Restoring balance: principled under/oversampling of data for optimal classification

Emanuele Loffredo [* 1]   Mauro Pastore [* 1]   Simona Cocco [1]   Rémi Monasson [1]

Abstract

Class imbalance in real-world data poses a common bottleneck for machine learning tasks, since achieving good generalization on under-represented examples is often challenging. Mitigation strategies, such as under or oversampling the data depending on their abundances, are routinely proposed and tested empirically, but how they should adapt to the data statistics remains poorly understood. In this work, we determine exact analytical expressions of the generalization curves in the high-dimensional regime for linear classifiers (Support Vector Machines). We also provide a sharp prediction of the effects of under/oversampling strategies depending on class imbalance, first and second moments of the data, and the metrics of performance considered. We show that mixed strategies involving under and oversampling of data lead to performance improvement. Through numerical experiments, we show the relevance of our theoretical predictions on real datasets, on deeper architectures and with sampling strategies based on unsupervised probabilistic models.

1. Introduction

Many real-world classification tasks, *e.g.* in automated medical diagnostics (Krawczyk et al., 2016; Fotouhi et al., 2019), molecular biology (Wang et al., 2006; Yang et al., 2012; Cheng et al., 2015; Song et al., 2021; Ansari & White, 2023), text classification (Liu et al., 2009; 2017), ... are plagued by the so-called curse of class imbalance (Kubat & Matwin, 1997; Lemaître et al., 2017): one or more classes

in the training data are significantly under-represented, yet correct identification of these rare examples is crucial for performance (Weiss, 2004). For such imbalanced datasets, machine learning methods struggle to achieve good classification performances when tested fairly (He & Ma, 2013).

Due to its wide-spread and intrinsic nature, the issue of learning with class imbalance has long been studied in the computer science literature. Systematic empirical studies, assessing the performances of various architectures and algorithms with imbalanced data have been carried out (Japkowicz & Stephen, 2002; Akbani et al., 2004; Lemnaru & Potolea, 2012; Buda et al., 2018), with a renewed interest in the era of big data and deep learning (Ghosh et al., 2022; Johnson & Khoshgoftaar, 2019). Various strategies (Anand et al., 1993; Laurikkala, 2001; Chawla et al., 2002; Batista et al., 2004; He et al., 2008; Alshammari et al., 2022) based on restoring balance in the training set can be found in textbooks (He & Ma, 2013; Fernández et al., 2018). Despite this considerable body of work, a theoretical understanding of learning with class imbalance remains elusive, which possibly impedes the design of optimal, task-adapted strategies.

In this work we propose a probabilistic setting that faithfully takes into account the imbalance ratio and essential features of the data, such as their first and second moments, in which the performances of linear classifiers can be analytically derived in the high-dimensional limit using the framework of the statistical mechanics of learning (Engel & Van den Broeck, 2001). Based on this theory we can identify strategies to under or oversample the data in, respectively, the majority and minority classes offering optimal performances for the metrics of interest.

Summary of the main results. We report here the main results of our work.

In Sec. 2, we analytically characterize the asymptotic performances of linear classifiers trained on imbalanced datasets, when data points are discrete multi-state variables. These asymptotics are based on the so-called *replica method* from statistical physics.

We provide *quantitative* behaviors of various metrics within this setting – such as confusion matrix, accuracy (ACC), balanced accuracy (BA) and area under the ROC curve

---

*Equal contribution [1]Laboratoire de physique de l'École normale supérieure, CNRS-UMR8023, PSL University, Sorbonne University, Université Paris-Cité 24 rue Lhomond, 75005 Paris, France. Correspondence to: Emanuele Loffredo <emanuele.loffredo@phys.ens.fr>, Mauro Pastore <mauro.pastore@phys.ens.fr>.

(AUC). This allows us to show, for example, that AUC is rather insensitive to imbalance, and is therefore not a reliable predictor of some aspects of the performances captured by more sensitive metrics, e.g. BA.

In Sec. 3, we provide exact asymptotic predictions to assess the effect of simple under and/or oversampling strategies to mitigate the imbalance problem. We show that best performances are reached with mixed strategies restoring data balance (Fig. 1).

In Sec. 4, we provide empirical support that our theoretical findings qualitatively hold with deeper classifiers and more sophisticated under/oversampling methods. In particular, we propose and test an unsupervised probabilistic model, learned on the minority class, to generate new minority-class instances or to filter out majority-class data (Fig. 1).
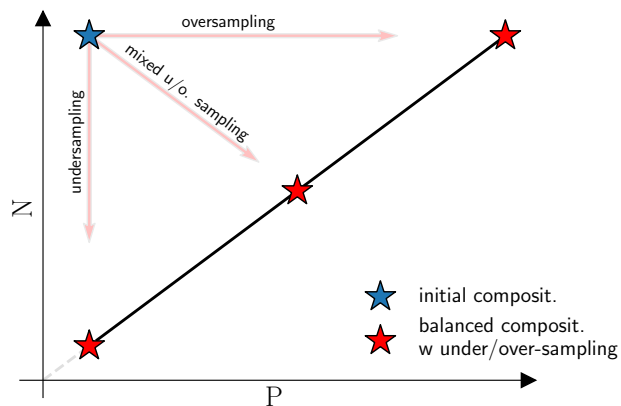

Figure 1. Illustration of our restoring balance procedure. Classifiers trained on datasets with severe imbalance (blue star) generally show poor generalization performances. Restoring balance by mixing under and oversampling improves classification performances (red stars across the line $P = N$). Here P, N indicate the sizes of the positive and negative classes, initially with $P \ll N$.

Related works. Learning imbalanced mixture models of data with linear classifiers has long been studied in the statistical physics literature (Del Giudice, P. et al., 1989). Recently, Loureiro et al. (2021); Mignacco et al. (2020); Pesce et al. (2023) evaluated rigorously the generalization error of a linear classifier on data from Gaussian mixture models, on test sets distributed as the training set; in the present work, we consider instead data in the form of multi-state discrete variables not necessarily normally distributed, we compare different performance metrics and study the impact of mitigation strategies on the learning protocol. For rigorous results that go beyond the hypothesis of Gaussian data from this approach, see the recent Dandi et al. (2023). Mannelli et al. (2023) revisited this problem in the context of the fairness of AI methods; at variance with them, we consider a classification task with labels coinciding with class membership, and we focus on dataset-preprocessing miti-

gation strategies more than on algorithmic-based methods (loss-reweighting, coupled neural networks). The dynamics of learning with gradient-based methods in presence of class imbalance, which we do not address here, was recently studied by Francazi et al. (2023).

Other theoretical assessments from complementary point of views were formulated in the recent past to obtain optimal oversampling ratios for imbalanced datasets (Shang et al., 2023; Chaudhuri et al., 2023). An interesting comparison with our work is given by Menon et al. (2013), which proves that classifiers trained with empirically-balanced losses attain optimal performances in the limit of infinite data: as discussed in Section 3.4, we can account for this setting within our formalism, but our work deals with the proportional asymptotics where both the input dimension and the training set size are large, for which we show that simple loss-reweighting methods are sub-optimal with respect to data augmentation techniques more sophisticated than random oversampling.

In addition, a plethora of methods have been proposed to effectively restore balance in datasets. SMOTE (Chawla et al., 2002), whose strategy is to oversample the minority class by linearly interpolating existing data points, is one of the most frequently applied algorithms and exists in the literature under more than 100 variations (Kovács, 2019). The use of unsupervised generative model to first learn and then augment the minority class has been actively explored, *e.g.* with generative adversarial networks (Douzas & Bacao, 2018), autoencoders (Mondal et al., 2023), variational autoencoders (Wan et al., 2017; Ai et al., 2023), etc. Our proposed approch is closely related to RBM-SMOTE, introduced by Zięba et al. (2015), where a Restricted Boltzmann Machine (RBM) is trained on the minority class and then used to generate new samples starting from intermediate points obtained by SMOTE. Similar restoring balance procedures have been showed in Mirza et al. (2021) to have a positive impact on different performances metrics.

Table 1. Notations and conventions used in this work.

L, Q	dimension of the data, size of the alphabet
P, N	positive and negative class sizes of training data
α^+, α^-	classes sizes scaled by data dimension
ρ^+, ρ^-	fractions of the two classes in the training dataset
φ^+, φ^-	fractions of the two classes in the test dataset
$\langle \cdot \rangle_\pm$	average over positive and negative data
$\mathbf{M}, \boldsymbol{\delta}$	midpoint and shift between the classes' centers
C	covariance matrix of the data
$\mathbf{J}, b$	weights and bias of the linear SVM
κ	margin of the linear SVM
Δ^+, Δ^-	pre-activations of the output neuron
$\ell(y, \Delta)$	training loss function
$g(y, \Delta)$	test loss function

2. Theoretical setting

We consider pairs $(\mathbf{v}^\nu, y^\nu) \in \mathbb{R}^{L \times Q} \times \{-1, +1\}$ with $\nu = 1, \ldots, P + N$ of training data points independently sampled, where L sets the dimension of data points and each symbol v_i can take Q values (Table 1). For example, in image recognition tasks, L is the number of pixels and Q the number of possible colors of each pixel; in molecular biology, L is the length of the biological sequence and Q the size of the alphabet, *e.g.* 20 for amino acids. In statistical physics, this kind of data are usually called Potts configurations; in computer science, they are routinely obtained via tokenization, for example in language models (Jurafsky & Martin, 2009) or image recognition (Wu et al., 2020). A convenient representation of these categorical variables is one-hot encoding: $s_i(t) = 1$ if $v_i = t$, 0 otherwise.

In particular, we focus on a binary classification task where the classes are labeled by $y = \pm 1$ and the training set is made of P positive and N negative examples, $\mathcal{T} = \{(\mathbf{s}_+^\mu, y^\mu = +1)\}_{\mu=1}^P \cup \{(\mathbf{s}_-^\nu, y^\nu = -1)\}_{\nu=1}^N$. The classes are *imbalanced*, *i.e.* in general $P \neq N$. We denote with $\alpha^\pm$ the classes' sizes scaled by the dimension of data, *e.g.* $\alpha^+ = P/L$, and with $\rho^\pm$ the fractions of positive and negative examples in the training set, *i.e.* $\rho^\pm = \alpha^\pm / (\alpha^+ + \alpha^-)$. We are interested in studying the statistical properties of the linear classifier

$$\hat{y} = f(\mathbf{s}) = \text{sgn}\left[\sum_{i,t} \frac{J_i(t)s_i(t)}{\sqrt{L}} - b\right], \quad (1)$$

where the weights and bias $(\mathbf{J}, b) \in \mathbb{R}^{L \times Q} \times \mathbb{R}$ define the model and are learnt over the training data. The scaling of the inner product with $\sqrt{L}$ is chosen to obtain components $J_i(t) \sim O(1)$ after regularization and training. Due to the property $\sum_t s_i(t) = 1$ of one-hot encoding, a linear classifier described by the set of parameters $(J_i(t), b)$ is identical to the one defined by $(J_i(t) + u_i, b + \sum_i u_i/\sqrt{L})$ for any arbitrary real numbers u_i. To avoid this overparametrization, we impose the zero-sum conditions $\sum_t J_i(t) = 0$ for all i.

Given the training set, the parameters of the model are learnt through the following Empirical Risk Minimization (ERM):

$$(\mathbf{J}^\star, b^\star) = \underset{\mathbf{J} \in \mathcal{S}, b \in \mathbb{R}}{\arg\min} \mathcal{L}(\mathbf{J}, b), \quad (2)$$

$$\mathcal{L}(\mathbf{J}, b) = \sum_{\mu=1}^P \ell\left[y^\mu = +1, \Delta_\mu^+(\mathbf{J}, b)\right]$$

$$+ \sum_{\nu=1}^N \ell\left[y^\nu = -1, \Delta_\nu^-(\mathbf{J}, b)\right], \quad (3)$$

$$\Delta_\mu^\pm(\mathbf{J}, b) = \sum_{i,t} \frac{J_i(t)s_{\pm,i}^\mu(t)}{\sqrt{L}} - b, \quad (4)$$

where $\ell : \{-1, +1\} \times \mathbb{R} \to \mathbb{R}^+$ is a loss function and the set $\mathcal{S}$, over which the optimization problem (2) is defined,

is the surface of the $(L \times Q)$-dimensional sphere (spherical regularization) compatible with the zero-sum conditions. Notice that the variables $\Delta_\mu^\pm$ defined in Eq. (4) are simply the pre-activations of the output neuron evaluated on the input points $\mathbf{s}_\pm^\mu$. A possible choice of the loss function, which we will use in the following as a case of study, is the hinge loss

$$\ell(y, \Delta) = \max(0, \kappa - y\Delta), \quad (5)$$

for some positive parameter κ called *margin*. High values of κ make the classifier less affected by noise in the training data, so that the learning protocol is more *stable*. The choice of the set $\mathcal{S}$ and of the hinge loss implies that the linear model (1) is a spherical perceptron with hinge loss (Franz et al., 2019), which is equivalent to a soft-margin support vector machine with L2 regularization, as detailed in the Appendix A.6.

In this work we assume that the input data points belonging to the two classes are sampled independently from a distribution having the following first and second moments:

$$\langle s_i(t) \rangle_\pm = M_i(t) \pm \frac{\delta_i(t)}{2\sqrt{L}},$$

$$\langle s_i(t)s_j(u) \rangle_\pm - \langle s_i(t) \rangle_\pm \langle s_j(u) \rangle_\pm = C_{ij}(t, u), \quad (6)$$

where the angle brackets stand for the expectations over the positive and negative classes. The vector $\mathbf{M}$ represents the global center-of-mass of the positive and negative distributions, $\boldsymbol{\delta}$ is the (rescaled) shift between their centers. The scaling by the factor $1/\sqrt{L}$ with $\delta_i(t) \sim O(1)$ ensures that the classification task is non-trivial for a linear classifier in the high-dimensional regime $L \gg 1$. Notice that, to simplify the analysis, we are assuming the two classes to have the same covariance, a condition often referred to as *homoscedasticity*. Higher-order statistics of the data is irrelevant for the asymptotic properties reported below under mild conditions on the cumulants (Monasson, 1992).

2.1. Statistical mechanics of the learning problem

The ERM problem introduced in the previous section can be rephrased in a statistical mechanics framework. We consider $(\mathbf{J}, b)$ as the configuration of a 'physical' system with energy $\mathcal{L}$. The partition function of this system reads

$$\Omega = \int db \, d\mu(\mathbf{J}) \exp[-\beta \mathcal{L}(\mathbf{J}, b)], \quad (7)$$

where $d\mu(\mathbf{J})$ is an opportune measure over the weights taking into account regularization and zero-sum conditions, as defined in Appendix A, Eq. (16). As the 'inverse temperature' $\beta \to \infty$, the integral in Eq. (7) is dominated by the solution of the optimization problem (2). Within this framework, we are able to derive the expected values of functions of $(\mathbf{J}^\star, b^\star)$ over the data. In particular, we can characterize the asymptotic performance of the linear classifer.

