Class Imbalance in Anomaly Detection: Learning from an Exactly Solvable Model

F.S. Pezzicoli
TAU team, LISN
Université Paris-Saclay
CNRS, Inria
91405 Orsay, France
francesco.pezzicoli@
universite-paris-saclay.fr

V. Ros
LPTMS
Université Paris Saclay
CNRS
91405 Orsay, France
valentina.ros@cnrs.fr

F.P. Landes
TAU team, LISN
Université Paris-Saclay
CNRS, Inria
91405 Orsay, France
francois.landes@inria.fr

M. Baity-Jesi
SIAM Department
Eawag (ETH)
8600 Dübendorf, Switzerland
marco.baityjesi@eawag.ch

Abstract

Class imbalance (CI) is a longstanding problem in machine learning, slowing down training and reducing performances. Although empirical remedies exist, it is often unclear which ones work best and when, due to the lack of an overarching theory. We address a common case of imbalance, that of anomaly (or outlier) detection. We provide a theoretical framework to analyze, interpret and address CI. It is based on an exact solution of the teacher-student perceptron model, through replica theory. Within this framework, one can distinguish several sources of CI: either intrinsic, train or test imbalance. Our analysis reveals that, depending on the specific problem setting, one source or another might dominate. We further find that the optimal train imbalance is generally different from 50%, with a non trivial dependence on the intrinsic imbalance, the abundance of data and on the noise in the learning. Moreover, there is a crossover between a small noise training regime where results are independent of the noise level to a high noise regime where performances quickly degrade with noise. Our results challenge some of the conventional wisdom on CI and pave the way for integrated approaches to the topic.

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

1 INTRODUCTION

Supervised learning under class imbalance (CI) is a fundamental challenge in modern machine learning, as many real-world datasets often exhibit varying degrees of imbalance [Yamanishi et al., 2000, Almeida et al., 2011, Kyathanahally et al., 2021, Schür et al., 2023]. Efforts to mitigate the detrimental effects of class imbalance have led to the development of various approaches, with the machine learning community establishing widely accepted heuristics based on empirical evaluations. These approaches can be broadly categorized into three types: those acting on the data distribution [Japkowicz and Stephen, 2002, Chawla et al., 2002, Ando and Huang, 2017, Zhang and Mani, 2003], those modifying the loss function [Xie and Manski, 1989, Kini et al., 2021, Behnia et al., 2023, Thrampoulidis et al., 2022, Menon et al., 2021], and those biasing the dynamics of the training process [Anand et al., 1993, Tang et al., 2020]. However, due to the lack of a theoretical framework for the analysis of CI, it is often unclear which of those methods work best and when or why. For this reason, recent studies tried to fill this theoretical gap, either by focusing on how imbalance influences the learning dynamics [Ye et al., 2021, Francazi et al., 2023, Kunstner et al., 2024], or how it influences the loss landscape [Mignacco et al., 2020, Loffredo et al., 2024, Mannelli et al., 2023].

In these works, CI is treated as a single phenomenon, which can be addressed through a single formal approach. We highlight, instead, that one should distinguish between (at least) two types of imbalance. What we call *Multiple Groups* imbalance (MGI) involves samples drawn from distinctly different distributions, with the imbalance arising either from the sampling process (*e.g.* the toxicity of certain chemicals is tested more often than others [Schür et al., 2023]) or being intrinsic to the data itself (*e.g.* some species

being more common within an ecosystem [Kyathanahally et al., 2021]). In contrast, *Outlier or Anomaly Detection* imbalance (ADI), is generally a binary problem. All examples are drawn from the same distribution, and one needs to identify outliers based on an unknown rule (*e.g.* only some of the components of a power grid will cause a failure, but we do not know what will cause it [Zhang et al., 2019]), with the rule itself determining the imbalance of the data. In this case, the imbalance is intrinsic to the problem at hand, as anomalies are naturally fewer in number than normal samples. As we will see, differently from MGI, ADI has an associated intrinsic imbalance scale, which we call ρ_0 and which represents the fraction of anomalies. If $\rho_0 = 0.5$, anomalies and normal samples appear with the same frequency.

