excluded from the figures. For numerical experiments, two different system sizes $N = 1024 (= 2^{10})$ and $8192 (= 2^{13})$ are used. The figure shows that the empirical distributions of both the weight vector and the logits at each iteration step are in good agreement even at $N = 1024$. Although the results of $N = 1024$ have a slight fluctuation due to the lack of sample size, the results of $N = 8192$ show almost complete agreement with the RS solution.

Macroscopic quantities: Figure 12 presents the results of a comparison between experiment of finite size systems and the theoretical predictions for macroscopic quantities such as the norm $\|\hat{\boldsymbol{w}}^{(t)}\|_2^2/N$, the inner product with the cluster center $\boldsymbol{v} \cdot \hat{\boldsymbol{w}}^{(t)}/N$, the bias $\hat{B}^{(t)}$, and the generalization error (5). Each panel shows the result of different label bias ρ. For numerical experiments, three different system sizes $N = 128, 1024$, and 8192 are used. Reported experimental results are averaged over several realization of D depending on the size N. The error bars represent standard deviations. The panels of $q^{(t)}, m^{(t)}$ and $B^{(t)}$ show that the experimental results and RS solutions are in good agreement even at $N = 1024$. Also, the standard deviation decreases as the system size grows, indicating the concentration of these quantities. However, in the panels of generalization error, there is still a discrepancy between experiment and theory at $N = 1024$. This is because $q^{(t)}$ becomes very small when t is large, causing even slight fluctuations of $\bar{q}^{(t)}$ and $\hat{B}^{(t)}$ to lead to significant deviations in the generalization error that depends on these quantities in the form of $\hat{B}^{(t)}/\sqrt{\bar{q}^{(t)}}$ as in (6). Hence a large system size such as $N = 8192$ is required to obtain a good agreement in the generalization error.

Appendix C. Increase of the Linear Susceptibility in Small Unlabeled Data Scenario

In this appendix, we complement the observations in subsection 4.1, which shows that the ST does not perform better when α_U is small. Figure 13 shows that as α_U decreases, the linear susceptibility $\chi^{(T)}$ becomes greater at large iteration T, which indicates that the training result is getting unstable against a small perturbation to the loss function. These observations indicate that ST is unstable when it repeatedly fits to pseudo-labels in overparameterized settings.

Appendix D. Derivation of Prediction 5

In this section, we outline the derivation of Prediction 5. Also, let $\lambda_U^{(t)}, \alpha_U^{(t)}, \rho_U^{(t)}$ and $\Delta_U^{(t)}$ be the values of the regularization parameter, the relative size of the unlabeled data, and the label imbalance and the size of the cluster at step $t \geq 1$. Although we mainly focus on the

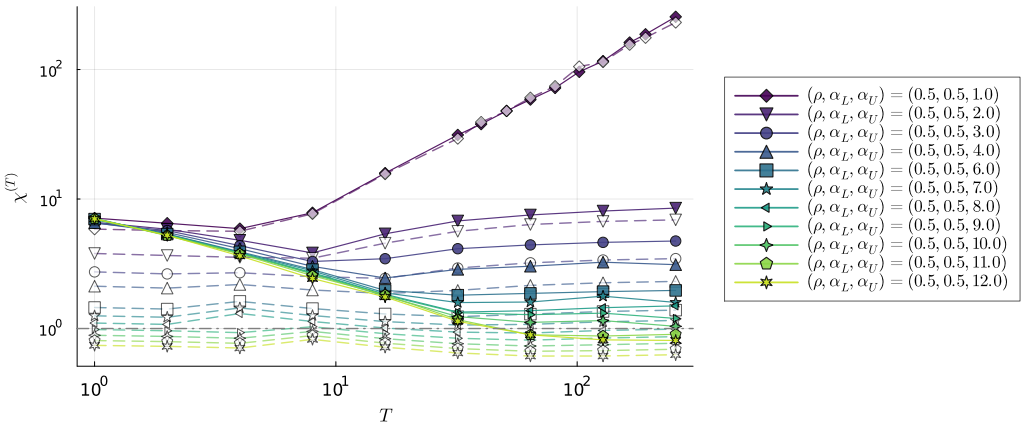


Figure 13: The behavior of the quantity $\chi^{(T)}$ when $(\rho, \Delta) = (1/2, 0.75^2)$. Different colors represent different α_U.

case of $\lambda_U^{(t)} = \lambda_U, \alpha_U^{(t)} = \alpha_U, \rho_U^{(t)} = \rho_U$ and $\Delta_U^{(t)} = \Delta_U$ in the main text paper, the following derivation allows these step-dependent cases.

The basic strategy is to expand $\Theta^{(t)}$ and $\hat{\Theta}$ as $\Theta^{(t)} = \Theta_0^{(t)} + \Theta_1^{(t)}\lambda_U^{(t)} + \dots$, and $\hat{\Theta}^{(t)} = \hat{\Theta}_0^{(t)} + \hat{\Theta}_1^{(t)}\lambda_U^{(t)} + \dots$, respectively, and then to consider the self-consistent equations order by order in λ_U. In the following, lower subscript represents the order of the expansion. For example, $q_0^{(t)}$ is the zero-th order term of $q^{(t)}$.

Before going into the details, we make two remarks here. The first point is regarding the sign of $\chi^{(t)}$. Since $\chi^{(t)}$ is defined as a limit of the variance as in (52), it is positive:

$$\chi^{(t)} > 0. \tag{55}$$

The second point is about $\hat{\mathsf{u}}^{(t)}$. Due to the convexity of the loss functions, $\hat{\mathsf{u}}^{(t)}, t \geq 1$ in (18) is determined by the following extreme condition:

$$\frac{\hat{\mathsf{u}}^{(t)}}{\chi^{(t)}\Delta_U^{(t)}} + \partial_2 \tilde{l}(\tilde{h}_u^{(t)}, h_u^{(t)} + \hat{\mathsf{u}}^{(t)}) = 0. \tag{56}$$

D.1 Zero-th Order

First, Prediction 2 and Assumption 2 immediately imply the following:

$$q_0^{(t)} = q^{(t-1)}, R_0^{(t)} = q^{(t-1)}, m_0^{(t)} = m^{(t-1)}. \tag{57}$$