Proposition 2.1. *Consider the ERM problem in Eq. (2) under the assumption of training data distribution (6). Let the fractions $\varphi^\pm$ be the composition of the test set and*

$$\mathcal{M}_{\text{gen.}}(\varphi^+, \varphi^-) = \sum_{y \in \{\pm\}} \varphi^y \langle g(\Delta^y, y) \rangle_{\Delta^y}, \quad (8)$$

a performance metric of choice, with $\Delta^\pm$ defined as in Eq. (4) over test data points and g a generic test loss function (see Table 2 for examples). In the asymptotic regime $L, P, N \to \infty$ at fixed ratios $\alpha^+ = P/L$ and $\alpha^- = N/L$, the expected value in Eq. (8) can be taken over the normal distribution

$$p(\Delta^\pm) = \mathcal{N}\left(\pm\frac{r}{2} - b, q\right), \quad (9)$$

whose parameters q, r, b are the solution of the saddle-point equations stemming from the extremization of the following quantity

$$\mathbb{E}_{\text{data}}[\log \Omega] \sim \frac{\beta L}{2} \underset{\substack{b,r,q,x \\ \hat{k},\hat{r},\hat{q},\hat{x}}}{\text{extr}} (G_S + \alpha^+ G_+ + \alpha^- G_-). \quad (10)$$

In the above proposition – fully derived in Appendices A.1-4, B – the functions $G_\pm$ and G_S in Eq (10) are defined as

$$G_\pm = -\sqrt{\frac{q}{2\pi}}\left(\frac{x - \mathcal{K}_\pm}{x}e^{-\frac{(x-\mathcal{K}_\pm)^2}{2q}} + \frac{\mathcal{K}_\pm}{x}e^{-\frac{\mathcal{K}_\pm^2}{2q}}\right)$$
$$+ \frac{(x-\mathcal{K}_\pm)^2 + q}{x}H\left(\frac{x-\mathcal{K}_\pm}{\sqrt{q}}\right) - \frac{\mathcal{K}_\pm^2 + q}{x}H\left(\frac{-\mathcal{K}_\pm}{\sqrt{q}}\right), \quad (11)$$

and

$$G_S = Q\hat{k} + \hat{x}q + \hat{q}x + 2\hat{r}r + \frac{\hat{r}^2}{L}\boldsymbol{\delta}^\top A^{-1}\boldsymbol{\delta}$$
$$- \frac{\hat{q}}{L}\operatorname{tr}(A^{-1}C), \quad (12)$$

where $\mathcal{K}_\pm = \kappa - r/2 \pm b$, $H(x) = \operatorname{erfc}(x/\sqrt{2})/2$ (erfc is the complementary error function), $A = \hat{k}\mathbb{I}_Q \otimes \mathbb{I}_L + \hat{x}C$ ($\otimes$ is the Kroenecker product). Here, algebraic operations as traces and vector/matrix multiplications are performed in the $(L \times Q)$-dimensional linear space spanned by the Potts and position indices.

We stress here that Proposition 2.1 holds under the hypotheses that higher cumulants in the distribution of the data can be neglected in Eq. (10) for large L with respect to the first and second cumulants reported in Eq. (6), and that the optimization problem (2) is strictly convex; the hypothesis of strict convexity is inessential and can be relaxed to convexity, in which case Eqs. (11) and (12) need to be substituted

Table 2. Analytical predictions for the elements of the confusion matrix and common performance metrics derived from them. Brackets indicate averages w.r.t. the distributions (9), the threshold γ is introduced to define the ROC and the PR curves, $\varphi^\pm$ indicates the composition of the test set and θ the Heaviside step function. Other metrics are reported in Appendix.

TPR(γ)	$\langle\theta(\Delta^+ - \gamma)\rangle_{\Delta^+}$	True positive rate
FPR(γ)	$\langle\theta(\Delta^- - \gamma)\rangle_{\Delta^-}$	False positive rate
TNR(γ)	$1 - \text{FPR}(\gamma)$	True negative rate
FNR(γ)	$1 - \text{TPR}(\gamma)$	False negative rate
PPV(γ)	$\dfrac{\varphi^+\text{TPR}(\gamma)}{\varphi^+\text{TPR}(\gamma) + \varphi^-\text{FPR}(\gamma)}$	Precision
ACC	$\varphi^+\text{TPR}(0) + \varphi^-\text{TNR}(0)$	Accuracy
BA	$[\text{TPR}(0) + \text{TNR}(0)]/2$	Balanced accuracy
AUC	$\int_{-\infty}^{\infty}d\gamma\|\text{FPR}'(\gamma)\|\text{TPR}(\gamma)$	Area under ROC
AUPRC	$\int_{-\infty}^{\infty}d\gamma\|\text{TPR}'(\gamma)\|\text{PPV}(\gamma)$	Area under PRC

with (35) and (30), as discussed in Appendix A. The extremization in Eq. (10) is performed numerically with respect to the bias b, the *order parameters* x, q, r, and the conjugate variables $\hat{x}$, $\hat{q}$, $\hat{r}$, $\hat{k}$. This procedure provides our analytical predictions for training and generalization metrics as a function of α^+, α^- and the data statistics, $\mathbf{M}$, $\boldsymbol{\delta}$ and C. The order parameters, whose precise definition in terms of replica-symmetric overlap matrices is detailed in the Appendix, have a simple physical interpretation in terms of the low-order statistics of the variables $\Delta_{\text{test}}^\pm(\mathbf{J}^\star, b^\star)$, defined over *test* data points. The order parameters can also be interpreted in geometric terms. In particular, r measures how well $\mathbf{J}^\star$ is aligned along the vector $\boldsymbol{\delta}$, q measures the inner product $(\mathbf{J}^\star)^\top C\mathbf{J}^\star/L$, and b is the optimal value of the bias in Eq. (1).

Detailed calculations and more general results, such as the expression for the functions G_S, $G_\pm$ in the linearly separable phase of the classifier, are reported in Appendix A.2-3.

2.2. Performance metrics

The approach above, and in particular Eq. (9) evaluated on the values of the order parameters obtained from the extremization (10), provides analytical expressions for the generalization performance, assessed through the elements of the confusion matrix from Table 2. For example, the accuracy ACC(φ^+, φ^-) (or 0/1 accuracy) referred to as ACC in Table 2 is obtained from Eq. (8) with $g(y, \Delta^y) = \theta(y \Delta^y)$, where θ is the Heaviside step function. The choice $\varphi^\pm = 0.5$ corresponds to the *balanced accuracy* (BA), a popular measure of performance with imbalanced datasets, while $\varphi^+ = \rho^+$ stands for test as imbalanced as training.

The expression of the balanced accuracy BA is generally involved, but simplifies for large margin $\kappa \gg 1$. In this regime, we obtain explicit rates of convergence in terms of the training set size (α^+, α^-). In the case of $Q = 2$ states $(t = \pm)$, position-independent $M_i(t = \pm) = (1 \pm m)/2$ and diagonal covariance matrix $C_{ij}(t = +, u = +) = \delta_{ij}(1 - m^2)/4$, we obtain the peak value of the BA at $\alpha^+ = \alpha^-$ as

$$\text{BA}(\alpha^+ = \alpha^-) \sim 1 - \frac{1}{2}\,\text{erfc}\left(\frac{\|\boldsymbol{\delta}\|_2}{2\sqrt{2L(1 - m^2)}}\right) + \frac{\mathcal{C}}{2\alpha^+}, \tag{13}$$

where $\mathcal{C}$ is a constant term, while $\text{BA}(\alpha^+ \neq \alpha^-) = 0$.

3. Theoretical results on imbalanced learning

3.1. Behaviour of common performance metrics

In this section we analyze the metrics defined in Table 2 based on the theory explained above. We first notice, from their definition, that these metrics are impacted by imbalance both *explicitly*, due to the dependence on the ratios $\varphi^\pm$ expressing the composition of the test set, and *implicitly*, due to the fact that they are evaluated on a model trained on an imbalanced dataset. Common choices for the test set are $\varphi^+ = \rho^+$ (test set distributed as the training set) or $\varphi^+ = 0.5$ (balanced test set).

We report in Fig. 2a the analytical curves as a function of the training imbalance ratio ρ^+, obtained in the theoretical setting explained above, for $\varphi^+ = \rho^+$. We conclude that:

- ACC, even if exhibiting a peak around the balanced point $\rho^+ = 0.5$, predicts best performances when the dataset is heavily imbalanced. This is expected: due to the explicit dependence on $\varphi^\pm$ and the fact that $\varphi^+ = \rho^+$, ACC only evaluates how well the majority class is predicted, even though the model behaves as a random classifier on the minority class.

- BA, which has no explicit dependence on the composition of the test set (it weights in the same way the probabilities of true label prediction on the majority and minority classes), assign best performances to the model trained on a balanced dataset.

- AUC predicts best performances away from the balanced point. AUC is defined as the area under the receiver operating characteristic (ROC) curve, that is the parametric curve $(\text{FPR}(\gamma), \text{TPR}(\gamma))$ as a function of the threshold on the predictor γ. As $\gamma \in (-\infty, \infty)$, AUC is independent from the bias b in Eq. (9), which is the observable most impacted by imbalance (see its trend obtained from our theory in Appendix A, Fig. 5, and many observations in literature, *e.g.* Chaudhuri et al. (2023)).

- AUPRC, *i.e.* the area under the Precision-Recall curve (PRC – the parametric curve $(\text{TPR}(\gamma), \text{PPV}(\gamma))$ as a function of γ), is often proposed as a better alternative to AUC in presence of imbalance. As AUC, it does not depend on the bias b, and thus does not directly measure the parameter most sensitive to imbalance. However, due to the explicit dependence on the test set (notice that $\text{PPV}(\infty) = 1$, $\text{PPV}(-\infty) = \varphi^+$), AUPRC is always bounded from below by $\varphi^+ = \rho^+$, increasing "by definition" as the positive class ratio in the dataset.

For more of these metrics, we address the reader to Appendix B and Table 3. We conclude from this analysis that, for different reasons we can explain within our theoretical setting, standard generalization metrics can be misleading about the performance of a model trained under imbalance. For the rest of this paper we will mostly make use of BA, which is one of the most robust in predicting best performances around the balanced point.

3.2. Agreement with results on real datasets.

We show in Figure 2b that the theory devised in this work is able to *quantitatively* predict the BA curves (as well as other generalization curves) of linear SVMs trained on standard benchmark datasets. In particular, we validated our predictions on (i) the Parity MNIST (pMNIST) dataset, having odd and even digits classes; (ii) Fashion MNIST (FMNIST) with classes containing "Pullover" and "Shirt" images; and (iii) CelebA with classes containing faces with "Straight hair" and "Wavy hair". Note that our theory assumes the two classes in the training dataset to have the same covariance matrices: while this is not generally true on all datasets, we observe agreement between numerical experiments and asymptotic predictions, by feeding the mean empirical covariance of positive and negative data, $C = (C_{\text{emp}}^+ + C_{\text{emp}}^-)/2$, in our theory. Such agreement remains quantitatively valid provided that the covariance matrices C_{emp}^+, C_{emp}^- are similar, and is only qualitative for more diverse covariances (*e.g.* 0/1 MNIST). Moreover, we report in Appendix E a benchmark dataset with Lattice Proteins where diagonal covariances are sufficient to predict quantitatively the numerical results.

3.3. Balance-to-performance trade-off

Using analytical results from Sec. 2, we can assess to what extent the imbalance of the dataset and the stability of the classifier impact the performance. We evaluate results in terms of the BA for different compositions of the training set. The generalization curves in Figure 2c exhibit the presence of a balance-to-performance trade-off: for high values of κ, the accuracy is extremely sensitive to the composition of the training set, and the outputs of the classifier are essentially random unless $\rho^- = \rho^+ \sim 0.5$. Increasing the margin κ
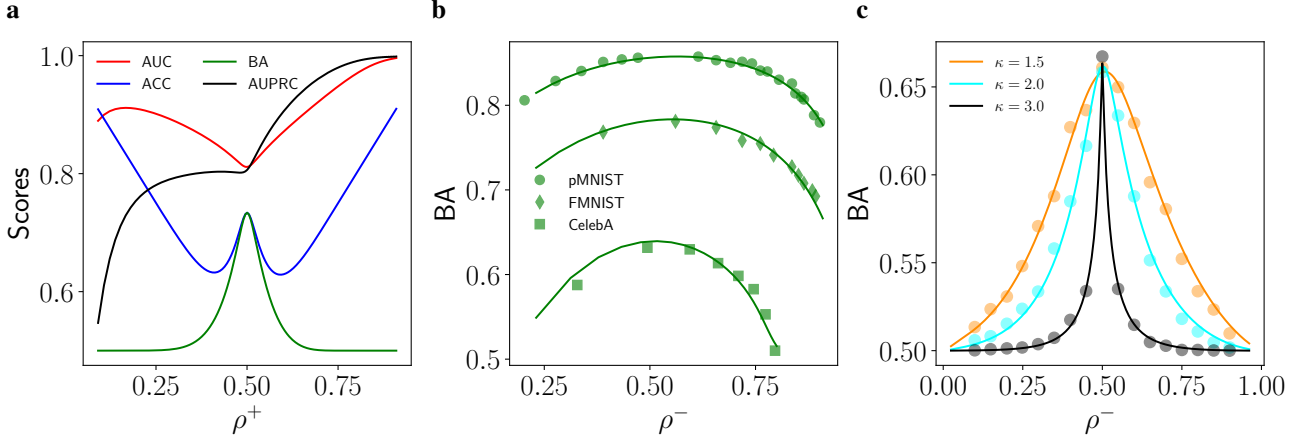
Figure 2. Analytical results derived within our framework. **a)** Different metrics on synthetic data evaluated on a test set having same train set composition. Here $L = 100$, $\kappa = 2$, $\alpha^- = 5$ with C, δ sampled randomly. **b)** Analytical predictions for benchmark datasets (MNIST, FashionMNIST and CelebA). Dots show numerical simulations, averaged over 10 trials for $\kappa = 0.5$. **c)** Balanced accuracy curves as a function of ρ^-. The margin κ of the algorithm controls the balance-to-performance trade-off. Dots correspond to numerical simulations with $\alpha^+ = 2$, $Q = 2$, $L = 100$, position-independent $\mathbf{M}$, normally-distributed δ, and diagonal covariance C averaged over 50 trials.

results in an absolute gain of accuracy as a function of ρ^-, but narrows down the window outside of which the classifier acts as a random one.

We can get intuition on the existence of such trade-off by looking at the high-κ and $\rho^+ = 0.5$ case. In this regime, the bias of the algorithm b is then factoring out the center-of-mass $\mathbf{M}$ of the data, while the weight vector $\mathbf{J}$ becomes fully aligned with the displacement δ between the two classes, yielding the optimal solution for the classification task. However, as soon as one moves away from the balanced training set, the extremely steep behaviour of the parameters $\mathbf{J}$ and b around $\rho^+ = \rho^- = 0.5$ – due to the high margin – makes the performance drop down (see Figure 5a in the Appendix A). For milder values of κ, the effect is less pronounced, even though $\mathbf{J}$ and b are not optimal.

For real data, the value of the margin κ of the linear classifier has to be compared to the data statistics to determine how tight the balance-to-performance trade-off is. Theory suggests that what matters is an effective margin depending on the ratio $\eta(C)/\|\delta\|_2^2$, where $\eta(C)$ denotes the largest eigenvalue of the covariance matrix. We check this on pMNIST and CelebA datasets, where the ratio $\eta(C)/\|\delta\|_2^2$ is larger for CelebA data and thus results in a tighter good performance window (see Figure 2b).

Hence, from the interplay between κ and data statistics and composition arises the balance-to-performance trade-off.

3.4. Insights on standard over vs. undersampling

The theoretical framework of Sec. 2 allows us to gain some understanding on *whether it is better to restore balance by oversampling the minority class or undersampling the*

majority class. With no loss of generality we assume the minority class to be the positive one; hence, the negative class size is $\alpha^- > \alpha^+$.
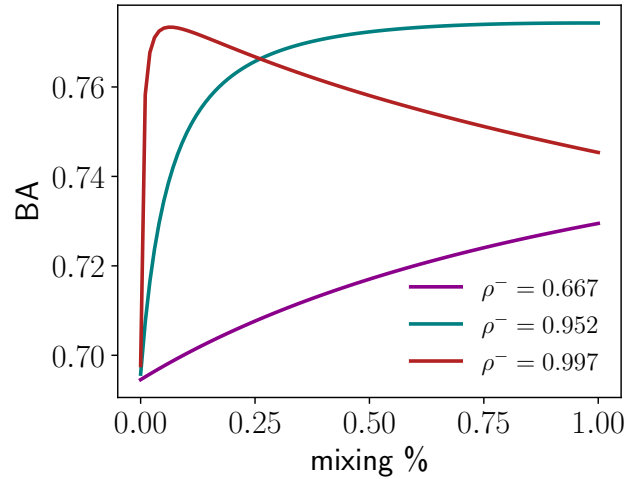

Figure 3. Optimal mixing strategy. We report theoretical predictions for the BA metric as a function of the mixing under/oversampling percentage in the training set. Depending on the initial training set composition (ρ^+, ρ^-), one can select the optimal strategy to restore balance. Here $L = 100$, $\kappa = 0.5$, with C and δ sampled randomly.