While most of the theoretical literature on class imbalance implicitly treats MGI, we are not aware of theoretical work targeting how ADI affects the loss landscape. This requires a different theoretical setup, yields different results, and is the aim of our study.

We study the effects of ADI on the training and test landscape in a paradigmatic analytically tractable model. Specifically, we study a modified version of the Teacher-Student (TS) spherical perceptron [Gardner and Derrida, 1989, Seung et al., 1992], where one can tune the amount of CI, and study its effect on learning. Studying a tractable model, where the ground truth is known, allows us to disentangle the various reasons why a high performance is reached or not, providing interpretable results.

Contributions. The main contributions of our work are the following :

1. By solving the Teacher-Student spherical perceptron in the presence of ADI, we provide an **interpretable framework to characterize ADI**. This allows us to elucidate the role of three sources of imbalance: intrinsic imbalance ρ_0, the train imbalance ρ_{train}, and the test set imbalance ρ_{test}. As a function of these quantities, we examine how various commonly used performance metrics vary and are able to track the quality of the learnt model.

2. We contribute to challenging the conventional wisdom and common practice that a perfectly rebalanced training set ($\rho_{\text{train}} = 0.5$) is optimal, and instead we find that **the optimal value of ρ_{train} is not 0.5**. Factors influencing this value and its relevance include the abundance of data, the amount of noise in the dynamics, and the intrinsic imbalance ρ_0.

3. **Dynamics with lower noise are less susceptible to CI**. We identify two distinct regions. For low noises, the performance is optimal, and largely unaffected by the exact level of noise. For large noises, the performance degrades and is sensitive to additional amounts of noise. This effect is related to how well the student can guess the intrinsic ρ_0, which is also affected by ρ_{train}.

Related work. Although recent analytical work covered class imbalance with approaches similar to ours, this does not explicitly distinguish between different types of imbalance, and instead implicitly focuses on MGI. These studies assume the presence of two distinct sub-populations in the data distribution and explore the effects of class imbalance from the perspective of the loss landscape, along with potential mitigation strategies. [Mignacco et al., 2020] focus on the role of regularization, showing that imbalance impedes achieving Bayes-optimal performances. [Mannelli et al., 2023] focus on fairness implications, studying how imbalance affects the performances across the sub-populations. They also introduce a mitigation strategy based on coupled neural networks trained on subsets of the full training dataset. [Loffredo et al., 2024] investigate the effect of imbalance on various accuracy metrics and in particular show that the AUC score is rather insensitive to imbalance while the Balanced Accuracy is a better suited metrics to study imbalanced problems. They focus also on the effectiveness of re-sampling techniques and prove that mixed strategies of random over-sampling/under-sampling are the most effective. Their setup explicitly covers MGI, with a learning problem where it is impossible to obtain zero loss.

2 FRAMEWORK

2.1 Type of Imbalance

We now make more precise the nuance between MGI and ADI mentioned in Sec. 1.

Multiple Groups Imbalance In MGI, there are $K \geq 2$ separate populations, with distinct feature's distribution, *i.e.* we have a mixture of densities: $\mathrm{d}\mu(\mathbf{S}) = \sum_k^K \pi_k \, \mathrm{d}\mu(\mathbf{S}|\text{class} = k)$, where π_k is the class frequency of class k and $\mathbf{S}$ represents the data. Both the distributions $\mathrm{d}\mu(\mathbf{S}|\text{class} = k)$ and the π_k can be different for different k. Since the π_ks are arbitrary, they can be selected in such a way that some classes are more present than others.