Inserting these conditions into the expression of $h_u^{(t)}$, it is shown that the lowest order of $h_u^{(t)}$ is equal to $\tilde{h}_u^{(t)}$:

$$h_{u,0}^{(t)} = \tilde{h}_u^{(t)}.$$

Then, the zero-th order of the condition (56) for determining $\hat{\mathsf{u}}^{(t)}$ is given as follows

$$\frac{\hat{\mathsf{u}}_0^{(t)}}{\chi_0^{(t)}\Delta_U^{(t)}} + \partial_2 \tilde{l}(\tilde{h}_{u,t}, \tilde{h}_{u,t} + \hat{\mathsf{u}}_0^{(t)}) = 0.$$

From (36) in Assumption 1, this equation is satisfied trivially when $\hat{\mathfrak{u}}_0^{(t)} = 0$. Thus, it is concluded that the zero-th order of $\hat{\mathfrak{u}}_0^{(t)}$ vanishes:

$$\hat{\mathfrak{u}}_0^{(t)} = 0.$$

By inserting this condition into the self-consistent equations (24) and (23), it is shown that the zero-th order terms of $\hat{m}^{(t)}$ and $\hat{\chi}^{(t)}$ also vanish:

$$\hat{m}_0^{(t)} = 0, \tag{58}$$

$$\hat{\chi}_0^{(t)} = 0. \tag{59}$$

On the other hand, the zero-th order of the self-consistent equations for $q^{(t)}, m^{(t)}$ and $R^{(t)}$ are given as follows:

$$\begin{aligned} q_0^{(t)} &= \left(\frac{\hat{R}_0^{(t)}}{\hat{Q}_0^{(t)}}\right)^2 q^{(t-1)}, \\ m_0^{(t)} &= \frac{\hat{R}_0^{(t)}}{\hat{Q}_0^{(t)}} m^{(t-1)}, \\ R_0^{(t)} &= \frac{\hat{R}_0^{(t)}}{\hat{Q}_0^{(t)}} q^{(t-1)}. \end{aligned}$$

Because the condition (57) should be satisfied self-consistently, the zero-th order of $\hat{Q}^{(t)}$ and $\hat{R}^{(t)}$ are equal:

$$\hat{Q}_0^{(t)} = \hat{R}_0^{(t)}.$$

Finally, the zero-th order expression of $\chi^{(t)}$ yields the condition

$$\chi_0^{(t)} = \frac{1}{\hat{Q}_0^{(t)}}.$$

Since $\chi_0^{(t)}$ is a value of $\chi^{(t)}$ at $\lambda_U = 0$, it is non-negative from (55). Hence $\hat{Q}_0^{(t)}$ is also non-negative.

D.2 First Order

First, we show that $h_u^{(t)}$ does not contain non-integer order terms at least up to the first order, i.e., it can be expanded as $h_u^{(t)} = h_{u,0}^{(t)} + h_{u,1}^{(t)} \lambda_U^{(t)} + \dots$. By inserting (58) and (59) into (21) and (22), $q_1^{(t)} - 2R_1^{(t)} = \hat{\chi}_1^{(t)} / (\hat{Q}_0^{(t)})^2$ is obtained. Since $h_u^{(t)}$ can be written as

$$\begin{aligned} h_u^{(t)} = \tilde{h}_u^{(t)} &+ \left(B_1^{(t)} + (2y^{(t)} - 1) m_1^{(t)} + \frac{R_1^{(t)}}{\sqrt{q^{(t-1)}}} \xi_{u,1}^{(t)} \right) \lambda_U^{(t)} \\ &+ \sqrt{(q_1^{(t)} - 2R_1^{(t)}) \lambda_U^{(t)} + \mathcal{O}((\lambda_U^{(t)})^2)} \xi_{u,2}^{(t)} + \mathcal{O}((\lambda_U^{(t)})^2), \end{aligned} \tag{60}$$

$\hat{\mathsf{u}}^{(t)}$ that is determined by the condition (56) may be expanded as

$$\hat{\mathsf{u}}^{(t)} = \hat{\mathsf{u}}^{(t)}_{1/2}(\lambda_U^{(t)})^{1/2} + \hat{\mathsf{u}}^{(t)}_1 \lambda_U + \dots .$$

However, by inserting this expression and (60) into the condition (56) and expanding by λ_U, one can show that $\hat{\mathsf{u}}^{(t)}_{1/2} \propto \sqrt{\hat{\chi}_1^{(t)}}$. This implies that the RHS of (23) is also proportional to $\hat{\chi}_1^{(t)}$ at the first order of $\lambda_U^{(t)}$, which leads to the condition $\hat{\chi}_1^{(t)} \propto \hat{\chi}_1^{(t)}$. Thus, we obtain the following:

$$\hat{\chi}_0^{(t)} = \hat{\chi}_1^{(t)} = 0. \tag{61}$$

From this, it is evident that $h_u^{(t)}$ can be expanded as $h_u^{(t)} = h_{u,0}^{(t)} + h_{u,1}^{(t)} \lambda_U^{(t)} + \dots$, where

$$h_{u,1}^{(t)} = B_1^{(t)} + (2y^{(t)} - 1)m_1^{(t)} + \frac{R_1^{(t)}}{\sqrt{q^{(t-1)}}}\xi_{u,1}^{(t)} + \sqrt{q_2^{(t)} - \frac{(R_1^{(t)})^2}{q^{(t-1)}} - R_2^{(t)}}\,\xi_{u,2}^{(t)}.$$

From the above result, it also follows that the lowest order of the expansion of $\hat{\mathsf{u}}^{(t)}$ is the first order: $\hat{\mathsf{u}}^{(t)} = \hat{\mathsf{u}}^{(t)}_1 \lambda_U^{(t)} + \dots$.

The consequence of the above result is that the contributions from the noise terms $\sqrt{\hat{\chi}^{(t)}}\xi_w^{(t)}$ in (28) and $\sqrt{q^{(t)} - \frac{(R^{(t)})^2}{q^{(t-1)}}}\xi_{u,2}^{(t)}$ in (35) are higher order in $\lambda_U^{(t)}$ because these terms appear only in the form of a square (second order in $\lambda_U^{(t)}$) or the raw average (equals to zero).