The full undersampling strategy consists in randomly removing negative data points down to $\alpha^- = \alpha^+$, a strategy called Random Undersampling (RUS). To oversample positive data, we modify Eq. (3) by introducing a factor $c_\mu \geq 1$ in front of the loss function for positive samples, accounting for their multiplicity due to duplication. For the sake of simplicity, we assume $\langle c_\mu \rangle = c \in \mathbb{R}$, *i.e.* we uniformly

sample each positive data c times, with $c \in [1, \alpha^-/\alpha^+]$. The limit case $c = 1$ corresponds to leaving the dataset as it is (RUS), while when $c = \alpha^-/\alpha^+$ the minority class is duplicated up to the majority class size, a strategy known as Random Oversampling (ROS). In this respect, ROS is theoretically equivalent to a simple loss-reweighting strategy that scales the contribution of each class to the total loss by the inverse of their size. Intermediate strategies mixing under and oversampling are associated to intermediate values of c. Based on our observation from the previous section, we study only the case of balanced training set looking at BA – even though our theory allows to make prediction for any training set composition and metrics. Thus we define the mixing percentage as

$$\text{mixing } \% = \frac{\alpha^+}{\alpha^- - \alpha^+}(c - 1), \qquad (14)$$

and we ask what is the optimal mixing percentage in terms of BA, *i.e.* what is the best c value?

In Figure 3 we observe that for all tested cases, full undersampling is a sub-optimal technique among restoring balance protocols, while mixed under/oversampling and oversampling should be preferred. In particular, for severe imbalance a mixed strategy gives highest BA compared to full oversampling, which should be preferred when the imbalance is milder: this is also numerically observed on real data in Figure 4b, dashed lines. However, we expect that in general the optimal strategy strongly depends not only on the imbalance ratio between classes, but on absolute dataset size, on classes' similarity in feature space and on the margin κ used in the linear classifier.

In this respect, our theoretical framework can be useful as it offers a quantitative way to evaluate *a priori* the performances based on the first- and second-order statistic of the data, thus providing a guideline in practical applications.

4. Numerical experiments

In this section we investigate with numerical experiments if the phenomenological findings at the theoretical level hold when we go beyond the limits of our theory. Specifically, we developed a framework for liner SVMs and for ROS/RUS techniques. We now ask what happens with deeper neural networks and more involved under/oversampling techniques. We address these two questions separately, to factor out any other element.

4.1. Improved sampling strategies

First, we examine the effect of sampling strategies more involved than random sampling, resorting to Restricted Boltzmann Machines (RBMs) that allow to under and oversample at the same time with a protocol we call Likelihood-

Informed Sampling (LIS). We train the model on the positive (minor) class solely and use it to generate new positive digits and to subsample negative digits based on their likelihood. A closely related approach to generate new examples with RBMs has been introduced in Zięba et al. (2015).

Restricted Boltzmann machines. RBMs are bipartite graphical models, and include L visible units $\mathbf{v} = (v_1, \cdots, v_L)$ and M latent (or hidden) units $\mathbf{z} = (z_1, \cdots, z_M)$. Only connections between visible and latent units are allowed through the interaction weights $w_{i\mu}$. RBMs define a joint probability distribution over $\mathbf{v}$ and $\mathbf{z}$ as the Gibbs distribution

$$p(\mathbf{v}, \mathbf{z}) = \frac{1}{Z} \exp \left\{ \sum_{i=1}^{L} h_i(v_i) - \sum_{\mu=1}^{M} U_\mu(z_\mu) \right.$$
$$\left. + \sum_{\mu=1}^{M} \sum_{i=1}^{L} w_{i\mu}(v_i) z_\mu \right\}, \quad (15)$$

where the h_i's and U_μ's act on, respectively, visible and latent units, and the last term couples the latent and visible layers. Learning of the model parameters is achieved by maximizing the marginal likelihood $p(\mathbf{v}) \equiv \int d\mathbf{z}\, p(\mathbf{v}, \mathbf{z})$ over the training set. Details about the numerical implementation can be found in Appendix G. After training and validation, the model can be used to score any new data $\mathbf{v}'$ through $\log p(\mathbf{v}')$.

Likelihood-informed sampling. We use RBM scores to form an appropriate balanced dataset by oversampling the minority class or undersampling the majority one, as follows.

(i) for oversampling, we rely on the generative power of our unsupervised architecture. Starting from an initial configuration, we perform Gibbs sampling in data space based on the model scores. We retain only samples having scores similar to the ones in the minority data class, as low-score samples are considered not representative of the class.

(ii) for undersampling, we filter out majority class samples at the boundaries of the log-likelihood distribution, as they are potentially less informative to the classification task.

Results. We perform numerical experiments on MNIST (classes with smaller and larger than digit 5) and with linear SVMs. Using RBMs for LIS we show that there is an improvement of performances as it was happening for ROS/RUS, even when starting from low positive class sizes; moreover, this sampling strategy yields better results compared to naive random under/oversamplings (see Figure 4a). The effect of restoring balance of data can be seen through a

geometric visualization of the problem. We train separately two linear SVMs with imbalanced ERM and with ERM after restoring balance between the two classes of the training set. We then project the test set data in the 2d space defined by the two optimal decision boundaries found by the classifier (see Figure 4b). We observe that data balance between classes allows the model to place a more effective decision boundary (gold line), while imbalanced ERM (purple line) only learns the majority class.

4.2. Deep classifiers

We consider an ImageNet pretrained ResNet-50 and finetune it on a binarized version of Cifar10 (*i.e.* images are split in two classes – positive and negative – depending if their label is smaller or larger than 5). To understand if the imbalance plays a relevant role similarly to the case with linear SVMs, we run two parallel experiments with imbalanced and balanced training set, $P = 0.1N$ and $P = N$ respectively. All network parameters in the two experiments stay the same. We show that the balanced experiments achieves higher accuracy by looking at the last feature layer of the network (see Figure 4c).

5. Discussion and perspectives

In this work we devise a theoretical framework to study the generalization performance of linear models for binary classification tasks under imbalanced training datasets. We show which metrics are more informative than others in imbalance learning and we give sharp estimations on the optimal mixing of under and oversampling strategy given the data statistics. We extend our study beyond the limit of the theoretical analysis, supporting our findings with numerical investigations.

To the best of our knowledge, this is the first asymptotic characterization of linear classifiers trained over imbalanced datasets for any generalization metrics and categorical data. Looking forward, we believe that while restoring data balance techniques are unanimously empirically beneficial for ERM, their theoretical understanding falls behind, and our work can help closing this gap. In this regard, our study can be extended along several directions. First, our theoretical setting assumes that the two classes of data differ by their first-order statistics while share the same second-order one. We would like to pursue this direction and derive analytical predictions for *heteroscedastic* classes, motivated by the stream of results obtained for high-dimensional data with non-trivial structures, such as correlated patterns (Monasson, 1992; Lopez et al., 1995), Gaussian covariate and mixture models (Loureiro et al., 2021), random features models (Goldt et al., 2020), object manifolds (Chung et al., 2018), simplex learning (Pastore et al., 2020; Rotondo et al., 2020; Pastore, 2021; Baroffio et al., 2024).

Moreover, it would be interesting to investigate, in light of our findings, the behavior of deep classifiers: how does the sensitivity to class imbalance scale with the expressive power of the architecture? For a recent review on the phenomenology, see Ghosh et al. (2022). While deep learning models in general remain analytically intractable, theoretical understanding has been reached in certain asymptotic regimes (mainly, the infinite-width limit and the proportional scaling between width and size of the training set), both for fully connected (Jacot et al., 2018; Lee et al., 2018; Pacelli et al., 2023; Cui et al., 2023; 2024) and convolutional architectures (Naveh & Ringel, 2021; Aiudi et al., 2023).

Analysis in a framework similar to the one we provide has been proven to work not only for linear models, but also for kernel machines (Dietrich et al., 1999; Gerace et al., 2020; Bordelon et al., 2020; Aguirre-López et al., 2024). We expect our approach to be easily generalizable to these cases, by simply replacing the input space we consider here with the feature space of those models, and the input statistics (means and covariances) with the one of the features (means of the features, kernel in feature space). Deep neural networks have been proven equivalent to kernel machines in certain cases, namely the infinite-width limit mentioned above; in this sense, kernels provide lower bounds on the performance of deep neural networks that work in more generic regimes of feature learning. Therefore, our findings, which are sharp for linear models and could be extended to kernel machines, can provide insights on generalization also for deep neural networks. Considering the large use of deep learning in practical applications, a theoretical assessment of its properties in the case of imbalanced data is very much needed.

Lastly, our theory could be extended beyond binary classification as in several applications one faces imbalance within the context of multiclass classification tasks.

As for numerical experiments, we introduced a strategy to restore balance based on unsupervised probabilistic models proving that it improves performance compared to naive imbalanced ERM. At a speculative level, RBMs have the potential to capture interpretable features of the minority class, while saving computational cost compared to deeper architectures. This restoring balance approach could also benefit from recent progress on efficient out-of-equilibrium sampling with RBMs (Agoritsas et al., 2023; Carbone et al., 2023). We believe that similar increase in performances would follow from restoring data balance with other methods (Mirza et al., 2021). Geometrically, we showed that restoring balance allowed one to remove the bias of the decision boundary towards the majority class only, as also observed in Chaudhuri et al. (2023).
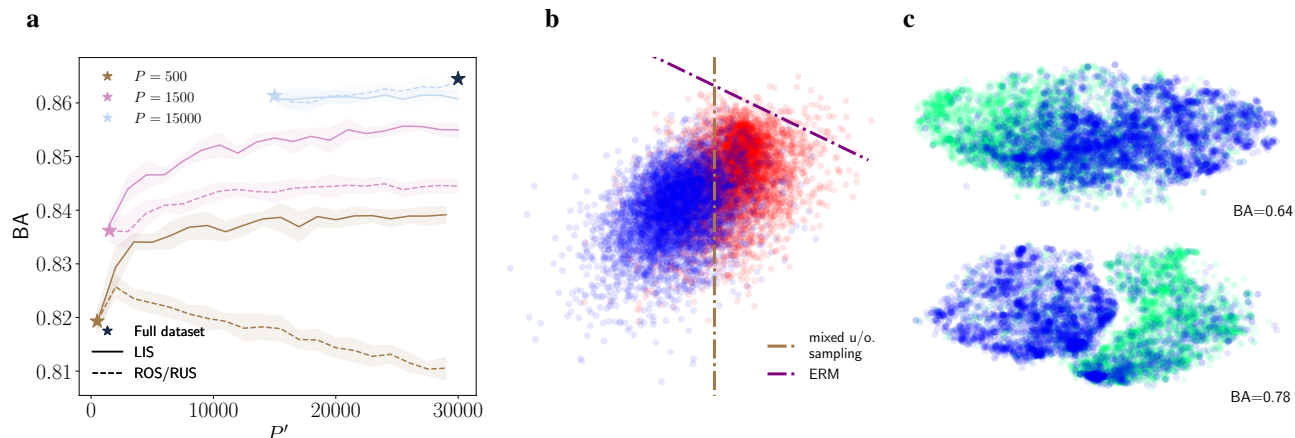
Figure 4. Numerical investigation on improved sampling techniques and deep classifier ResNet-50. **a)** Mixed sampling strategies to obtain a balanced training set for classification with a linear SVM on binary MNIST, as a function of the new sample size P'. As random sampling techniques, also higher level methods lead to an increase of performance. **b)** Geometrical interpretation of the effect of restoring balance (gold line) on the decision boundary of a linear SVM compared to imbalanced ERM (purple line). Data points are the MNIST test set. **c)** We visualize test data classification in the last feature layer of the network through tSNE, for imbalanced and balanced training set (top and bottom, respectively). The network trained on balanced data achieves improved performances, separating better the two classes.

Impact Statement

This paper aims to advance the field of Machine Learning. Class imbalance in datasets represents a common issue in many Machine Learning applications, since it undermines performance and makes predictions hard to assess with adequate metrics. We consider crucial to have a theoretical control on how imbalance impacts learning, and how it can be alleviated.

Acknowledgements

References

Agoritsas, E., Biroli, G., Urbani, P., and Zamponi, F. 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.

Agoritsas, E., Catania, G., Decelle, A., and Seoane, B. Explaining the effects of non-convergent MCMC in the training of energy-based models. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 322–336. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/

v202/agoritsas23a.html.

Aguirre-López, F., Franz, S., and Pastore, M. Random features and polynomial rules, 2024.

Ai, Q., Wang, P., He, L., Wen, L., Pan, L., and Xu, Z. Generative oversampling for imbalanced data via majority-guided vae. In Ruiz, F., Dy, J., and van de Meent, J.-W. (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 3315–3330. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/ai23a.html.

Aiudi, R., Pacelli, R., Vezzani, A., Burioni, R., and Rotondo, P. Local kernel renormalization as a mechanism for feature learning in overparametrized convolutional neural networks, 2023.

Akbani, R., Kwek, S., and Japkowicz, N. Applying support vector machines to imbalanced datasets. In Boulicaut, J.-F., Esposito, F., Giannotti, F., and Pedreschi, D. (eds.), *Machine Learning: ECML 2004*, pp. 39–50, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg. ISBN 978-3-540-30115-8. doi: 10.1007/978-3-540-30115-8_7.

Alshammari, S., Wang, Y.-X., Ramanan, D., and Kong, S. Long-tailed recognition via weight balancing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6897–6907, June 2022.

Anand, R., Mehrotra, K., Mohan, C., and Ranka, S. An improved algorithm for neural network classification of imbalanced training sets. *IEEE Transactions on Neural Networks*, 4(6):962–969, 1993. doi: 10.1109/72.286891.

Ansari, M. and White, A. D. Learning peptide properties with positive examples only, 2023. URL https://www.biorxiv.org/content/early/2023/06/05/2023.06.01.543289.

Baroffio, A., Rotondo, P., and Gherardi, M. Resolution of similar patterns in a solvable model of unsupervised deep learning with structured data. *Chaos, Solitons & Fractals*, 182:114848, 2024. ISSN 0960-0779. doi: https://doi.org/10.1016/j.chaos.2024.114848. URL https://www.sciencedirect.com/science/article/pii/S0960077924004004.

Batista, G. E. A. P. A., Prati, R. C., and Monard, M. C. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explor. Newsl.*, 6(1):20–29, jun 2004. ISSN 1931-0145. doi: 10.1145/1007730.1007735. URL https://doi.org/10.1145/1007730.1007735.

Bordelon, B., Canatar, A., and Pehlevan, C. Spectrum dependent learning curves in kernel regression and wide neural networks. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1024–1034. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/bordelon20a.html.

Buda, M., Maki, A., and Mazurowski, M. A. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018. ISSN 0893-6080. doi: 10.1016/j.neunet.2018.07.011. URL https://www.sciencedirect.com/science/article/pii/S0893608018302107.

Carbone, A., Decelle, A., Rosset, L., and Seoane, B. Fast and functional structured data generator. In *ICML 2023 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2023. URL https://openreview.net/forum?id=uXkfPvjYeM.

Chaudhuri, K., Ahuja, K., Arjovsky, M., and Lopez-Paz, D. Why does throwing away data improve worst-group error? In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 4144–4188. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/chaudhuri23a.html.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002. doi: 10.1613/jair.953.

Cheng, Z., Zhou, S., and Guan, J. Computationally predicting protein-RNA interactions using only positive and unlabeled examples. *Journal of Bioinformatics and Computational Biology*, 13(03):1541005, 2015. doi: 10.1142/S021972001541005X. URL https://doi.org/10.1142/S021972001541005X.

Chung, S., Lee, D. D., and Sompolinsky, H. Classification and geometry of general perceptual manifolds. *Phys. Rev. X*, 8:031003, Jul 2018. doi: 10.1103/PhysRevX.8.031003. URL https://link.aps.org/doi/10.1103/PhysRevX.8.031003.

Cui, H., Krzakala, F., and Zdeborová, L. Bayes-optimal learning of deep random networks of extensive-width, 2023.

Cui, H., Pesce, L., Dandi, Y., Krzakala, F., Lu, Y. M., Zdeborová, L., and Loureiro, B. Asymptotics of feature learning in two-layer networks after one gradient-step, 2024.

Dandi, Y., Stephan, L., Krzakala, F., Loureiro, B., and Zdeborová, L. Universality laws for gaussian mixtures in generalized linear models. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 54754–54768. Curran Associates, Inc., 2023.

Del Giudice, P., Franz, S., and Virasoro, M. A. Perceptron beyond the limit of capacity. *J. Phys. France*, 50(2):121–134, 1989. doi: 10.1051/jphys:01989005002012100. URL https://doi.org/10.1051/jphys:01989005002012100.

Dietrich, R., Opper, M., and Sompolinsky, H. Statistical mechanics of support vector networks. *Phys. Rev. Lett.*, 82:2975–2978, 04 1999. doi: 10.1103/PhysRevLett.82.2975. URL https://link.aps.org/doi/10.1103/PhysRevLett.82.2975.

Douzas, G. and Bacao, F. Effective data generation for imbalanced learning using conditional generative adversarial networks. *Expert Systems with Applications*, 91:464–471, 2018. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2017.09.030. URL https://www.sciencedirect.com/science/article/pii/S0957417417306346.

Engel, A. and Van den Broeck, C. *Statistical Mechanics of Learning*. Cambridge University Press, 2001. doi: 10.1017/CBO9781139164542.

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., and Herrera, F. *Learning from Imbalanced Data Sets*. Springer Cham, 2018. doi: 10.1007/978-3-319-98074-4.

Fotouhi, S., Asadi, S., and Kattan, M. W. A comprehensive data level analysis for cancer diagnosis on imbalanced data. *Journal of Biomedical Informatics*, 90:103089, 2019. ISSN 1532-0464. doi: 10.1016/j.jbi.2018.12.003. URL https://www.sciencedirect.com/science/article/pii/S1532046418302302.

Francazi, E., Baity-Jesi, M., and Lucchi, A. A theoretical analysis of the learning dynamics under class imbalance, 2023.

Franz, S., Parisi, G., Sevelev, M., Urbani, P., and Zamponi, F. Universality of the SAT-UNSAT (jamming) threshold in non-convex continuous constraint satisfaction problems. *SciPost Phys.*, 2:019, 2017. doi: 10.21468/SciPostPhys.2.3.019. URL https://scipost.org/10.21468/SciPostPhys.2.3.019.

Franz, S., Sclocchi, A., and Urbani, P. Critical jammed phase of the linear perceptron. *Phys. Rev. Lett.*, 123:115702, Sep 2019. doi: 10.1103/PhysRevLett.123.115702. URL https://link.aps.org/doi/10.1103/PhysRevLett.123.115702.

Gardner, E. The space of interactions in neural network models. *Journal of Physics A: Mathematical and General*, 21(1):257, jan 1988. doi: 10.1088/0305-4470/21/1/030. URL https://dx.doi.org/10.1088/0305-4470/21/1/030.

Gerace, F., Loureiro, B., Krzakala, F., Mezard, M., and Zdeborova, L. Generalisation error in learning with random features and the hidden manifold model. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3452–3462. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/gerace20a.html.

Ghosh, K., Bellinger, C., Corizzo, R., Branco, P., Krawczyk, B., and Japkowicz, N. The class imbalance problem in deep learning. *Machine Learning*, Dec 2022. ISSN 1573-0565. URL https://doi.org/10.1007/s10994-022-06268-8.

Goldt, S., Mézard, M., Krzakala, F., and Zdeborová, L. Modeling the influence of data structure on learning in neural networks: The hidden manifold model. *Phys. Rev. X*, 10:041044, Dec 2020. doi: 10.1103/PhysRevX.10.041044. URL https://link.aps.org/doi/10.1103/PhysRevX.10.041044.

He, H. and Ma, Y. (eds.). *Imbalanced Learning: Foundations, Algorithms, and Applications*. John Wiley & Sons, Ltd, 2013. ISBN 9781118646106. doi: 10.1002/9781118646106. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118646106.

He, H., Bai, Y., Garcia, E. A., and Li, S. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pp. 1322–1328, 2008. doi: 10.1109/IJCNN.2008.4633969.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/5a4be1fa34e62bb8a6ec6b91d2462f5a-Paper.pdf.

Japkowicz, N. and Stephen, S. The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6:429–449, 2002. ISSN 1571-4128. doi: 10.3233/IDA-2002-6504. URL https://doi.org/10.3233/IDA-2002-6504. 5.

Johnson, J. M. and Khoshgoftaar, T. M. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):27, Mar 2019. ISSN 2196-1115. doi: 10.1186/s40537-019-0192-5. URL https://doi.org/10.1186/s40537-019-0192-5.

Jones, E., Oliphant, T., and Peterson, P. SciPy: open source scientific tools for Python, 2001. URL http://www.scipy.org.

Jurafsky, D. and Martin, J. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall series in artificial intelligence. Pearson Prentice Hall, 2009. ISBN 9780131873216. URL https://books.google.fr/books?id=fZmj5UNK8AQC.

Kepler, T. B. and Abbott, L. Domains of attraction in neural networks. *J. Phys. France*, 49(10):1657–1662, 1988. doi: 10.1051/jphys:0198800490100165700. URL https://doi.org/10.1051/jphys:0198800490100165700.

Kovács, G. An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets. *Applied Soft Computing*, 83:105662, 2019. ISSN 1568-4946. doi: https://doi.org/10.1016/j.asoc.2019.105662. URL https://www.sciencedirect.com/science/article/pii/S1568494619304429.

Krawczyk, B., Galar, M., Łukasz Jeleń, and Herrera, F. Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy. *Applied Soft Computing*, 38:714–726, 2016. ISSN 1568-4946. doi: https://doi.org/10.1016/j.asoc.2015.08. 060. URL https://www.sciencedirect.com/ science/article/pii/S1568494615005815.

Kubat, M. and Matwin, S. Addressing the curse of imbalanced training sets: one-sided selection. In *14th International Conference on Machine Learning(ICML97)*, pp. 179–186, 1997. ISBN 1558604863.

Laurikkala, J. Improving identification of difficult small classes by balancing class distribution. In Quaglini, S., Barahona, P., and Andreassen, S. (eds.), *Artificial Intelligence in Medicine*, pp. 63–66, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg. ISBN 978-3-540-48229-1.

Lee, J., Sohl-dickstein, J., Pennington, J., Novak, R., Schoenholz, S., and Bahri, Y. Deep neural networks as gaussian processes. In *International Conference on Learning Representations*, 2018. URL https:// openreview.net/forum?id=B1EA-M-0Z.

Lemaître, G., Nogueira, F., and Aridas, C. K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017. URL http: //jmlr.org/papers/v18/16-365.html.

Lemnaru, C. and Potolea, R. Imbalanced classification problems: Systematic study, issues and best practices. In Zhang, R., Zhang, J., Zhang, Z., Filipe, J., and Cordeiro, J. (eds.), *Enterprise Information Systems*, pp. 35–50, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-29958-2. doi: 10.1007/978-3-642-29958-2_3.

Li, H., Helling, R., Tang, C., and Wingreen, N. Emergence of preferred structures in a simple model of protein folding. *Science*, 273(5275):666–669, 1996.

Liu, S., Wang, Y., Zhang, J., Chen, C., and Xiang, Y. Addressing the class imbalance problem in twitter spam detection using ensemble learning. *Computers & Security*, 69:35–49, 2017. ISSN 0167-4048. doi: 10.1016/j.cose.2016.12. 004. URL https://www.sciencedirect.com/ science/article/pii/S0167404816301754.

Liu, Y., Loh, H. T., and Sun, A. Imbalanced text classification: A term weighting approach. *Expert Systems with Applications*, 36(1):690–701, 2009. ISSN 0957-4174. doi: 10.1016/j.eswa.2007.10. 042. URL https://www.sciencedirect.com/ science/article/pii/S0957417407005350.

Lopez, B., Schroder, M., and Opper, M. Storage of correlated patterns in a perceptron. *Journal of Physics A: Mathematical and General*, 28(16):L447, aug 1995. doi: 10.1088/0305-4470/28/16/005. URL https://dx. doi.org/10.1088/0305-4470/28/16/005.

Loureiro, B., Sicuro, G., Gerbelot, C., Pacco, A., Krzakala, F., and Zdeborová, L. Learning gaussian mixtures with generalized linear models: Precise asymptotics in high-dimensions. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 10144–10157. Curran Associates, Inc., 2021. URL https://proceedings.neurips. cc/paper_files/paper/2021/file/ 543e83748234f7cbab21aa0ade66565f-Paper. pdf.

Mannelli, S. S., Gerace, F., Rostamzadeh, N., and Saglietti, L. Bias-inducing geometries: an exactly solvable data model with fairness implications, 2023.

Menon, A., Narasimhan, H., Agarwal, S., and Chawla, S. On the statistical consistency of algorithms for binary classification under class imbalance. In *International Conference on Machine Learning*, pp. 603–611. PMLR, 2013.

Mezard, M. The space of interactions in neural networks: Gardner's computation with the cavity method. *Journal of Physics A: Mathematical and General*, 22 (12):2181, jun 1989. doi: 10.1088/0305-4470/22/ 12/018. URL https://dx.doi.org/10.1088/ 0305-4470/22/12/018.

Mignacco, F., Krzakala, F., Lu, Y., Urbani, P., and Zdeborova, L. The role of regularization in classification of high-dimensional noisy Gaussian mixture. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6874–6883. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/ v119/mignacco20a.html.

Mirny, L. and Shakhnovich, E. Protein folding theory: from lattice to all-atom models. *Annual review of biophysics and biomolecular structure*, 30(1):361–396, 2001.

Mirza, B., Haroon, D., Khan, B., Padhani, A., and Syed, T. Q. Deep generative models to counter class imbalance: A model-metric mapping with proportion calibration methodology. *IEEE Access*, 9:55879–55897, 2021.

Monasson, R. Properties of neural networks storing spatially correlated patterns. *Journal of Physics A: Mathematical and General*, 25(13):3701, jul 1992. doi:

10.1088/0305-4470/25/13/019. URL https://dx.doi.org/10.1088/0305-4470/25/13/019.

Mondal, A. K., Singhal, L., Tiwary, P., Singla, P., and AP, P. Minority oversampling for imbalanced data via class-preserving regularized auto-encoders. In Ruiz, F., Dy, J., and van de Meent, J.-W. (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 3440–3465. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/mondal23a.html.

Nadal, J.-P. and Rau, A. Storage capacity of a potts-perceptron. *J. Phys. I France*, 1(8):1109–1121, 1991. doi: 10.1051/jp1:1991104. URL https://doi.org/10.1051/jp1:1991104.

Naveh, G. and Ringel, Z. A self consistent theory of gaussian processes captures feature learning effects in finite CNNs. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 21352–21364. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/b24d21019de5e59da180f1661904f49a-Paper.pdf.

Pacelli, R., Ariosto, S., Pastore, M., Ginelli, F., Gherardi, M., and Rotondo, P. A statistical mechanics framework for bayesian deep neural networks beyond the infinite-width limit, Dec 2023. ISSN 2522-5839. URL https://doi.org/10.1038/s42256-023-00767-6.

Pastore, M. Critical properties of the sat/unsat transitions in the classification problem of structured data. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(11):113301, 11 2021. doi: 10.1088/1742-5468/ac312b. URL https://dx.doi.org/10.1088/1742-5468/ac312b.

Pastore, M., Rotondo, P., Erba, V., and Gherardi, M. Statistical learning theory of structured data. *Phys. Rev. E*, 102:032119, Sep 2020. doi: 10.1103/PhysRevE.102.032119. URL https://link.aps.org/doi/10.1103/PhysRevE.102.032119.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. URL http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.

Pesce, L., Krzakala, F., Loureiro, B., and Stephan, L. Are Gaussian data all you need? The extents and limits of universality in high-dimensional generalized linear estimation. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 27680–27708. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/pesce23a.html.

Rotondo, P., Pastore, M., and Gherardi, M. Beyond the storage capacity: Data-driven satisfiability transition. *Phys. Rev. Lett.*, 125:120601, Sep 2020. doi: 10.1103/PhysRevLett.125.120601. URL https://link.aps.org/doi/10.1103/PhysRevLett.125.120601.

Sauvola, J. and Pietikäinen, M. Adaptive document image binarization. *Pattern Recognition*, 33(2):225–236, 2000. ISSN 0031-3203. doi: https://doi.org/10.1016/S0031-3203(99)00055-2. URL https://www.sciencedirect.com/science/article/pii/S0031320399000552.

Shakhnovich, E. and Gutin, A. Enumeration of all compact conformations of copolymers with random sequence of links. *The Journal of Chemical Physics*, 93(8):5967–5971, 1990.

Shang, H., Langlois, J.-M., Tsioutsiouliklis, K., and Kang, C. Precision/recall on imbalanced test data. In Ruiz, F., Dy, J., and van de Meent, J.-W. (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 9879–9891. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/shang23a.html.

Song, H., Bremer, B. J., Hinds, E. C., Raskutti, G., and Romero, P. A. Inferring protein sequence-function relationships with large-scale positive-unlabeled learning. *Cell Systems*, 12(1):92–101.e8, 2021. ISSN 2405-4712. doi: 10.1016/j.cels.2020.10.007. URL https://www.sciencedirect.com/science/article/pii/S2405471220304142.

Tieleman, T. Training restricted boltzmann machines using approximations to the likelihood gradient. In *Proceedings of the 25th international conference on Machine learning*, pp. 1064–1071, 2008.

Wan, Z., Zhang, Y., and He, H. Variational autoencoder based synthetic data generation for imbalanced learning. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1–7, 2017. doi: 10.1109/SSCI.2017.8285168.

Wang, C., Ding, C., Meraz, R. F., and Holbrook, S. R. PSoL: a positive sample only learning algorithm for finding non-coding RNA genes. *Bioinformatics*, 22 (21):2590–2596, 08 2006. ISSN 1367-4803. doi: 10.1093/bioinformatics/btl441. URL https://doi.org/10.1093/bioinformatics/btl441.

Weiss, G. M. Mining with rarity: A unifying framework. *SIGKDD Explor. Newsl.*, 6(1):7–19, jun 2004. ISSN 1931-0145. doi: 10.1145/1007730.1007734. URL https://doi.org/10.1145/1007730.1007734.

Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Yan, Z., Tomizuka, M., Gonzalez, J., Keutzer, K., and Vajda, P. Visual transformers: Token-based image representation and processing for computer vision, 2020.

Yang, P., Li, X.-L., Mei, J.-P., Kwoh, C.-K., and Ng, S.-K. Positive-unlabeled learning for disease gene identification. *Bioinformatics*, 28(20):2640–2647, 08 2012. ISSN 1367-4803. doi: 10.1093/bioinformatics/bts504. URL https://doi.org/10.1093/bioinformatics/bts504.

Zięba, M., Tomczak, J. M., and Gonczarek, A. RBM-SMOTE: Restricted boltzmann machines for synthetic minority oversampling technique. In Nguyen, N. T., Trawiński, B., and Kosala, R. (eds.), *Intelligent Information and Database Systems*, pp. 377–386, Cham, 2015. Springer International Publishing. ISBN 978-3-319-15702-3.

A. Details on the replica calculation

We give here all the details in order to get the theoretical predictions sketched in Sec. 2 of the main text. The measure over the weights for the two cases of the spherical perceptron and the SVM in Eq. (7) is defined as

$$
d\mu(\mathbf{J}) = \begin{cases} \prod_{i,t} \dfrac{dJ_i(t)}{\Omega_0} \delta[\sum_u J_i(u)]\delta(\|\mathbf{J}\|_2^2 - LQ) & \text{(perceptron)}, \\[2ex] \prod_{i,t} \dfrac{dJ_i(t)}{\Omega_0} \delta[\sum_u J_i(u)] \exp\left(-\dfrac{\beta\lambda}{2}\|\mathbf{J}\|_2^2\right) & \text{(SVM)}, \end{cases} \tag{16}
$$

where Ω_0 is the corresponding normalization factor and for convenience we included the L2 regularization in the measure of the SVM. The common delta functions are enforcing the zero-sum conditions $\sum_t J_i(t) = 0$ for all i mentioned in Sec. 2. We report in the following the case of the spherical perceptron, mentioning where the calculation differs crucially from the case of the SVM in Sec. A.6.