Appendix E. Derivation of Prediction 6

The settings are the same with Appendix D. Let us start from the first order of the perturbative expansion considered in Appendix D. Let us consider the self-consistent equations determining $q^{(t)}, R^{(t)}, m^{(t)}, B^{(t)}$. Inserting (58) and (61) into (21)-(22), we obtain the first order expansion of $m^{(t)}, R^{(t)}$ and $q^{(t)}$ as follows:

$$m_1^{(t)} = \frac{1}{\hat{Q}_0^{(t)}}\left(\hat{m}_1^{(t)} - (1 - \hat{\delta}_1^{(t)})m^{(t-1)}\right), \tag{62}$$

$$R_1^{(t)} = \frac{1}{\hat{Q}_0^{(t)}}\left(\hat{m}_1^{(t)} m^{(t-1)} - (1 - \hat{\delta}_1^{(t)})q^{(t-1)}\right),$$

$$q_1^{(t)} = 2R_1^{(t)}, \tag{63}$$

$$\hat{\delta}_1^{(t)} = \hat{R}_1^{(t)} - \hat{Q}_1^{(t)},$$

Importantly, the noise term

$$\sqrt{q_2^{(t)} - \frac{(R_1^{(t)})^2}{q^{(t-1)}} - R_2^{(t)}}\,\xi_{u,2}^{(t)},$$

in $h_u^{(t)}$ does not contribute to the above result at all because this term appears only in the form of a square (higher order in $\lambda_U^{(t)}$) or raw average (equals to zero). Therefore, combined

with the result of $\hat{\chi}^{(t)}$ in (61), the update of the order parameters are intrinsically noiseless up to the first order of $\lambda_U^{(t)}$.

By combining equations (62)-(63), the next equations are obtained:

$$M^{(t)} = M^{(t-1)} + M_1^{(t)}\lambda_U^{(t)} + \dots,$$
$$M_1^{(t)} = \underbrace{\frac{2}{\hat{Q}_0^{(t)}}\frac{\hat{m}_1^{(t)}}{m^{(t-1)}}}_{\equiv C^{(t)}} M^{(t-1)}(1 - M^{(t-1)}).$$

Note that $\hat{Q}_0^{(t)} = 1/\chi_0^{(t)} > 0$ as already described above. Since $\hat{m}_1^{(t)}$ is the leading order of $\hat{m}^{(t)}$ that represents the signal component of $\hat{\mathsf{w}}^{(t)}$ accumulated at step t as depicted in (28), $C^{(t)}$ is positive if the training yields a positive accumulation of the signal component to $\hat{\mathsf{w}}^{(t)}$ when $m^{(t-1)} = \boldsymbol{v} \cdot \hat{\boldsymbol{w}}^{(t-1)}/N > 0$, i.e., the classification plane at step $t-1$ is positively correlated with the cluster center, which seems to naturally hold when using a legitimate loss function. This is a natural result of the elimination of noise contributions as seen above.

We also remark that the condition $0 = \mathbb{E}[\hat{\mathsf{u}}^{(t)}]$ that determines $B^{(t)}$ is not used to derive the above result.

Appendix F. Derivation of Prediction 8

In this section, we outline the derivation of Prediction 8. As in Appendix E, let $\lambda_U^{(t)}, \alpha_U^{(t)}, \rho_U^{(t)}$ and $\Delta_U^{(t)}$ be the values of the regularization parameter, the relative size of the unlabeled data, and the label imbalance and the size of the cluster at step $t \geq 1$, although we mainly focus on the case of $\lambda_U^{(t)} = \lambda_U, \alpha_U^{(t)} = \alpha_U, \rho_U^{(t)} = \rho_U$ and $\Delta_U^{(t)} = \Delta_U$ in the main text. Dropping the time indices yields Prediction 8.

When $\tilde{l}_{\tilde{\Gamma}}(y, x) = 1/2(y - x)^2$, we can obtain an explicit expression of $\hat{\mathsf{u}}^{(t)}$ as follows:

$$\hat{\mathsf{u}}^{(t)} = \frac{\tilde{h}_u^{(t)} - h_u^{(t)}}{\frac{1}{\chi^{(t)}\Delta_U^{(t)}} + 1}.$$

From this, the self-consistent equations that determine $\hat{Q}^{(t)}, \hat{R}^{(t)}, \hat{m}^{(t)}, B^{(t)}$ also have the following explicit expressions:

$$\hat{Q}^{(t)} = \hat{R}^{(t)} = \frac{\alpha_U^{(t)}\Delta_U^{(t)}}{1 + \chi^{(t)}\Delta_U^{(t)}},$$
$$\hat{m}^{(t)} = \frac{\alpha_U^{(t)}\Delta_U^{(t)}}{1 + \chi^{(t)}\Delta_U^{(t)}}\left((2\rho_U^{(t)} - 1)(B^{(t-1)} - B^{(t)}) + (m^{(t-1)} - m^{(t)})\right),$$
$$B^{(t)} = B^{(t-1)} + (2\rho_U^{(t)} - 1)(m^{(t-1)} - m^{(t)}).$$

Since $\hat{Q}^{(t)}$ does not involve average, the expansion coefficient of $\hat{Q}^{(t)}$ and $\chi^{(t)}$ at the zero-th order can be explicitly solved as

$$\hat{Q}_0^{(t)} = \Delta_U^{(t)}(\alpha_U^{(t)} - 1),$$
$$\chi_0^{(t)} = \frac{1}{\Delta_U^{(t)}(\alpha_U^{(t)} - 1)}.$$

By using the above explicit expressions, similar perturbative analysis made in Appendix E yields the expansion coefficients at the first order as follows:

$$m_1^{(t)} = -\frac{1}{\Delta_U^{(t)} + V_U^{(t)}} \frac{1}{\alpha_U^{(t)} - 1} m^{(t-1)},$$
$$M_1^{(t)} = C^{(t)} M^{(t-1)}(1 - M^{(t-1)}), \quad C^{(t)} = \frac{2}{\alpha_U^{(t)} - 1} \frac{V_U^{(t)}}{\Delta_U^{(t)} + V_U^{(t)}},$$
$$B_1^{(t)} = -(2\rho_U^{(t)} - 1) m_1^{(t)}.$$

Solving these equations at the continuum limit yields Prediction 8.

Since an explicit expression of $\hat{m}_1^{(t)}$ as $\hat{m}_1^{(t)} = V_U^{(t)}/(\Delta_U^{(t)} + V_U^{(t)}) m^{(t-1)}$ is obtained, the time evolution of $q^{(t)}$ can also be derived, which indicates the exponential shrinkage of the norm directly. However, the resultant expression is rather complex.