A.1. Replicated partition function

To evaluate the expected value of the logarithm of the partition function appearing in Eq. (10), we resort to the so-called replica trick from statistical physics: writing $\log \Omega = \lim_{n\to 0^+} (\Omega^n - 1)/n$, we convert the quenched averaged over the training data into an annealed average of the n-times replicated partition function Ω^n. This quantity can be written as

$$
\mathbb{E}_{\text{data}}[\Omega^n] = \int \prod_{a=1}^n db^a \, d\mu(\mathbf{J}^a) \prod_{y\in\{\pm\}} \int \prod_{\mu=1}^{\alpha^y L} \frac{d\Delta_\mu^a \, d\hat{\Delta}_\mu^a}{2\pi} e^{-\beta(\kappa-y\Delta_\mu^a)\theta(\kappa-y\Delta_\mu^a)} e^{i\sum_{a,\mu}\hat{\Delta}_\mu^a(\Delta_\mu^a+b)} \left\langle e^{-i\sum_{a,\mu}\hat{\Delta}_\mu^a \frac{\mathbf{J}^a\cdot\mathbf{s}^\mu}{\sqrt{L}}} \right\rangle_y
$$

(17)

where we inserted delta-functions for the variables (replicated versions of the ones defined in Eq. (4))

$$
\Delta_\mu^a = \frac{\mathbf{J}^a \cdot \mathbf{s}^\mu}{\sqrt{L}} - b, \tag{18}
$$

using their Fourier-conjugates $\hat{\Delta}_\mu^a$. We introduce the order parameters

$$
r^a = \sum_{i,t} \frac{J_i^a(t)\delta_i(t)}{L}, \qquad q^{ab} = \sum_{i,j,t,u} \frac{J_i^a(t)J_j^b(u)}{L} C_{ij}(t,u), \tag{19}
$$

so that the average over the samples gives, for large L,

$$
\left\langle e^{-i\sum_{a,\mu}\hat{\Delta}_\mu^a \frac{\mathbf{J}^a\cdot\mathbf{s}^\mu}{\sqrt{L}}} \right\rangle_\pm \sim \exp\left[-\frac{1}{2}\sum_{a,b}\hat{\Delta}_\mu^a\hat{\Delta}_\mu^b q^{ab} \mp \frac{i}{2}\sum_a \hat{\Delta}_\mu^a r^a - \frac{i}{\sqrt{L}}\sum_a \hat{\Delta}_\mu^a \sum_{i,t} J_i^a(t)M_i(t) \right]. \tag{20}
$$

This asymptotic expansions is justified as long as higher moments in the distribution of the data can be neglected with respect to the first and second moments. The full integral to work out becomes

$$
\mathbb{E}_{\text{data}}[\Omega^n] = \int \prod_a db^a \prod_{i,t} \frac{dJ_i^a(t)}{\Omega_0} \delta\Big[\sum_u J_i^a(u)\Big]\delta(\|\mathbf{J}^a\|_2^2 - LQ) \int \prod_a dr^a \delta(\mathbf{J}^a\cdot\boldsymbol{\delta} - Lr^a)
$$

$$
\times \int \prod_{a\leq b} dq^{ab}\delta\Big(\sum_{i,j,t,u} J_i^a(t)J_i^b(u)C_{ij}(t,u) - Lq^{ab}\Big)
$$

(21)

$$
\times \prod_{y\in\{\pm\}} \int \prod_{a,\mu} \frac{d\Delta_\mu^a \, d\hat{\Delta}_\mu^a}{2\pi} e^{-\beta(\kappa-y\Delta_\mu^a)\theta(\kappa-y\Delta_\mu^a) - \frac{1}{2}\sum_{\mu,a,b}\hat{\Delta}_\mu^a\hat{\Delta}_\mu^b q^{ab} + i\sum_{\mu,a}\hat{\Delta}_\mu^a\left[\Delta_\mu^a + b^a - y\frac{r^a}{2} - \frac{1}{\sqrt{L}}\sum_{i,t} J_i^a(t)M_i(t)\right]}
$$

The last line depends explicitly on the weights $\mathbf{J}^a$ only through the combination $\mathbf{J}^a \cdot \mathbf{M}/\sqrt{L}$; however, as the threshold b^a is a parameter to be optimized, it is always possible to shift it with

$$
b^a - \frac{1}{\sqrt{L}}\sum_{i,t} J_i^a(t)M_i(t) \to b^a, \tag{22}
$$

15

in order to absorb this term.

To perform the integrals over the weights, we express all the delta-functions in their Fourier representation, introducing additional variables g_i^a (for the zero-sum conditions), $\hat{k}^a$ (for the spherical constraint) and $\hat{q}^{ab}$, $\hat{r}^a$ for the deltas enforcing the definition of the order parameters. Moreover, we impose the replica symmetric (RS) Ansatz

$$b^a = b, \quad r^a = r, \quad \hat{k}^a = \hat{k}, \quad \hat{r}^a = \hat{r}, \quad q^{ab} = q_0 \delta_{ab} + q(1 - \delta_{ab}), \quad \hat{q}^{ab} = \hat{q}_0 \delta_{ab} + \hat{q}(1 - \delta_{ab}) \tag{23}$$

on all the tensors with replica indices. This Ansatz is justified as long as the optimization problem in Eq. (2) is convex; though one may find a non-convex regime arising for some values of $\alpha^\pm$, $\boldsymbol{\delta}$ and C, as discussed in Franz et al. (2019), we expect it to be exact for many choices of the parameters. In practice, we find excellent agreement between our theoretical predictions and numerical experiments throughout this work (see Fig. 8).

Under the Ansatz (23), the high-dimensional integral over weights and threshold can be carried out and reduces to a four-dimensional integral over the order parameters b, r, q, q_0. The calculation can be split in an entropic contribution, from the integrals over the weights, and an energetic term, from the training data. At the end, the expected value in Eq. (21) can be written as

$$\mathbb{E}_{\text{data}}[\Omega^n] = \int \mathrm{d}b\,\mathrm{d}r\mathrm{d}q_0\mathrm{d}q \, \exp\left[\frac{n\beta L}{2}\left(G_S + \alpha^+ G_+ + \alpha^- G_-\right)\right], \tag{24}$$

where the functions of the order parameters G_S (entropic term) and $G_\pm$ (energetic term) are detailed in the following. The remaining integral over the order parameters can be estimated for L large through the saddle-point method.

A.2. Entropic term

First we evaluate the normalization of the measure $\mu(\mathbf{J})$ in Eq. (16). With our zero-sum conditions, compatible with the ones of Nadal & Rau (1991), this factor is given by

$$\begin{aligned}
\Omega_0 &= \int \prod_{i,t} dJ_i(t)\, \delta\Big[\sum_{i,t} J_i^2(t) - LQ\Big] \prod_i \delta\Big[\sum_t J_i(t)\Big] \\
&= \int \frac{d\hat{k}}{2\pi} \prod_i \frac{dg_i}{2\pi} \prod_t dJ_i(t) \, \exp\Big[i\hat{k}\sum_{i,t} J_i^2(t) - LQi\hat{k} + i\sum_i g_i \sum_t J_i(t)\Big] \\
&\sim \exp\Big[\frac{L(Q-1)}{2}[1 + \log(2\pi)] - \frac{L(Q-1)}{2}\log\Big(\frac{Q-1}{Q}\Big) - \frac{L}{2}\log(Q)\Big],
\end{aligned} \tag{25}$$

as the integrals first over $\mathbf{J}$ and then over $\mathbf{g}$ are Gaussian and the large-L result for the integral over $\hat{k}$ is obtained at the saddle point, which is located in $\hat{k} = i(Q-1)/(2Q)$.

In the RS case, the integral over the weights becomes

$$I = \int \prod_{a,i} \frac{\mathrm{d}g_i^a}{2\pi} \prod_t \mathrm{d}J_i^a(t) \exp\Big[-\frac{1}{2}\sum_{a,b}\sum_{t,u} J_i^a(t)\Sigma_{ij}^{ab}(t,u)J_i^b(u) - i\hat{r}\sum_{a,t} J_i^a(t)\delta_i(t) + i\sum_a g_i^a \sum_t J_i^a(t)\Big], \tag{26}$$

where the quadratic form Σ is given by

$$\Sigma_{ij}^{ab}(t,u) = \hat{k}\delta_{ij}\delta_{tu}\delta_{ab} + (2\hat{q}_0 - \hat{q})\delta_{ab}C_{ij}(t,u) + \hat{q}C_{ij}(t,u) \tag{27}$$

(to write this equation, we rescaled all the hat variables by $-i$). The evaluation of this Gaussian integral is straightforward but tedious. It can be performed, for example, factorizing the double sums over replica indices via Gaussian linearization (Hubbard-Stratonovich transformation). The result for small n is given by

$$I = \exp\left\{\frac{n}{2}\left[L(Q-1)\log(2\pi) - L\log\left(\frac{Q}{\hat{k}}\right) - \log\det A - \hat{q}\,\mathrm{tr}\big(A^{-1}C\big) + \hat{r}^2\boldsymbol{\delta}^\top A^{-1}\boldsymbol{\delta}\right]\right\}, \tag{28}$$

where all the algebraic operations (determinants, traces, vector/matrix multiplications) are defined on the space spanned by Potts and input indices, the matrix A is defined as

$$A = \hat{k}\mathbb{I}_Q \otimes \mathbb{I}_L + \hat{w}C \tag{29}$$

and $\hat{w} = 2\hat{q}_0 - \hat{q}$. Notice that, even if the matrix A is not invertible for $\hat{k} = 0$ (the matrix C has a spectrum with L zero eigenvalues, corresponding to the eigenvectors $v^{(\ell)}$ with components $v_i^{(\ell)}(t) = \delta_{i\ell}1$), the combinations appearing in this formula are always well defined; in particular, the factor $L \log \hat{k}$ comes from the zero-sum conditions, which regularize the divergence in $\log \det A$.

Summing over the input dimensions and normalizing by (25) we get, for the function G_S appearing in Eq. (24),

$$
\frac{n\beta L G_S}{2} = \log \int_{-i\infty}^{i\infty} \frac{\mathrm{d}\hat{k}}{4\pi} \frac{\mathrm{d}\hat{w}}{2\pi} \frac{\mathrm{d}\hat{q}}{2\pi} \frac{\mathrm{d}\hat{r}}{2\pi} e^{\frac{nL}{2}\left[Q\hat{k} + \hat{w}w + \hat{w}q + \hat{q}w + 2\hat{r}r\right] - \frac{nL(Q-1)}{2}\left[1 - \log\left(\frac{Q-1}{Q}\right)\right]}
$$

$$
\times e^{-\frac{n}{2}\left[\log \det A - L \log \hat{k} + \hat{q} \operatorname{tr}\left(A^{-1}C\right) - \hat{r}^2 \boldsymbol{\delta}^\top A^{-1} \boldsymbol{\delta}\right]}, \tag{30}
$$

where $w = q_0 - q$.

A.3. Energetic term

SAT phase. When the problem is linearly separable (low density α^+ and α^- of the classes, large distance $\boldsymbol{\delta}$ between their center), the perceptron at equilibrium reaches one of the many configurations corresponding to zero loss. This means that the optimization problem (2) becomes equivalent to a constraint statisfaction problem (CSP) where all the training data are required to be correctly classified by the decision boundary, namely in its satisfiable (SAT) phase. Formally, this means that the terms depending on the loss in Eq. (21) can be written as follows:

$$
e^{-\beta \sum_\mu (\kappa - y\Delta_\mu^a)\theta(\kappa - y\Delta_\mu^a)} \xrightarrow[\beta\to\infty]{} \prod_\mu \theta(y\Delta_\mu^a - \kappa). \tag{31}
$$

Thus, the part of the integral depending on the constraints from the training data reduces to the evaluation (respectively, P and N times) of the following two integrals

$$
I_\pm = \int \prod_a \frac{\mathrm{d}\Delta^a \, \mathrm{d}\hat{\Delta}^a}{2\pi} \theta(\pm\Delta^a - \kappa) e^{-\frac{w}{2}\sum_a (\hat{\Delta}^a)^2 - \frac{q}{2}\left(\sum_a \hat{\Delta}^a\right)^2 + i\sum_a \hat{\Delta}^a\left(\Delta^a + b \mp \frac{r}{2}\right)}. \tag{32}
$$

By Gaussian linearization,

$$
I_\pm = \int \mathcal{D}_q\xi \left[\int \frac{\mathrm{d}\Delta \, \mathrm{d}\hat{\Delta}}{2\pi} \theta(\pm\Delta - \kappa) e^{-\frac{w}{2}\hat{\Delta}^2 + i\hat{\Delta}\left(\Delta + b + \xi \mp \frac{r}{2}\right)}\right]^n, \tag{33}
$$

where we defined the Gaussian measure as $\mathcal{D}_q\xi = \mathrm{d}\xi\,\mathcal{N}(\xi|0, q)$. For small n, the final result is

$$
\log I_\pm \approx \frac{n\beta}{2} G_\pm, \tag{34}
$$

with

$$
\frac{\beta G_\pm}{2} = \int \mathcal{D}_q\xi \log\left[H\left(\frac{\kappa \pm b \mp \xi - r/2}{\sqrt{w}}\right)\right],
$$

$$
H(x) = \frac{1}{2}\operatorname{erfc}\left(\frac{x}{\sqrt{2}}\right) = 1 - H(-x). \tag{35}
$$

This equation defines the functions $G_\pm$ in Eq. (24) in the linearly separable phase.

UNSAT phase. When the training set is not linearly separable, the corresponding CSP is in its unsatisfiable phase. In this case, the loss gives a non-trivial energetic contribution for each mis-classified input point. Calculations can be done as in the previous paragraph, by noticing that

$$
e^{-\beta(\kappa \mp \lambda)\theta(\kappa \mp \lambda)} = e^{-\beta(\kappa \mp \lambda)} + \left[1 - e^{-\beta(\kappa \mp \lambda)}\right]\theta(-\kappa \pm \lambda). \tag{36}
$$

The result for the free energy contributions is

$$
\frac{\beta G_\pm}{2} = \int \mathcal{D}_q\xi \log\left\{\exp\left[-\beta(\kappa \pm b \mp \xi - r/2) + \frac{1}{2}\beta^2 w\right] H\left(-\frac{\kappa \pm b \mp \xi - r/2 - \beta w}{\sqrt{w}}\right) + H\left(\frac{\kappa \pm b \mp \xi - r/2}{\sqrt{w}}\right)\right\}. \tag{37}
$$

We will assume in the following that the space of equilibrium configurations of the model shrinks to a single point as $\beta \to \infty$. As the variance of this space is given, in the replica symmetric framework, by the order parameter $w = q_0 - q$ (see Sec. B), for $\beta \to \infty$ we take the following scaling Ansatz:

$$w \to \frac{x}{\beta}, \qquad q \to q, \qquad r \to r,$$
$$\hat{w} \to \beta\hat{x}, \qquad \hat{q} \to \beta^2\hat{q}, \qquad \hat{r} \to \beta\hat{r}, \qquad \hat{k} \to \beta\hat{k}, \tag{38}$$

where we re-defined all the order parameters to their infinite-β value except w and $\hat{w}$, for which we introduced the quantities $x, \hat{x}$ to avoid ambiguities. The scaling of all the others variables is fixed once w scales as $1/\beta$ and we search for finite values in the limit.

We can push forward the analysis using the asymptotic expansion of the complementary error function, which gives for β large

$$H(\sqrt{\beta}a) \sim \theta(-a) + \frac{e^{-\frac{\beta a^2}{2}}}{\sqrt{2\pi\beta}a}\theta(a), \tag{39}$$

so that the logarithm in Eq. (37) can be expressed as

$$\log\{\cdots\} \sim \begin{cases} -\beta\left(\kappa \pm b \mp \xi - r/2 - x/2\right) & \text{if} \quad \mp\xi > x + r/2 - \kappa \mp b, \\ -\beta(\kappa \pm b - r/2 \mp \xi)^2/(2x) & \text{if} \quad \mp\xi < x + r/2 - \kappa \mp b \quad \wedge \quad \mp\xi > r/2 - \kappa \mp b, \\ 0 & \text{if} \quad \mp\xi < r/2 - \kappa \mp b. \end{cases} \tag{40}$$

As a result, we obtain Eq. (11) for the functions $G_\pm$.

A.4. Saddle-point equations for large L

For large L, the integral in Eq. (24) can be evaluated via the saddle-point method. The stationary points of the functions G_S, $G_\pm$ with respect to the parameters can be found by solving the following equations. We distinguish the two cases SAT/UNSAT we reported above.

SAT phase. We obtain stationary equations for the exponent in Eq. (24) by deriving with respect to all the order parameters and their conjugate variables the function at exponent in Eq. (30) and the functions in Eq. (35). Deriving with respect to the hat variables,

$$Q = \hat{k}\left[\frac{1}{L}\operatorname{tr}(A^{-2}) - \frac{1}{\hat{k}^2}\right] + \frac{\hat{w}}{L}\operatorname{tr}(A^{-2}C) - \frac{\hat{q}}{L}\operatorname{tr}(A^{-2}C) + \frac{\hat{r}^2}{L}\boldsymbol{\delta}^\top A^{-2}\boldsymbol{\delta},$$
$$q = -\frac{\hat{q}}{L}\operatorname{tr}(A^{-2}C^2) + \frac{\hat{r}^2}{L}\boldsymbol{\delta}^\top A^{-2}C\boldsymbol{\delta},$$
$$w = \frac{\hat{k}}{L}\operatorname{tr}(A^{-2}C) + \frac{\hat{w}}{L}\operatorname{tr}(A^{-2}C^2), \tag{41}$$
$$r = -\frac{\hat{r}\hat{k}}{L}\boldsymbol{\delta}^\top A^{-2}\boldsymbol{\delta} - \frac{\hat{r}\hat{w}}{L}\boldsymbol{\delta}^\top A^{-2}C\boldsymbol{\delta},$$

where we used the useful formulas

$$A^{-1} = \hat{k}A^{-2} + \hat{w}A^{-2}C, \qquad A^2 = \hat{k}^2\mathbb{I}_Q \otimes \mathbb{I}_L + 2\hat{k}\hat{w}C + \hat{w}^2C^2. \tag{42}$$

The equations obtained deriving with respect to the other set of parameters are

$$\hat{w} = -\alpha_+\partial_q(\beta G_+) - \alpha_-\partial_q(\beta G_-),$$
$$\hat{w} + \hat{q} = -\alpha_+\partial_w(\beta G_+) - \alpha_-\partial_w(\beta G_-),$$
$$\hat{r} = [-\alpha_+\partial_r(\beta G_+) - \alpha_-\partial_r(\beta G_-)]/2, \tag{43}$$
$$0 = \alpha_+\partial_b(\beta G_+) + \alpha_-\partial_b(\beta G_-),$$

with $\beta G_\pm$ given by (35).

UNSAT phase. Using the scalings (38), we can redefine the matrix $A \to A/\beta$, so that now

$$A = \hat{k}\mathbb{I}_Q \otimes \mathbb{I}_L + \hat{x}C.$$

(44)

At leading order in β, the stationary equations for the hat variables in the scaling regime become

$$Q = -\frac{\hat{q}}{L}\operatorname{tr}(A^{-2}C) + \frac{\hat{r}^2}{L}\boldsymbol{\delta}^\top A^{-2}\boldsymbol{\delta},$$

$$q = -\frac{\hat{q}}{L}\operatorname{tr}(A^{-2}C^2) + \frac{\hat{r}^2}{L}\boldsymbol{\delta}^\top A^{-2}C\boldsymbol{\delta},$$

$$x = \frac{\hat{k}}{L}\operatorname{tr}(A^{-2}C) + \frac{\hat{x}}{L}\operatorname{tr}(A^{-2}C^2),$$

$$r = -\frac{\hat{r}\hat{k}}{L}\boldsymbol{\delta}^\top A^{-2}\boldsymbol{\delta} - \frac{\hat{r}\hat{x}}{L}\boldsymbol{\delta}^\top A^{-2}C\boldsymbol{\delta},$$

(45)

while the other set of equations is

$$\hat{x} = -\alpha_+\partial_q G_+ - \alpha_-\partial_q G_-,$$
$$\hat{q} = -\alpha_+\partial_x G_+ - \alpha_-\partial_x G_-,$$
$$\hat{r} = [-\alpha_+\partial_r G_+ - \alpha_-\partial_r G_-]/2,$$
$$0 = \alpha_+\partial_b G_+ + \alpha_-\partial_b G_-,$$

(46)

with $G_\pm$ given by Eq. (11). These last equations admit a semi-analytical expression for the derivatives:

$$\partial_x G_\pm = -\frac{G_\pm}{x} + \frac{x - \mathcal{K}_\pm}{x}\operatorname{erfc}\left(\frac{x - \mathcal{K}_\pm}{\sqrt{2q}}\right) - \frac{2}{x}\sqrt{\frac{q}{2\pi}}e^{-\frac{(\mathcal{K}_\pm - x)^2}{2q}},$$

$$\partial_q G_\pm = -\frac{1}{2x}\left[\operatorname{erfc}\left(-\frac{\mathcal{K}_\pm}{\sqrt{2q}}\right) - \operatorname{erfc}\left(\frac{x - \mathcal{K}_\pm}{\sqrt{2q}}\right)\right],$$

$$\partial_r G_\pm = \frac{\mathcal{K}_\pm}{2x}\operatorname{erfc}\left(-\frac{\mathcal{K}_\pm}{\sqrt{2q}}\right) + \frac{x - \mathcal{K}_\pm}{2x}\operatorname{erfc}\left(\frac{x - \mathcal{K}_\pm}{\sqrt{2q}}\right) + \frac{1}{x}\sqrt{\frac{q}{2\pi}}\left(e^{-\frac{\mathcal{K}_\pm^2}{2q}} - e^{-\frac{(\mathcal{K}_\pm - x)^2}{2q}}\right),$$

$$\partial_b G_\pm = \mp\frac{\mathcal{K}_\pm}{x}\operatorname{erfc}\left(-\frac{\mathcal{K}_\pm}{\sqrt{2q}}\right) \mp \frac{x - \mathcal{K}_\pm}{x}\operatorname{erfc}\left(\frac{x - \mathcal{K}_\pm}{\sqrt{2q}}\right) \mp \frac{2}{x}\sqrt{\frac{q}{2\pi}}\left(e^{-\frac{\mathcal{K}_\pm^2}{2q}} - e^{-\frac{(\mathcal{K}_\pm - x)^2}{2q}}\right).$$

(47)

SAT/UNSAT surface. The equations above can be used to draw the phase diagram of the model with respect to the external control parameters $\alpha^\pm$, κ (and, possibly, to the vectors $\mathbf{M}$, $\boldsymbol{\delta}$). In this space, the SAT/UNSAT transition surface expressing the critical value of one of these control parameters as a function of the others can be plotted requiring that $w \to 0$ (approaching the transition from the SAT phase) or $x \to \infty$ (approaching the transition from the UNSAT phase).

A.5. Ising case ($Q = 2$)

The case $Q = 2$ (Ising configurations, with the index $t = \pm 1$) admits a much simpler solution that dates back to Gardner (1988), Monasson (1992). In this case, the probability distribution of the data can be parametrized as

$$M_i(\pm 1) = \frac{1 \pm m_i}{2}, \qquad \delta_i(\pm 1) = \pm\frac{\delta_i}{2},$$

$$C_{ij}(+1, +1) = C_{ij}(-1, -1) = -C_{ij}(+1, -1) = -C_{ij}(-1, +1) = \frac{\Gamma_{ij}}{4},$$

(48)

where now m_i, δ_i, Γ_{ij} are scalars. The saddle-point equations for the hat variables become

$$1 = -\frac{\hat{q}}{L}\operatorname{tr}(A^{-2}\Gamma) + \frac{\hat{r}^2}{L}\boldsymbol{\delta}^\top A^{-2}\boldsymbol{\delta},$$

$$q = -\frac{\hat{q}}{L}\operatorname{tr}(A^{-2}\Gamma^2) + \frac{\hat{r}^2}{L}\boldsymbol{\delta}^\top A^{-2}\Gamma\boldsymbol{\delta},$$

$$x = \frac{\hat{k}}{L}\operatorname{tr}(A^{-2}\Gamma) + \frac{\hat{x}}{L}\operatorname{tr}(A^{-2}\Gamma^2),$$

$$r = -\frac{\hat{r}\hat{k}}{L}\boldsymbol{\delta}^\top A^{-2}\boldsymbol{\delta} - \frac{\hat{r}\hat{x}}{L}\boldsymbol{\delta}^\top A^{-2}\Gamma\boldsymbol{\delta},$$

(49)

Table 3. Additional common performance metrics that can be computed from Table 2

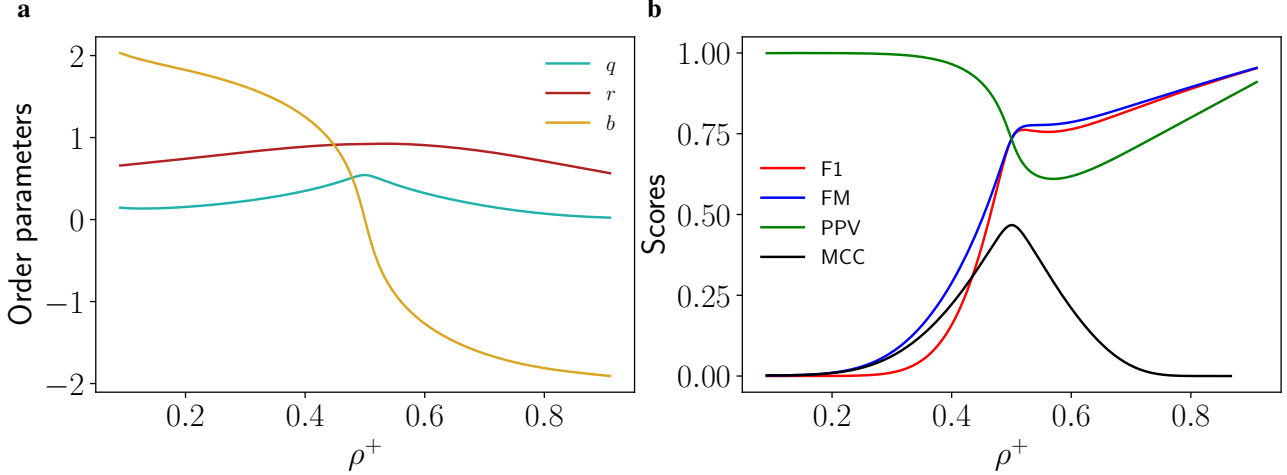|		
NPV	$\varphi^{-}\mathrm{TNR}(0)/(\varphi^{-}\mathrm{TNR}(0) + \varphi^{+}\mathrm{FNR}(0))$	Negative predicted value
FDR	$1 - \mathrm{PPV}$	False discovery rate
FOR	$1 - \mathrm{NPV}$	False omission rate
MCC	$\sqrt{\mathrm{TPR}(0) \cdot \mathrm{TNR}(0) \cdot \mathrm{PPV} \cdot \mathrm{NPV}} - \sqrt{\mathrm{FNR}(0) \cdot \mathrm{FPR}(0) \cdot \mathrm{FOR} \cdot \mathrm{FDR}}$	Matthews correlation coefficient
F_1	$2 \cdot \mathrm{PPV} \cdot \mathrm{TPR}(0)/(\mathrm{PPV} + \mathrm{TPR}(0))$	F_1 score
FM	$2 \cdot \sqrt{\mathrm{PPV} \cdot \mathrm{TPR}(0)}$	Fowlkes–Mallows index

Figure 5. **a)** Behaviors of the order parameters q, r, b for the data statistics in Figure 2a. The bias b becomes steep around $\rho^{+} \sim 0.5$ and this effect is more evident the larger the value of κ is. **b)** Additional metric behaviors defined in this work (see Tables 2 and 3), for synthetic data as in Figure 2a.

where now

$$A = \hat{k}\mathbb{I}_L + \hat{x}\Gamma, \tag{50}$$

The other set of equations remains unchanged and the case of the SAT phase can be obtained as before.

A.6. SVM

The case of the soft-margin SVM can be cast in its standard form (see Sec. G.2) by starting from the loss defined in Eq. (3), and scaling $\mathbf{J} \to \kappa\mathbf{J}$, $b \to \kappa b$, $\lambda \to \tilde{\lambda}/\kappa$, obtaining the optimization problem (we now move the L2 regularization from the measure (16) to the training energy, as this is the standard formulation for SVMs)

$$(\mathbf{J}^{\star}, b^{\star}) = \underset{\mathbf{J}\in\mathcal{S}, b\in\mathbb{R}}{\arg\min}\left\{\sum_{\mu=1}^{P}\max\left[0, 1 - \Delta_{\mu}^{+}(\mathbf{J}, b)\right] + \sum_{\nu=1}^{N}\max\left[0, 1 + \Delta_{\nu}^{-}(\mathbf{J}, b)\right] + \tilde{\lambda}\frac{\|\mathbf{J}\|_2^2}{2}\right\}. \tag{51}$$

From this, we can define the associated finite-temperature statistical mechanics model from the SVM measure in Eq. (16) and proceed as before. The calculation is the same, provided that we take $\kappa \to 1$, we substitute $\hat{k}$ from (26) with $\beta\tilde{\lambda}$, and we do not integrate over it (being now fixed as an external hyperparameter). The scaling of $\tilde{\lambda}$ with β mirrors the fact that, at finite regularization, the SVM is finding a single solution of the optimization problem (51) even in the linearly separable case, the *max-margin solution*, so that the scalings (38) should be enforced in the whole phase diagram.

B. Details on the theoretical derivation of training and generalization metrics

The order parameters introduced in Sec. 2 and obtained from the replica method explained above have a clear interpretation in terms of the low-order statistics – with respect to thermal fluctuations and quenched disorder – of the collective variables

Δ^y ($y \in \{-1, +1\}$) after training. We recall here the definition of Δ^y for convenience

$$\Delta^y(\mathbf{J}, b) = \sum_{i,t} \frac{J_i(t) s_i^y(t)}{\sqrt{L}} - b, \tag{52}$$

where $(\mathbf{J}, b)$ is an equilibrium configuration of the canonical ensemble described by the partition function (7) and $\mathbf{s}^y$ is an input point from the probability distributions described by Eq. (6). We need to distinguish the two cases of $\mathbf{s}^y$ being part of the training set or $\mathbf{s}^y$ being part of the test set (*i.e.*, independent from – but identically distributed as – the training set), denoting Δ_{train}^y and Δ_{test}^y the corresponding collective variables. It can be shown that their distribution for L large is given by

$$p(\Delta_{\text{train}}^y | \xi) = \frac{e^{-\beta \ell(y, \Delta_{\text{train}}^y)} \mathcal{N}(\Delta_{\text{train}}^y | \xi - b + yr/2, q_0 - q)}{\int d\Delta \, e^{-\beta \ell(y, \Delta)} \mathcal{N}(\Delta | \xi - b + yr/2, q_0 - q)}, \tag{53}$$

$$p(\Delta_{\text{test}}^y | \xi) = \mathcal{N}(\Delta_{\text{test}}^y | \xi - b + yr/2, q_0 - q), \tag{54}$$

where ξ is an auxiliary normal random variable, in both cases and for both classes distributed as

$$p(\xi) = \mathcal{N}(\xi | 0, q). \tag{55}$$

Note that in the test case the marginal of Δ can be obtained explicitly, as

$$p(\Delta_{\text{test}}^y) = \mathcal{N}\left(\Delta_{\text{test}}^y \left| \frac{yr}{2} - b, q_0 \right.\right), \tag{56}$$

which reduces to Eq. (9) for large β and in the UNSAT phase where $q_0 = q + x/\beta$. The Gaussian nature of Δ_{test}^y is due to the fact that the weights learned by the classifier and the points in the test set are independent, and from central limit theorems. Eq. (53) and (54) are classical results from the statistical mechanics of disordered systems, and can be obtained by mapping the replica approach we devised so far to the cavity method from Mezard (1989). For a more recent derivation, see Agoritsas et al. (2018); for a way to obtain the train distribution directly from the replica approach, see Kepler & Abbott (1988), extended in a more general non-convex setting in Franz et al. (2017).