Appendix G. Numerical Procedure for Evaluating RS Saddle Point

In this section, we sketch the numerical treatment of the self-consistent equations in Definition 1. Recall that $\Theta^{(t)}$ and $\hat{\Theta}^{(t)}$ are defined as follows:

$$\Theta^{(0)} = (q^{(0)}, \chi^{(0)}, m^{(0)}, B^0),$$
$$\hat{\Theta}^{(0)} = (\hat{Q}^{(0)}, \hat{\chi}^{(0)}, \hat{m}^{(0)}),$$
$$\Theta^{(t)} = (q^{(t)}, \chi^{(t)}, R^{(t)}, m^{(t)}, B^t),$$
$$\hat{\Theta}^{(t)} = (\hat{Q}^{(t)}, \hat{\chi}^{(t)}, \hat{R}^{(t)}, \hat{m}^{(t)}).$$

G.1 Solving the Self-Consistent Equations

To solve the self-consistent equations in Definition 1, the fixed-point iteration algorithm is used. Since the self-consistent equations are already written in a form of $(\Theta, \hat{\Theta}) = \mathcal{F}(\Theta, \hat{\Theta})$, the definition of the fixed-point iteration is straightforward. In particular, $(\Theta^{(t)}, \hat{\Theta}^{(t)})$ only depends on the parameters at t and $t-1$, the equations are solved successively. Let us define the right-hand side of the self-consistent equations at step t as $\Theta^{(t)} = \mathcal{F}^{(t)}(\Theta^{(t-1)}, \hat{\Theta}^{(t)})$ and $\hat{\Theta}^{(t)} = \hat{\mathcal{F}}^{(t)}(\Theta^{(t)})$. Then our fixed-point iteration can be summarized as in Algorithm 1. For evaluating the integral in updating $\hat{\Theta}^{(t)}$, the Gauss-Hermite quadrature is used when $\Gamma = 0$ and the Monte-Carlo integral is used in other cases. In Monte-Carlo integral, the samples of size $10^5 - 10^7$ are used depending on the cases. For solving the nonlinear equations in (53) and (54), the Newton method is used. Note that the derivative in (25) includes the

Algorithm 1 Fixed-point iteration of solving the self-consistent equations

Require: The damping parameter $\eta_{\rm d} \in (0,1)$, the convergence criterion $\epsilon_{\rm tol}$ and the maximum number of iterations $s_{\max}$.

// solving for $t = 0$
Select initial value $\hat{\Theta}_0^{(0)}$ of $\hat{\Theta}^{(0)}$.
$\Theta_0^{(0)} \leftarrow \mathcal{F}^{(0)}(\hat{\Theta}_0^{(0)})$
for $s = 1, 2, \ldots, s_{\max}$ **do**
  $\Theta_s^{(0)} \leftarrow \mathcal{F}^{(0)}(\hat{\Theta}_{s-1}^{(0)})$
  $\hat{\Theta}_s^{(0)} \leftarrow (1-\eta_{\rm d})\hat{\Theta}_{s-1}^{(0)} + \eta_{\rm d}\hat{\mathcal{F}}^{(0)}(\Theta_{s-1}^{(0)})$
  if $\max|\Theta_s^{(0)} - \Theta_{s-1}^{(0)}| < \epsilon_{\rm tol}$ **then**
    $s \leftarrow s_{\max}$
    break
  end if
end for
// solving for $t > 1$
for $t = 1, \ldots, T$ **do**
  $(\Theta_0^{(t)}, \hat{\Theta}_0^{(t)}) \leftarrow (\Theta_{s_{\max}}^{(t-1)}, \hat{\Theta}_{s_{\max}}^{(t-1)})$
  for $s = 1, 2, \ldots, s_{\max}$ **do**
    $\Theta_s^{(t)} \leftarrow \mathcal{F}^{(t)}(\Theta_{s-1}^{(t-1)}, \hat{\Theta}_{s-1}^{(t)})$
    $\hat{\Theta}_s^{(t)} \leftarrow (1-\eta_{\rm d})\hat{\Theta}_{s-1}^{(t)} + \eta_{\rm d}\hat{\mathcal{F}}^{(t)}(\Theta_{s-1}^{(t)})$
    if $\max|\Theta_s^{(t)} - \Theta_{s-1}^{(t)}| < \epsilon_{\rm tol}$ **then**
      $s \leftarrow s_{\max}$
      break
    end if
  end for
end for

derivative of the indicator function $\mathbb{1}(|\tilde{h}_u^{(t)}| > \Gamma\sqrt{q^{(t-1)}})$, which yields the delta functions at $\tilde{h}_u^{(t)} = \pm\Gamma\sqrt{q^{(t-1)}}$. Although including this contribution by Gauss-Hermite or Monte-Carlo integral is difficult, we can easily include it by hand because it only requires the values of the integrand at two points, i.e, the boundaries of the indicator.

We remark that the fixed-point iteration of the self-consistent equations of the type considered above may be closely related to the efficient approximate inference algorithms. Indeed, when considering a simple empirical risk minimization or the Bayesian inference in linear models, it corresponds to the state-evolution formula of the approximate message passing (AMP) algorithm (Zdeborová and Krzakala, 2016; Takahashi and Kabashima, 2020, 2022), which is an efficient algorithm with fast convergence. Hence we expect that the above apparently naive fixed-point iteration yields a moderately good convergence.

G.2 Optimization of the Regularization Parameter

Since the generalization error can be numerically evaluated for each value of the hyper parameters by solving the self-consistent equation using Algorithm 1, we can treat the

generalization error (33) as a function of the hyper parameters. However, the explicit form of such dependence is unknown. Thus, to optimize the generalization error, the Nelder-Mead algorithm, which is a black-box optimization algorithm, implemented in Julia language (Mogensen and Riseth, 2018) is used. By solving the self-consistent equation at each step of optimization, it can be straightforwardly implemented.

Appendix H. Pathological Finite Size Effect at Large Regularization

As already reported in the previous literature (Dobriban and Wager, 2018; Mignacco et al., 2020a), the ridge-regularized logistic regression with true labels yields the Bayes-optimal classifier when $\rho = 1/2$ and the infinitely large regularization parameter is used. However, this result is valid only in the asymptotic regime, and it is known that the experimental results of finite-size systems significantly deviate from theoretical predictions as reported in (Mignacco et al., 2020a). This is because the norm of the weight vector becomes extremely small for large regularization parameters, making the numerical computation challenging and even small fluctuations can lead to catastrophic results. Therefore, predictions from sharp asymptotics may be irrelevant to practical results if we unbound the regularization parameters in our ST as well.