Through these probability distributions, we obtain theoretical predictions for training and generalization metrics; in particular, train error and mis-classification error (corresponding to $1 - $ ACC) follow from

$$\mathcal{E}_{\text{train.}}(\rho^+, \rho^-) = \sum_{y \in \{\pm\}} \rho^y \langle \theta(-y\Delta_{\text{train}}^y) \rangle_{\Delta_{\text{train}}^y}, \qquad \mathcal{E}_{\text{gen.}}(\varphi^+, \varphi^-) = \sum_{y \in \{\pm\}} \varphi^y \langle \theta(-y\Delta_{\text{test}}^y) \rangle_{\Delta_{\text{test}}^y}, \tag{57}$$

while training and generalization losses are given by

$$\mathcal{L}_{\text{train.}}(\rho^+, \rho^-) = \sum_{y \in \{\pm\}} \rho^y \langle \ell(y, \Delta_{\text{train}}^y) \rangle_{\Delta_{\text{train}}^y}, \qquad \mathcal{L}_{\text{gen.}}(\varphi^+, \varphi^-) = \sum_{y \in \{\pm\}} \varphi^y \langle \ell(y, \Delta_{\text{test}}^y) \rangle_{\Delta_{\text{test}}^y}. \tag{58}$$

Alternatively, the training loss can be obtained in a more simple way, as the $\beta \to \infty$ limit of the free-energy per-sample of the physical model, so that from Eq. (10) we directly obtain

$$\mathcal{L}_{\text{train.}}(\rho^+, \rho^-) = \frac{1}{2(\alpha^+ + \alpha^-)} (G_S + \alpha^+ G_+ + \alpha^- G_-), \tag{59}$$

where the functions of the order parameters are evaluated on their saddle-point values. Other common performance metrics that can be computed in this way can be found in Tables 2 and 3, from Proposition 2.1.

B.1. Convergence rate of the mis-classification error $\mathcal{E}_{\text{gen.}}$

Here we fully derive the analytical expression for the balanced mis-classification error as a function of the composition of the training set, in the limit where the training set size is large. To obtain the expression in Eq. (13), we work in the Ising case ($Q = 2$) using the notation of section Sec. A.5, setting $m_i = m$, $\Gamma_{ij} = \delta_{ij}(1 - m^2)$. In this scenario the saddle point

equations are simpler and can be carried out without any numerical support. In particular, for the hat variables the system (49) becomes

$$1 = \frac{\hat{r}^2 u^2 - \hat{q}(1 - m^2)}{[\hat{k} + \hat{x}(1 - m^2)]^2}, \qquad x = \frac{1 - m^2}{\hat{k} + \hat{x}(1 - m^2)},$$

$$\frac{q}{1 - m^2} = \frac{\hat{r}^2 u^2 - \hat{q}(1 - m^2)}{[\hat{k} + \hat{x}(1 - m^2)]^2}, \qquad r = -\frac{\hat{r} u^2}{\hat{k} + \hat{x}(1 - m^2)},$$

(60)

where we defined the quantity $u = \|\boldsymbol{\delta}\|_2/\sqrt{L}$. The system is overdetermined due to the choice of a position-independent vector $m_i = m$. Indeed, the constraint enforced by the definition of q_0 is equivalent to the spherical constraint on the weights J_i, hence here $q_0^\star = (1 - m^2)$; consequently, the conjugate variable $\hat{q}_0$ plays no role and we can set $\hat{q}_0 = 0$, from which $\hat{x} = -\hat{q}$ follows. The system of equation (60) gives

$$q^\star = 1 - m^2, \qquad \hat{k} = \frac{q^\star}{x} + \hat{q}q^\star,$$

$$\hat{q} = \frac{q^\star}{x^2}\left(\frac{r^2}{u^2} - 1\right), \qquad \hat{r} = -\frac{rq^\star}{xu^2}.$$

(61)

Replacing the hat variables in the entropic contribution, we have to perform the extremization over (x, r, b) of the training loss

$$\mathcal{L}_{\text{train.}}(\alpha^+, \alpha^-) = \frac{1}{2(\alpha^+ + \alpha^-)} \operatorname*{extr}_{x,r,b}\left[\frac{q^\star}{xu^2}\left(u^2 - r^2\right) + \alpha^+ G_+ + \alpha^- G_-\right],$$

(62)

where $G_\pm$ are given by Eq. (11) with $q = q^\star$. In addition, we fix α^+ and minimize the training loss over the number of negative samples α^-, to obtain the optimal imbalance ratio in terms of training performances. Now we first derive Eq. (62) w.r.t. to x obtaining

$$-\frac{1}{x^2}\left\{\frac{q^\star(u^2 - r^2)}{u^2} + \sum_{y=\pm}\alpha^y\sqrt{\frac{q}{2\pi}}\left[(\mathcal{K}_y - x)e^{-\frac{1}{2q}(x - \mathcal{K}_y)^2} - \mathcal{K}_y e^{-\frac{1}{2q}\mathcal{K}_y^2}\right]\right.$$

$$\left. - \sum_{y=\pm}\alpha^y\left[(\mathcal{K}_y^2 - x^2 + q)\operatorname{erfc}\left(\frac{x - \mathcal{K}_y}{\sqrt{2q}}\right) - (\mathcal{K}_y^2 + q)\operatorname{erfc}\left(-\frac{\mathcal{K}_y}{\sqrt{2q}}\right)\right]\right\} = 0.$$

(63)

We discard the solution $x \to \infty$ as it is only valid when approaching the SAT/UNSAT transition, and we call $x = x^\star(r, b, \alpha^+, \alpha^-)$ the other solution. The stationary conditions of $\mathcal{L}_{\text{train.}}$ in Eq. (62) using Eq. (63) w.r.t. (α^-, r, b) are

$$\alpha^- : \qquad \frac{\alpha^+}{2(\alpha^+ + \alpha^-)^2}\left[(x - \mathcal{K}_-)\operatorname{erfc}\left(\frac{x - \mathcal{K}_-}{\sqrt{2q}}\right) - (x - \mathcal{K}_+)\operatorname{erfc}\left(\frac{x - \mathcal{K}_+}{\sqrt{2q}}\right)\right] = 0,$$

$$r : \qquad \left[\frac{\partial x^\star(r, b)}{\partial r} - \frac{1}{2}\right]\sum_{y=\pm}\alpha^y\left[\operatorname{erfc}\left(\frac{x - \mathcal{K}_y}{\sqrt{2q}}\right) - \frac{2(x - \mathcal{K}_y)}{\sqrt{2q\pi}}\exp\left\{-\frac{(x - \mathcal{K}_y)^2}{2q}\right\}\right] = 0, \qquad (64)$$

$$b : \qquad \sum_{y=\pm}\alpha^y\left(\frac{\partial x^\star(r, b)}{\partial r} - y\right)\left[\operatorname{erfc}\left(\frac{x - \mathcal{K}_y}{\sqrt{2q}}\right) - \frac{2(x - \mathcal{K}_y)}{\sqrt{2q\pi}}\exp\left\{-\frac{(x - \mathcal{K}_y)^2}{2q}\right\}\right] = 0.$$

The first equation gives the condition $\mathcal{K}_- = \mathcal{K}_+$, yielding $b^\star = 0$. In the large limit $\kappa \gg 1$, where we have shown to achieve the absolute best generalization performance, we get $\mathcal{K}_\pm \to \infty$. Hence Eq. (63) admits the explicit solution

$$x^\star = \frac{1}{u}\sqrt{\frac{q^\star(u^2 - (r^\star)^2)}{\alpha^+ + \alpha^-}},$$

(65)

with

$$r^\star = u\sqrt{\frac{(\alpha^+ + \alpha^-)u^2}{4q^\star + u^2(\alpha^+ + \alpha^-)}}, \qquad \alpha^- = \alpha^+.$$

(66)

We conclude that optimal classification over the training set is reached at balanced number of samples, for high margin and infinite dataset size. For large $\alpha^{\pm}$, we finally derive the convergence rate of the mis-classification error in the balanced case ($\varphi^{\pm} = 0.5$), as

$$\mathcal{E}_{\text{gen.}} \underset{\alpha^{\pm} \gg 1}{\sim} \frac{1}{2} \operatorname{erfc}\left(\frac{\|\boldsymbol{\delta}\|_2}{2\sqrt{2L(1-m^2)}}\right)\delta(\alpha^+ - \alpha^-) + \frac{\mathcal{C}}{\alpha^+ + \alpha^-}\delta(\alpha^+ - \alpha^-) + O((\alpha^{\pm})^{-2}), \tag{67}$$

where

$$\mathcal{C} = \sqrt{1-m^2}\exp\left\{-\frac{\|\boldsymbol{\delta}\|_2^2}{8L(1-m^2)}\right\}. \tag{68}$$

C. Details on the theory of oversampling strategy

Here we provide details about the main steps to adapt the analytical computation of Sec. A when the minority class is oversampled c times. We assume, as in most of the paper, that the minority class is the positive one. The mathematical formulation in this setting follows immediately from Eq. (17) by introducing the oversampling factor c, accounting for the multiplicity of each positive sample:

$$\exp\left\{-\beta\sum_{\mu=1}^{\alpha^+L}(\kappa - \Delta_\mu^a)\theta(\kappa - \Delta_\mu^a)\right\} \longrightarrow \exp\left\{-\beta c\sum_{\mu=1}^{\alpha^+L}(\kappa - \Delta_\mu^a)\theta(\kappa - \Delta_\mu^a)\right\}. \tag{69}$$

In the phase where the dataset is linearly separable, oversampling the positive class is pointless, as the perceptron already correctly classifies all the training data and adding equivalent ones does not yield any benefit. On the other hand, in the UNSAT phase the presence of the factor c becomes relevant and affects the energetic contribution from the positive class, see Eq. (11); in practice, the quantities β, x are multiplied by the degree of oversampling c. The positive function G_+ becomes

$$G_+ = -\sqrt{\frac{q}{2\pi}}\left(\frac{cx - \mathcal{K}_+}{cx}e^{-\frac{(cx-\mathcal{K}_+)^2}{2q}} + \frac{\mathcal{K}_+}{cx}e^{-\frac{\mathcal{K}_+^2}{2q}}\right) + \frac{(cx - \mathcal{K}_+)^2 + q}{cx}H\left(\frac{cx - \mathcal{K}_+}{\sqrt{q}}\right) - \frac{\mathcal{K}_+^2 + q}{cx}H\left(\frac{-\mathcal{K}_+}{\sqrt{q}}\right) \tag{70}$$

and all its derivatives in Eq. (47) are changed accordingly. The procedure to solve the saddle-point equations remains analogous.

D. Oversampling vs undersampling: the choice depends on the performance metric

In Section 3.4 we discussed the advantage of restoring balance in the dataset, oversampling or undersampling, evaluating the generalization performances in terms of the BA metrics. However, same data distributions and sizes can lead to different optimal strategies depending on how performances are evaluated, *i.e.* on the specific performance metric adopted. We discuss this phenomenon within our theoretical framework, comparing the mis-classification error $\mathcal{E}_{\text{gen.}}$ and a distance-based metric, namely the generalization loss $\mathcal{L}_{\text{gen.}}$ defined in Eq. (58). Both metrics are such that the lower the better, as they measure wrongly classified data points.

The so-defined generalization loss gives the average distance of the mis-classified examples from the decision boundary set by the classifier. In practice, low $\mathcal{L}_{\text{gen.}}$ and high $\mathcal{E}_{\text{gen.}}$ values correspond to many mis-classified examples close to the decision boundary, while the opposite refers to situations where there are few mis-classified examples far away from the decision boundary. Using the same notation as in Section 3.4 with $c \in [1, \alpha^-/\alpha^+]$, we compare the metrics $\mathcal{E}_{\text{gen.}}$ and $\mathcal{L}_{\text{gen.}}$ under two different protocols, defined as follows

(IP) *Imbalanced Protocol*: we leave the majority class as it is and augment the minority class size, thus performing training with $c\alpha^+$ positives and α^- negatives;

(BP) *Balanced Protocol*: we undersample the negatives down to the same size of positives, hence dealing with a balanced training set of total size $2c\alpha^+$. Note that this is the protocol adopted elsewhere in the work and it is the only one restoring balance between classes.

We show results in Figure 6: in the same conditions, the optimal strategy minimizing the generalization loss is the BP protocol with $c = 1$ (*i.e.*, random undersampling of the majority class), while the optimal one for the mis-classification error is the BP protocol with $c = \alpha^-/\alpha^+$ (*i.e.*, random oversampling of the minority class).
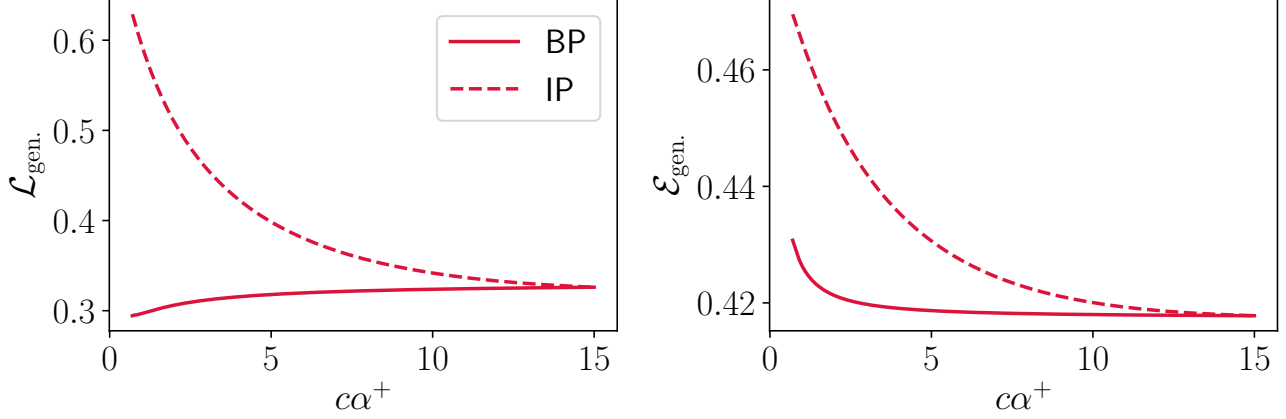
Figure 6. Comparison between the mis-classification error and generalization loss as a function of the actual minority training size (including oversampling) $c\alpha^+$ for protocols IP and BP. Here $L = 100$, $Q = 2$, $\kappa = 0.4$, $\alpha^- = 15$, $\alpha^+ = 0.5$ with $m_i = 0.4$ over each position, δ_i are i.i.d. as $\mathcal{N}(0.625, 0.75)$ and covariances are diagonal. Note that for these metrics the lower the better.

E. Lattice Proteins

In many real-world applications tokenization and categorical variables are routinely encountered in machine learning. For example, biological applications are one of such cases, where Potts indices run over the amino acid or nucleotides alphabet, *i.e.* $Q = 20$ or $Q = 4$ respectively. In this section we report an analysis carried out on synthetic protein data – namely Lattice Proteins – where we aim to distinguish between two Lattice Proteins in a bounded or unbounded state. In this analysis we note that our theory quantitatively predicts numerical results even if we ignore the covariance matrix, *i.e.* $C_{ij}(t, u) = [M_i^2(t) - M_i(t)M_i(u)]\delta_{ij}$; in other terms, the first-order statistics of the data is sufficient to reproduce numerical results, despite the data contain non-trivial correlations.

Lattice Proteins (LPs) models consist of artificial protein sequences used to reproduce relevant features of real proteins and investigate their folding properties (Li et al., 1996; Mirny & Shakhnovich, 2001). A LP is defined as a self-avoiding path of over a $3 \times 3 \times 3$ lattice cube. There are 103,406 possible configurations (structures) over the lattice cube, excluding global symmetries (Shakhnovich & Gutin, 1990). A sequence $\mathbf{v}$ is given by the chain of $L = 27$ amino acids sitting on each site of the lattice cube. A structure S is defined by its contact matrix $\mathbf{c}^S$ such that - given two sites i,j on the lattice cube

$$c_{ij}^S = \begin{cases} 1 & i, j \quad \text{in contact,} \\ 0 & \text{otherwise.} \end{cases} \tag{71}$$

The probability that sequence $\mathbf{v}$ folds into structure S is

$$p_{\text{nat}}(S|\mathbf{v}) = \frac{e^{-E(\mathbf{v}|S)}}{\sum_{S'=1}^{R} e^{-E(\mathbf{v}|S')}}, \tag{72}$$

where R is a representative subset of all possible structures ($R = 10000$ in this case). The energy of the sequence in a structure $E(\mathbf{v}|S)$ is given by

$$E(\mathbf{v}|S) = \sum_{i<j} c_{ij}^S E_{MJ}(v_i, v_j), \tag{73}$$

where residues in contact interact via the Miyazawa-Jernigan matrix E_{MJ}, a proxy containing effective interaction energies for each amino acid pair. Eventually we let two sequences $\mathbf{v}_1$, $\mathbf{v}_2$ – folded in structure S_1, S_2 respectively – interact to form a unique compound via the interaction energy

$$I(\mathbf{v}_1, \mathbf{v}_2|S_1 + S_2, \pi) = \sum_{i<j} c_{ij}^{(S_1+S_2, \pi)} E_{MJ}(v_i, v_j), \tag{74}$$
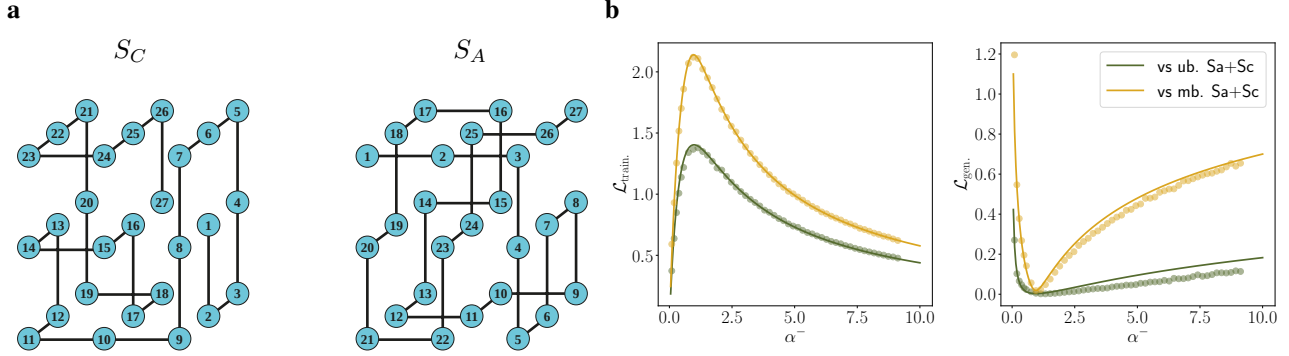
a

b

Figure 7. **a)** Graphical representation of the compound $S_A + S_C$ made of two Lattice Proteins. The solid black lines highlight the backbone structures over the lattice cubes. **b)** Theoretical predictions (solid lines) and numerical simulations (dots) on LP data of training and generalization losses as defined in Eq. (58) (left and right panel, respectively), for two different distributions of negative examples. In practice, we use data $S_A + S_C$ in the bounded state as positives; negatives are data $S_A + S_C$ in the unbounded or mildly bounded state (green and gold, respectively). Here $\alpha^+ = 19$, $\kappa = 5$, $Q = 20$, $L = 54$, with $\mathbf{M}, \boldsymbol{\delta}$ estimated from the data. Numerical simulations are averaged over 20 trials.

where now the contact matrix $\mathbf{c}^{(S_1+S_2)}$ is defined over both structures and the sum runs over all sites of both lattice cubes. The index π in Eq. (74) labels a specific orientation of the interaction. In analogy with Equation (72), the probability that sequences $\mathbf{v}_1, \mathbf{v}_2$ folded into the compound $S_1 + S_2$ interact is

$$p_{\text{int}}(\pi = 0, S_1 + S_2 | \mathbf{v}_1, \mathbf{v}_2) = \left[\frac{e^{-I(\mathbf{v}_1, \mathbf{v}_2 | S_1+S_2, \pi=0)}}{\sum_{\pi'=1}^{R'} e^{-I(\mathbf{v}_1, \mathbf{v}_2 | S_1+S_2, \pi')}} \right]^{\chi}, \tag{75}$$

where R' is the total number of orientations ($R' = 144$ in this case) and we added the exponent χ to tune the interaction strength. Depending on the value of χ, we refer to the compound $S_1 + S_2$ as bounded ($\chi = 5$), mildly bounded ($\chi = 0.1$) or unbounded ($\chi = 0$).

Given two structures S_1, S_2, we collect sequences through MCMC dynamics such that $p_{\text{nat}}(S_1 | \mathbf{v}_1)$, $p_{\text{nat}}(S_2 | \mathbf{v}_2) > 0.99$. Here we randomly choose two specific compounds, namely $S_A + S_C$ and $S_B + S_C$ (see Figure 7a to visualize one of the two compounds). We group lists of sequences to form the dataset used in this work as follows:

- sequences $\mathbf{v}_A, \mathbf{v}_C$ in the bounded state $S_A + S_C$ represent positive data;

- sequences $\mathbf{v}_A, \mathbf{v}_C$ in the unbounded state $S_A + S_C$ represent negative data;

- sequences $\mathbf{v}_B, \mathbf{v}_C$ in the bounded state $S_B + S_C$ or sequences $\mathbf{v}_A, \mathbf{v}_C$ in the mildly bounded state $S_A + S_C$ represent out-of-sample data.

We stress again explicitly that this so-defined model of LPs is not a model with independent features as it involves inter- and intra-structures couplings in the energy terms (73) and (74). Yet, we find that our theoretical predictions with diagonal covariance matrix closely reproduce numerical simulations (see Fig. 7b).

F. Additional numerical results

F.1. Numerical validation of replica calculation

Here we numerically validate our predictions of the quantities $\mathcal{L}_{\text{train.}}$, $\mathcal{L}_{\text{gen.}}$, $\mathcal{E}_{\text{train.}}$ within the RS Ansatz. The numerical solution of the saddle-point equations of interest can be obtained via a simple iterative method and allows to get the theoretical predictions. The code to reproduce the theoretical curves can be found on github. We check the agreement of numerical simulation with theory over three different synthetic datasets $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3$, composed as follows. Note that here all the datasets contain diagonal covariance matrices, *i.e.* we are setting $C_{i,j}(t,u) = [M_i(t)^2 - M_i(t)M_j(u)]\delta_{ij}$.

1) The dataset $\mathcal{D}_1$ has $L = 100$, $Q = 2$, $\alpha^+ = 1$ and $\kappa = 0.4$. We sample positive and negative data using $m_i = 0.4$ and $\delta_i = 1$ over each position.
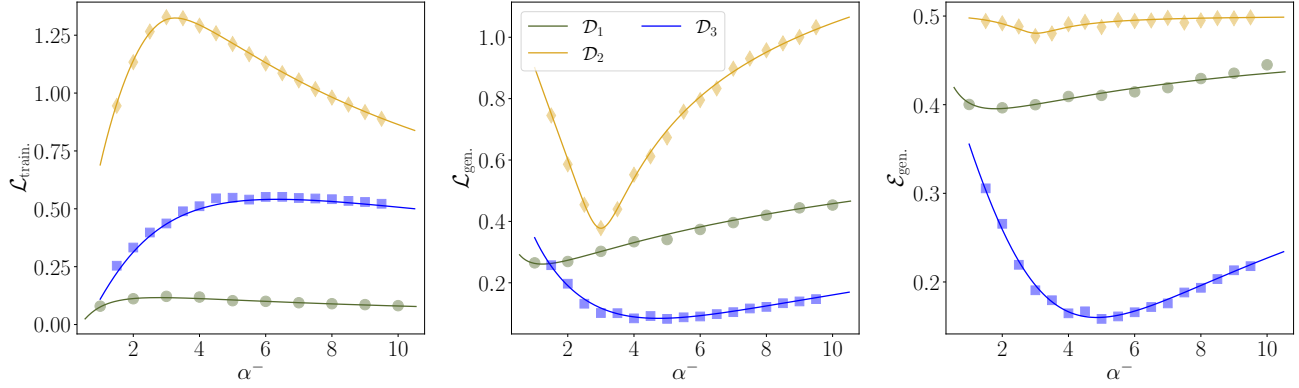
Figure 8. Theoretical curves with their numerical validation for the quantities $\mathcal{L}_{\text{train.}}$, $\mathcal{L}_{\text{gen.}}$ and $\mathcal{E}_{\text{gen.}}$ – defined in Eq. (57), (58) – in panels **a)**, **b)** and **c)**, respectively. We test our predictions within three different scenarios, labelled $\mathcal{D}_1$, $\mathcal{D}_2$, $\mathcal{D}_3$ and composed as explained in the text. All numerical experiments are averaged over 20 trials.

2) The dataset $\mathcal{D}_2$ has $L = 200$, $Q = 10$, $\alpha^+ = 3$ and $\kappa = 2$. We sample the tensor $\mathbf{M}$ as i.i.d. variables uniform in $[0.1, 0.3]$ and $\boldsymbol{\delta}$ as i.i.d. variables uniform in $[0, 0.1]$. We then enforce the normalization conditions setting the last component over each position as

$$M_i(Q) = 1 - \sum_{t=1}^{Q-1} M_i(t), \qquad \delta_i(Q) = -\sum_{t=1}^{Q-1} \delta_i(t). \tag{76}$$

3) The dataset $\mathcal{D}_3$ has $L = 200$, $Q = 10$, $\alpha^+ = 4.5$ and $\kappa = 2$. We sample the tensors

$$\mathbf{M}^{\pm} = \mathbf{M} \pm \frac{\boldsymbol{\delta}}{2\sqrt{L}} \sim \mathcal{N}(0.8 \pm 0.2, 0.1), \tag{77}$$

and rescale them so that on each position the vector sums to one (normalization condition). We then obtain $\mathbf{M}$, $\boldsymbol{\delta}$ from $\mathbf{M}^{\pm}$ by sum and difference.

We show the results in Figure 8.

G. On numerical experiments

G.1. Implementation details of RBM

We train the RBM using a gradient-based method to maximize the log-likelihood $\mathcal{L}_{\text{RBM}} = \sum_{k=1}^{P} \log p(\mathbf{v}_k)$, where the sum runs over the training (positive) data. Given a generic parameter ω to learn, we solve the momentum matching condition

$$\frac{\partial \mathcal{L}_{\text{RBM}}}{\partial \omega} = \left\langle \frac{\partial E}{\partial \omega} \right\rangle_{\text{M}} - \left\langle \frac{\partial E}{\partial \omega} \right\rangle_{\text{D}}, \tag{78}$$

where the r.h.s. expectation values are over the model and data distribution, respectively. The parameters are updated with persistent-contrastive-divergence (PCD) (Tieleman, 2008) algorithm. Averages over the model are estimated running $T = 100$ parallel Markov chains, while averages over the data are computed on mini-batches of size T sampled from the training set. The RBM weights w are initialized to small random Gaussian values, with a standard deviation equal to $0.1/\sqrt{L}$, where L is the number of visible units. Eventually, we include a regularization during the training procedure by introducing a L2 penalty term $\mathcal{L}_{\text{RBM}} \to \mathcal{L}_{\text{RBM}} - \lambda \|\omega\|_2^2$. For the MNIST dataset we used the following parameters: 784 visible units, 100 hidden units, L2 strength 0.001 and 250 Epochs.

Gibbs Sampling. To generate a new sample with the RBM, we use Gibbs Sampling as follows

(i) We start from an initial configuration $\mathbf{v}_0$ (either random or taken from the training set);

(ii) We sample the next configuration $\mathbf{v}_1$ based on $p(\mathbf{v}_0|\mathbf{z})$;

(iii) We repeat step (ii) for K times and return the configuration $\mathbf{v}_K$ as a new sample.

G.2. Implementation details of supervised classifiers

Perceptron. To validate the theoretical predictions (*e.g.* Figure 2b), we minimize the following cost function

$$\mathcal{L}(\mathbf{J}, b) = \sum_{\mu} (\kappa - y_{\mu} \Delta_{\mu}) \, \theta \, (\kappa - y_{\mu} \Delta_{\mu}) - \lambda \sum_{t} [J_i(t) - 1]^2 - \lambda \sum_{t} (\|\mathbf{J}\|_2^2 - LQ)^2, \tag{79}$$

where the last two terms on the r.h.s. enforce the spherical weights and zero-sum condition. We set $\lambda = LQ$ and minimize Eq. (79) using the L-BFGS-B routine of the Scipy library (Jones et al., 2001).

Linear SVM. To carry out the numerical experiments on MNIST dataset to compare different mixing balancing strategies (*e.g.* Figure 4a) we use Linear Support Vector Machines. We use the Scikit-learn package (Pedregosa et al., 2011) which implements the SVM as follows

$$\mathcal{L}(\mathbf{J}, b) = C \sum_{\mu} \max(0, 1 - y^{\mu}(\mathbf{J} \cdot \mathbf{x}^{\mu} + b)) + \frac{1}{2}\|\mathbf{J}\|_2^2, \tag{80}$$

where the sum runs over the examples in the dataset. We set the intercept value $b = 0$ (option "fit-intercept=False") and the regularization strength $C = 10$. For each minority class size we average over 100 realization to make the performance curves smooth enough. We evaluate results using the BA metric. Notice how we can compare numerical curves obtained in this way with theoretical predictions derived from Eq. (51), with $C \to \tilde{\lambda}^{-1}$, $b \to -b$ and rescaling the input data by $\sqrt{L}$.

ResNet-50 We take a pretrained ImageNet ResNet-50 architecture and add a fresh dense layer with 512 neurons and ReLU activation; we put an output neuron with sigmoid activation for the binary classification task. We fine-tune the network weights on Cifar10 data. We use RMSprop optimizer with a learning rate of 10^{-5} and decay of 10^{-5} and train for 100 epochs with a batch-size of 128.

G.3. Dataset preprocessing

CIFAR10. CIFAR10 for ResNet-50 fine-tuning are used without any preprocessing. To perform binary classification we split images into two classes according to what they represent – [airplane, car, bird, cat, deer] and [dog, frog, horse, ship, truck], respectively. The test set is balanced and has size 1000.

MNIST. MNIST data are first converted to binary values $\{0, 1\}$ setting the threshold value for the black/white mapping at 0.5. To perform binary classification we split the digits in two classes. Parity MNIST (pMNIST) has odd and even digits classes; while 5-MNIST has smaller and larger than digit five classes. Hence, the two classes contain ~ 30000 examples each. The test set is balanced and has size 1000.

FashionMINST. FMNIST data are first converted to binary values $\{0, 1\}$ setting the threshold value for the black/white mapping at 0.5. To perform binary classification we split the dataset in two classes representing "Pullover" and "Shirt" images. Hence, each class has 6000 examples. The test set is balanced and has size 1000.

CelebA. Images of the CelebA datasets are first converted to single (grayscale) channel by taking the average of the three color channels. Then data are binarized through the Sauvola–Pietikäinen adaptive image binarization algorithm (Sauvola & Pietikäinen, 2000), that sets the threshold for the black/white mapping for each pixel (x, y) as

$$T(x, y) = m(x, y) \left[1 + k \left(\frac{s(x, y)}{R} - 1 \right) \right]. \tag{81}$$

Here m, s refer to the mean and standard deviation of the intensity across a window of size 8 around the pixel (x, y). The two constant values are set as $R = \max s(x, y)$ over all pixels in the image and $k = 0.05$. To perform binary classification we only select images based on two attributes, namely "Straight hair" or "Wavy hair". Some randomly selected faces from the two classes are shown in Figure 9.

Figure 9. Randomly selected images of faces from the CelebA dataset after the data preprocessing described in the text. Images in the top row belong to the straight hair class, while bottom row displays faces with wavy hair.