To check this point, we conducted preliminary experiments. For simplicity, we consider the single-shot case with $T = 1$. In this case, we can find the optimal regularization parameters $(\lambda_*^{(0)}, \lambda_*^{(1)}) = (\lambda_L^*, \lambda_U^*)$ that minimize the generalization error $\epsilon_{\rm g}^{(1)}$ by a brute force optimization since numerically evaluating the generalization is easy. We optimize the regularization parameters in $(0, 10^4]^2$ and observed the behavior of the optimal regularization parameters and the generalization error $\epsilon_{\rm g}^{(1)}$. Here $\Delta_U = \Delta_L = \Delta, \rho_U = \rho_L = \rho$.

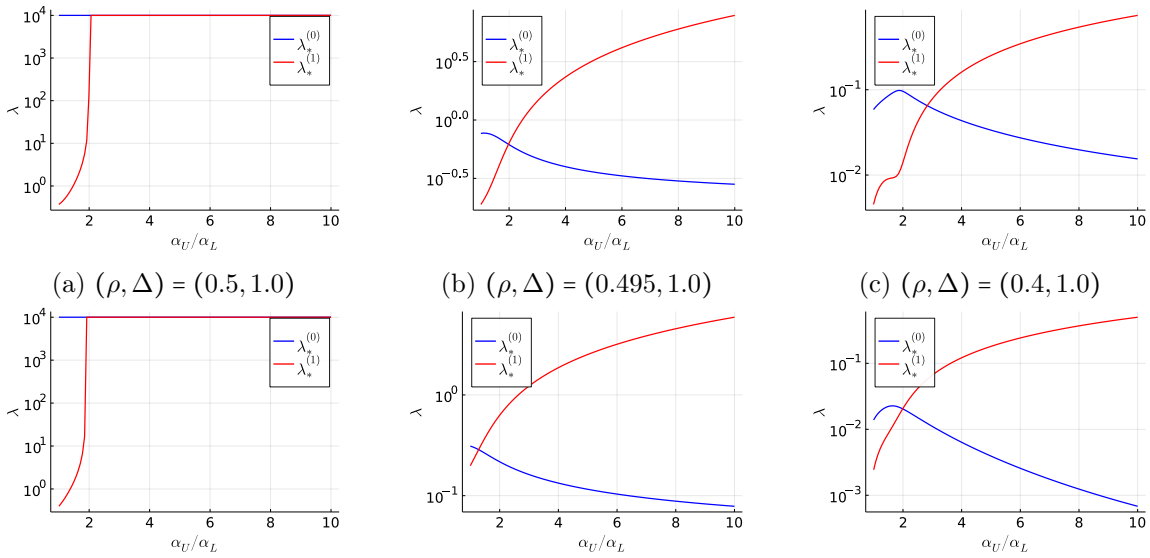
Figure 14 summarizes how the optimal regularization parameters $(\lambda_*^{(0)}, \lambda_*^{(1)})$, which minimize the generalization error $\epsilon_{\rm g}^{(1)}$, depend on the size of the unlabeled dataset α_U, the relative size of clusters ρ and the size of each cluster Δ. Each panel corresponds to different values of ρ and Δ. When $\rho = 1/2$, both $\lambda_*^{(0)}$ and $\lambda_*^{(1)}$ show diverging tendency as expected. Figure 15 shows the comparison with the experiments of the regularization parameter found in the above procedure. Similarly to the previous studies, the experimental results severely deviate from the prediction of the asymptotic result. Hence, we need to limit the range of the regularization parameters as in the main text.

References

Elisabeth Agoritsas, Giulio Biroli, Pierfrancesco Urbani, and Francesco Zamponi. Out-of-equilibrium dynamical mean-field equations for the perceptron model. Journal of Physics A: Mathematical and Theoretical, 51(8):085002, jan 2018. doi: 10.1088/1751-8121/aaa68d. URL https://dx.doi.org/10.1088/1751-8121/aaa68d.

Massih-Reza Amini, Vasilii Feofanov, Loic Pauletto, Emilie Devijver, and Yury Maximov. Self-training: A survey. arXiv preprint arXiv:2202.12040, 2022.

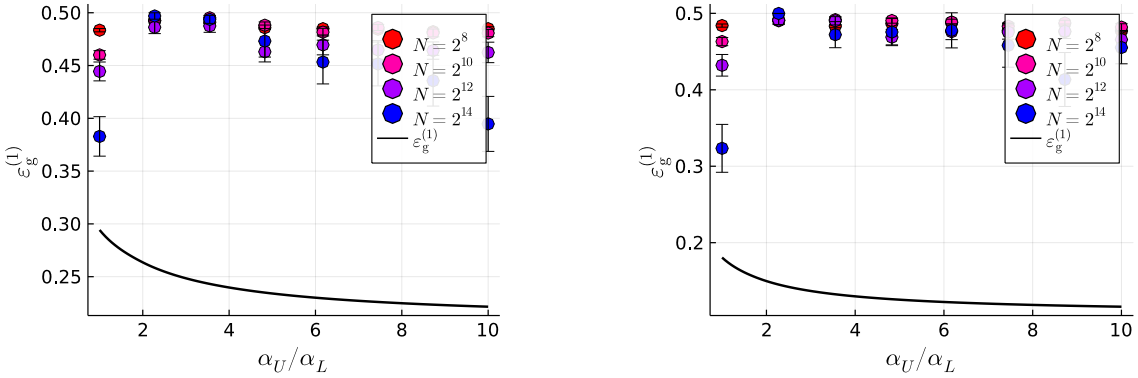
Carlo Baldassi, Christian Borgs, Jennifer T Chayes, Alessandro Ingrosso, Carlo Lucibello, Luca Saglietti, and Riccardo Zecchina. Unreasonable effectiveness of learning neural



(a) $(\rho, \Delta) = (0.5, 1.0)$ (b) $(\rho, \Delta) = (0.495, 1.0)$ (c) $(\rho, \Delta) = (0.4, 1.0)$

(d) $(\rho, \Delta) = (0.5, 0.75^2)$ (e) $(\rho, \Delta) = (0.495, 0.75^2)$ (f) $(\rho, \Delta) = (0.4, 0.75^2)$

Figure 14: The optimal regularization parameters $(\lambda_*^{(0)}, \lambda_*^{(1)})$, which minimize the generalization error $\epsilon_{\rm g}^{(1)}$, are plotted against the relative number of unlabeled data points α_U/α_L. Each panel corresponds to the different values of ρ and Δ.



(a) $(\rho, \Delta) = (0.5, 1.0)$ (b) $(\rho, \Delta) = (0.5, 0.75^2)$

Figure 15: Theoretical estimate for the optimal generalization error $\epsilon_{\rm g}^{(1)}$ for $\rho = 0.5$ is compared with the experiments. The black solid line is the theoretical estimate in LSL, and the symbols represent the experimental results. Here the regularization parameter is limited as $(\lambda^{(0)}, \lambda^{(1)}) \in (0, 10)^2$. For moderately large α_U, $(\lambda^{(0)}, \lambda^{(1)}) = (10, 10)$, however, for such large regularization parameters, the experimental results heavily suffer from large finite-size effects.

networks: From accessible states and robust ensembles to basic algorithmic schemes. Proceedings of the National Academy of Sciences, 113(48):E7655–E7662, 2016.

Afonso S Bandeira, Ahmed El Alaoui, Samuel B. Hopkins, Tselil Schramm, Alexander S Wein, and Ilias Zadik. The franz-parisi criterion and computational trade-offs in high dimensional statistics. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, Advances in Neural Information Processing Systems, 2022. URL https://openreview.net/forum?id=mzze3bubjk.

Jean Barbier and Nicolas Macris. The adaptive interpolation method: a simple scheme to prove replica formulas in Bayesian inference. Probability Theory and Related Fields, 174 (3):1133–1185, August 2019. ISSN 1432-2064. doi: 10.1007/s00440-018-0879-0. URL https://doi.org/10.1007/s00440-018-0879-0.

Jean Barbier, Mohamad Dia, Nicolas Macris, Florent Krzakala, Thibault Lesieur, and Lenka Zdeborová. Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula. In Advances in Neural Information Processing Systems, volume 29, 2016.

Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. Proceedings of the National Academy of Sciences, 116(12):5451–5460, 2019.

Blake Bordelon and Cengiz Pehlevan. Self-consistent dynamical field theory of kernel evolution in wide neural networks. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, Advances in Neural Information Processing Systems, volume 35, pages 32240–32256. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/d027a5c93d484a4312cc486d399c62c1-Paper-Conference.pdf.

Blake Bordelon and Cengiz Pehlevan. Dynamics of finite width kernel and prediction fluctuations in mean field neural networks. In Thirty-seventh Conference on Neural Information Processing Systems, 2023a. URL https://openreview.net/forum?id=fKwG6grp8o.

Blake Bordelon and Cengiz Pehlevan. The influence of learning rule on representation dynamics in wide neural networks. In The Eleventh International Conference on Learning Representations, 2023b. URL https://openreview.net/forum?id=nZ2NtpolC5-.

Tianle Cai, Ruiqi Gao, Jason Lee, and Qi Lei. A theory of label propagation for subpopulation shift. In Marina Meila and Tong Zhang, editors, Proceedings of the 38th International Conference on Machine Learning, volume 139 of Proceedings of Machine Learning Research, pages 1170–1182. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/cai21b.html.

Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Out-of-distribution generalization in kernel regression. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, Advances in Neural Information Processing Systems, volume 34, pages 12600–12612. Curran Associates, Inc., 2021a. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/691dcb1d65f31967a874d18383b9da75-Paper.pdf.

Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. Nature communications, 12(1):1–12, 2021b.

Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. Nature Communications, 12(1):2914, May 2021c. ISSN 2041-1723. doi: 10.1038/s41467-021-23103-1. URL https://doi.org/10.1038/s41467-021-23103-1.

Kabir Aladin Chandrasekher, Ashwin Pananjady, and Christos Thrampoulidis. Sharp global convergence guarantees for iterative nonconvex optimization with random data. The Annals of Statistics, 51(1):179–210, 2023.

Olivier Chapelle, Bernhard Schlkopf, and Alexander Zien. Semi-Supervised Learning. The MIT Press, 1st edition, 2010. ISBN 0262514125.

Patrick Charbonneau, Enzo Marinari, Marc Mézard, Giorgio Parisi, Federico Ricci-Tersenghi, Gabriele Sicuro, and Francesco Zamponi. Spin Glass Theory and Far Beyond. WORLD SCIENTIFIC, 2023. doi: 10.1142/13341. URL https://www.worldscientific.com/doi/abs/10.1142/13341.

Yining Chen, Colin Wei, Ananya Kumar, and Tengyu Ma. Self-training avoids using spurious features under domain shift. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems, volume 33, pages 21061–21071. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/f1298750ed09618717f9c10ea8d1d3b0-Paper.pdf.

Stéphane D'Ascoli, Maria Refinetti, Giulio Biroli, and Florent Krzakala. Double trouble in double descent: Bias and variance(s) in the lazy regime. In Hal Daumé III and Aarti Singh, editors, Proceedings of the 37th International Conference on Machine Learning, volume 119 of Proceedings of Machine Learning Research, pages 2280–2290. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/d-ascoli20a.html.

Rainer Dietrich, Manfred Opper, and Haim Sompolinsky. Statistical mechanics of support vector networks. Phys. Rev. Lett., 82:2975–2978, Apr 1999. doi: 10.1103/PhysRevLett.82.2975. URL https://link.aps.org/doi/10.1103/PhysRevLett.82.2975.

Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. The Annals of Statistics, 46(1):247–279, 2018. ISSN 00905364, 21688966. URL https://www.jstor.org/stable/26542784.

A. Engel and C. Van den Broeck. Statistical Mechanics of Learning. Cambridge University Press, 2001.

Silvio Franz and Giorgio Parisi. Phase diagram of coupled glassy systems: A mean-field study. Physical review letters, 79(13):2486, 1997.

Silvio Franz and Giorgio Parisi. Effective potential in glassy systems: theory and simulations. Physica A: Statistical Mechanics and its Applications, 261(3-4):317–339, 1998.

Silvio Franz and Giorgio Parisi. Quasi-equilibrium in glassy dynamics: an algebraic view. Journal of Statistical Mechanics: Theory and Experiment, 2013(02):P02003, feb 2013. doi: 10.1088/1742-5468/2013/02/P02003. URL https://dx.doi.org/10.1088/1742-5468/2013/02/P02003.

Spencer Frei, Difan Zou, Zixiang Chen, and Quanquan Gu. Self-training converts weak learners to strong learners in mixture models. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, Proceedings of The 25th International Conference on Artificial Intelligence and Statistics, volume 151 of Proceedings of Machine Learning Research, pages 8003–8021. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/frei22a.html.

Marylou Gabrié, Andre Manoel, Clément Luneau, jean barbier, Nicolas Macris, Florent Krzakala, and Lenka Zdeborová. Entropy and mutual information in models of deep neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/6d0f846348a856321729a2f36734d1a7-Paper.pdf.

Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mezard, and Lenka Zdeborova. Generalisation error in learning with random features and the hidden manifold model. In Hal Daumé III and Aarti Singh, editors, Proceedings of the 37th International Conference on Machine Learning, volume 119 of Proceedings of Machine Learning Research, pages 3452–3462. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/gerace20a.html.

Cédric Gerbelot, Alia Abbara, and Florent Krzakala. Asymptotic errors for high-dimensional convex penalized linear regression beyond gaussian matrices. In Conference on Learning Theory, pages 1682–1713. PMLR, 2020.

Cedric Gerbelot, Alia Abbara, and Florent Krzakala. Asymptotic errors for teacher-student convex generalized linear models (or: How to prove kabashima's replica formula). IEEE Transactions on Information Theory, 69(3):1824–1852, 2023. doi: 10.1109/TIT.2022.3222913.

Yves Grandvalet and Yoshua Bengio. Entropy regularization. Semi-Supervised Learning, 2006.

Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2(7), 2015.

Haiping Huang and Yoshiyuki Kabashima. Origin of the computational hardness for learning with binary synapses. Physical Review E, 90(5):052813, 2014.

Yuma Ichikawa and Koji Hukushima. Learning dynamics in linear VAE: Posterior collapse threshold, superfluous latent space pitfalls, and speedup with KL annealing. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, Proceedings of The 27th International Conference on Artificial Intelligence and Statistics, volume 238 of

Proceedings of Machine Learning Research, pages 1936–1944. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/ichikawa24a.html.

Ryo Karakida and Shotaro Akaho. Learning curves for continual learning in neural networks: Self-knowledge transfer and forgetting. In International Conference on Learning Representations, 2022. URL https://openreview.net/forum?id=tFgdrQbbaa.

Anders Krogh and John Hertz. A simple weight decay can improve generalization. In J. Moody, S. Hanson, and R.P. Lippmann, editors, Advances in Neural Information Processing Systems, volume 4. Morgan-Kaufmann, 1991. URL https://proceedings.neurips.cc/paper_files/paper/1991/file/8eefcfdf5990e441f0fb6f3fad709e21-Paper.pdf.

Florent Krzakala and Jorge Kurchan. Landscape analysis of constraint satisfaction problems. Phys. Rev. E, 76:021122, Aug 2007. doi: 10.1103/PhysRevE.76.021122. URL https://link.aps.org/doi/10.1103/PhysRevE.76.021122.

Ananya Kumar, Tengyu Ma, and Percy Liang. Understanding self-training for gradual domain adaptation. In Hal Daumé III and Aarti Singh, editors, Proceedings of the 37th International Conference on Machine Learning, volume 119 of Proceedings of Machine Learning Research, pages 5468–5479. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/kumar20c.html.

Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In Workshop on challenges in representation learning, ICML, page 896, 2013.

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. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, Advances in Neural Information Processing Systems, volume 34, pages 18137–18151. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/9704a4fc48ae88598dcbdcdf57f3fdef-Paper.pdf.

Bruno Loureiro, Cedric Gerbelot, Maria Refinetti, Gabriele Sicuro, and Florent Krzakala. Fluctuations, bias, variance & ensemble of learners: Exact asymptotics for convex losses in high-dimension. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, Proceedings of the 39th International Conference on Machine Learning, volume 162 of Proceedings of Machine Learning Research, pages 14283–14314. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/loureiro22a.html.

P. C. Martin, E. D. Siggia, and H. A. Rose. Statistical dynamics of classical systems. Phys. Rev. A, 8:423–437, Jul 1973. doi: 10.1103/PhysRevA.8.423. URL https://link.aps.org/doi/10.1103/PhysRevA.8.423.

G. J. McLachlan. Iterative reclassification procedure for constructing an asymptotically optimal rule of allocation in discriminant analysis. Journal of the American Statistical

Association, 70(350):365–369, 1975. doi: 10.1080/01621459.1975.10479874. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1975.10479874.

Marc Mézard, Giorgio Parisi, and Miguel Angel Virasoro. Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications, volume 9. World Scientific Publishing Company, 1987.

Francesca Mignacco, Florent Krzakala, Yue Lu, Pierfrancesco Urbani, and Lenka Zdeborova. The role of regularization in classification of high-dimensional noisy gaussian mixture. In International Conference on Machine Learning, pages 6874–6883. PMLR, 2020a.

Francesca Mignacco, Florent Krzakala, Pierfrancesco Urbani, and Lenka Zdeborová. Dynamical mean-field theory for stochastic gradient descent in gaussian mixture classification. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems, volume 33, pages 9540–9550. Curran Associates, Inc., 2020b. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6c81c83c4bd0b58850495f603ab45a93-Paper.pdf.

Francesca Mignacco, Pierfrancesco Urbani, and Lenka Zdeborová. Stochasticity helps to navigate rough landscapes: comparing gradient-descent-based algorithms in the phase retrieval problem. Machine Learning: Science and Technology, 2(3):035029, jul 2021. doi: 10.1088/2632-2153/ac0615. URL https://dx.doi.org/10.1088/2632-2153/ac0615.

Patrick Kofod Mogensen and Asbjørn Nilsen Riseth. Optim: A mathematical optimization package for Julia. Journal of Open Source Software, 3(24):615, 2018. doi: 10.21105/joss.00615.

Andrea Montanari and Subhabrata Sen. A friendly tutorial on mean-field spin glass techniques for non-physicists. Foundations and Trends® in Machine Learning, 17(1):1–173, 2024. ISSN 1935-8237. doi: 10.1561/2200000105. URL http://dx.doi.org/10.1561/2200000105.

Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. Text classification from labeled and unlabeled documents using em. Machine learning, 39(2):103–134, 2000.

Tomoyuki Obuchi and Yoshiyuki Kabashima. Cross validation in lasso and its acceleration. Journal of Statistical Mechanics: Theory and Experiment, 2016(5):053304, 2016.

Koki Okajima and Takashi Takahashi. Asymptotic dynamics of alternating minimization for bilinear regression. Journal of Statistical Mechanics: Theory and Experiment, 2025(5):053301, jun 2025. doi: 10.1088/1742-5468/add1ce. URL https://doi.org/10.1088/1742-5468/add1ce.

Koki Okajima, Xiangming Meng, Takashi Takahashi, and Yoshiyuki Kabashima. Average case analysis of lasso under ultra sparse conditions. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, Proceedings of The 26th International Conference on Artificial Intelligence and Statistics, volume 206 of Proceedings of Machine Learning

Research, pages 11317–11330. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/okajima23a.html.

Manfred Opper. Learning to generalize. Frontiers of Life, 3(part 2):763–775, 2001.

Manfred Opper and Wolfgang Kinzel. Statistical mechanics of generalization. In Models of Neural Networks III: Association, Generalization, and Representation, pages 151–209. Springer, 1996.

Samet Oymak and Talha Cihad Gulcu. Statistical and algorithmic insights for semi-supervised learning with self-training. arXiv preprint arXiv:2006.11006, 2020.

Samet Oymak and Talha Cihad Gulcu. A theoretical characterization of semi-supervised learning with self-training for gaussian mixture models. In International Conference on Artificial Intelligence and Statistics, pages 3601–3609. PMLR, 2021.

Giorgio Parisi, Pierfrancesco Urbani, and Francesco Zamponi. Theory of simple glasses: exact solutions in infinite dimensions. Cambridge University Press, 2020.

Cengiz Pehlevan and Blake Bordelon. Lecture notes on infinite-width limits of neural networks, 2023.

Mohammad Pezeshki, Amartya Mitra, Yoshua Bengio, and Guillaume Lajoie. Multi-scale feature learning dynamics: Insights for double descent. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, Proceedings of the 39th International Conference on Machine Learning, volume 162 of Proceedings of Machine Learning Research, pages 17669–17690. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/pezeshki22a.html.

Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V. Le. Meta pseudo labels. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 11557–11568, June 2021.

Farzad Pourkamali and Nicolas Macris. Bayesian extensive-rank matrix factorization with rotational invariant priors. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, Advances in Neural Information Processing Systems, volume 36, pages 24025–24073. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/4b8afc47273c746662a96dfdf562f87f-Paper-Conference.pdf.

Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In International Conference on Learning Representations, 2021. URL https://openreview.net/forum?id=-ODN6SbiUU.

Luca Saglietti and Lenka Zdeborova. Solvable model for inheriting the regularization through knowledge distillation. In Joan Bruna, Jan Hesthaven, and Lenka Zdeborova, editors, Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference, volume 145 of Proceedings of Machine Learning Research, pages 809–846. PMLR, 16–19 Aug 2022. URL https://proceedings.mlr.press/v145/saglietti22a.html.

H Schwarze. Learning a rule in a multilayer neural network. Journal of Physics A: Mathematical and General, 26(21):5781, nov 1993. doi: 10.1088/0305-4470/26/21/017. URL https://dx.doi.org/10.1088/0305-4470/26/21/017.

Henry Schwarze and John Hertz. Generalization in a large committee machine. EPL (Europhysics Letters), 20(4):375, 1992.

Henry Scudder. Probability of error of some adaptive pattern-recognition machines. IEEE Transactions on Information Theory, 11(3):363–371, 1965.

H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In Proceedings of the fifth annual workshop on Computational learning theory, pages 287–294, 1992.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, Advances in Neural Information Processing Systems, volume 35, pages 19523–19536. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/7b75da9b61eda40fa35453ee5d077df6-Paper-Conference.pdf.

Takashi Takahashi and Yoshiyuki Kabashima. Semi-analytic approximate stability selection for correlated data in generalized linear models. Journal of Statistical Mechanics: Theory and Experiment, 2020(9):093402, sep 2020. doi: 10.1088/1742-5468/ababff. URL https://dx.doi.org/10.1088/1742-5468/ababff.

Takashi Takahashi and Yoshiyuki Kabashima. Macroscopic analysis of vector approximate message passing in a model-mismatched setting. IEEE Transactions on Information Theory, 68(8):5579–5600, 2022. doi: 10.1109/TIT.2022.3163342.

Kaito Takanami, Takashi Takahashi, and Ayaka Sakata. The effect of optimal self-distillation in noisy gaussian mixture model. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, Advances in Neural Information Processing Systems, volume 38, pages 17213–17258. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/18ddfb199d71a8a24f83abc1ced077b7-Paper-Conference.pdf.

Michel Talagrand. Mean field models for spin glasses: Volume I: Basic examples, volume 54. Springer Science & Business Media, 2010.

Umberto M Tomasini, Antonio Sclocchi, and Matthieu Wyart. Failure and success of the spectral bias prediction for Laplace kernel ridge regression: the case of low-dimensional data. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, Proceedings of the 39th International Conference on Machine Learning, volume 162 of Proceedings of Machine Learning Research, pages 21548–21583. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/tomasini22a.html.

Colin Wei, Kendrick Shen, Yining Chen, and Tengyu Ma. Theoretical analysis of self-training with deep networks on unlabeled data. In International Conference on Learning Representations, 2021. URL https://openreview.net/forum?id=rC8sJ4i6kaH.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020.

I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. arXiv preprint arXiv:1905.00546, 2019.

Hajime Yoshino. From complex to simple: hierarchical free-energy landscape renormalized in deep neural networks. SciPost Physics Core, 2(2):005, 2020.

Jacob Zavatone-Veth and Cengiz Pehlevan. Learning curves for deep structured gaussian feature models. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, Advances in Neural Information Processing Systems, volume 36, pages 42866–42897. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/85d456fd41f3eec83bd3b0c337037a0e-Paper-Conference.pdf.

Lenka Zdeborová and Florent Krzakala. Statistical physics of inference: thresholds and algorithms. Advances in Physics, 65(5):453–552, 2016. doi: 10.1080/00018732.2016.1211393. URL https://doi.org/10.1080/00018732.2016.1211393.

Bowen Zhang, Yidong Wang, Wenxin Hou, HAO WU, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, Advances in Neural Information Processing Systems, volume 34, pages 18408–18419. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/995693c15f439e3d189b06e89d145dd5-Paper.pdf.

Shuai Zhang, Meng Wang, Sijia Liu, Pin-Yu Chen, and Jinjun Xiong. How unlabeled data improve generalization in self-training? a one-hidden-layer theoretical analysis. In International Conference on Learning Representations, 2022. URL https://openreview.net/forum?id=qiMXBIf4NfB.

Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In Proceedings of the European conference on computer vision (ECCV), pages 289–305, 2018.