Anomaly Detection Imbalance In ADI, there is one single population ($K = 1$), and the classification

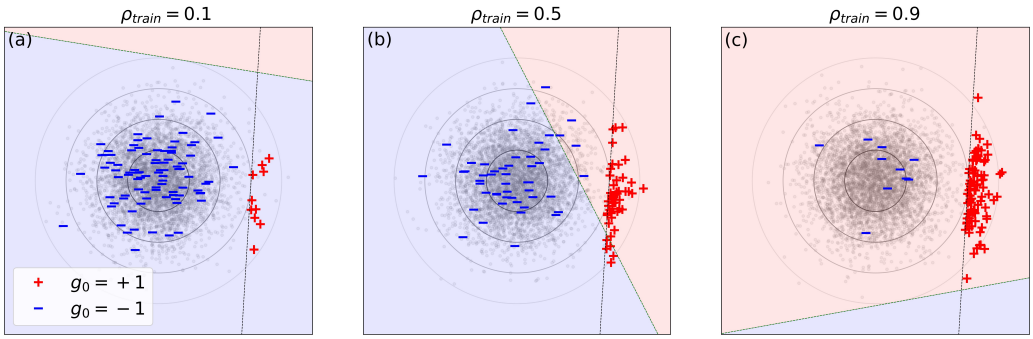

Figure 1: *Geometrical interpretation of learning an Anomaly Detection task under class imbalance,* with fixed ρ_0, and $\rho_{\text{train}} = 0.1, 0.5, 0.9$. Normal examples (negative label, $g_0^\ell = -1$) are represented with blue ($-$) symbols, anomalies (positive label, $g_0^\ell = +1$) with red ($+$). Shaded grey points depict the underlying Gaussian data distribution and grey circles locate contours at $1\sigma, 2\sigma$ and 3σ (σ is the standard deviation). The black dashed line represents the teacher hyperplane (decision boundary) which determines the ground-truth labels and the colored regions of the plane depict the classes predicted by the student (the model being trained). The three examples contain the same number of misclassified examples, but the learned model is very different. When the training set is strongly imbalanced ((a) and (c)) the student has an entropic incentive to learn a strong bias, completely discarding the alignment with the teacher: learning only a bias that matches the train imbalance is statistically favored due to the large number of possible directions for the student hyperplane that achieve a low error. Learning on a balanced training set (b) forces the student to learn the direction of the teacher because of the higher cost of entirely mis-classifying one of the two classes.

problem consists of identifying outliers in the density $\mathrm{d}\mu(\mathbf{S})$, where the definition of outlier depends on a given criterion/rule. Since this criterion defines both the classes and the frequency of the outliers, the imbalance of the problem (this is what we will call ρ_0) cannot be changed once the classes are defined. In this paper, our whole theoretical setup will be in the ADI setting.

2.2 Model

We consider the widely-studied Teacher-Student Spherical Perceptron [Gardner and Derrida, 1989, Györgyi, 1990, Seung et al., 1992, Fontanari and Meir, 1993, Nishimori, 2001]. In this set-up, diagrammatically represented in Fig. 6 in App.A, a *teacher* perceptron with a planted weight configuration assigns a label to each sample, while a *student* perceptron learns to mimic the teacher by adjusting its weights through Empirical Risk Minimization on the samples in the training set. We motivate these modeling hypotheses through empirical observations in App. G.3.1, by studying the distribution of input features in a realistic anomaly detection task, namely for the BTAD dataset [Mishra et al., 2021]. Those observations justify adopting the ADI setup (rather than MGI).

Problem Setting. Given an input sample, $\mathbf{S}^\ell \in \mathbb{R}^N$ (ℓ is the sample index and N the number of features), the Spherical Perceptron model related to the student assigns it a label, g^ℓ, through the relation

$$g^\ell = g\left(\frac{\mathbf{w}\cdot\mathbf{S}^\ell}{\sqrt{N}} + b\right), \tag{1}$$

where $\mathbf{w} = (w_1, \dots, w_N) \in \mathbb{R}^N$ are the model weights, subject to the constraint $\mathbf{w}^T \cdot \mathbf{w} = N$, $b \in \mathbb{R}$. This constraint is akin to a regularization, and ensures that each weight is of order 1. The activation function is usually chosen as $g(x) = \mathrm{sign}(x)$. Ground-truth labels g_0^ℓ are obtained through the teacher assignment:

$$g_0^\ell = g\left(\frac{\mathbf{w}^0\cdot\mathbf{S}^\ell}{\sqrt{N}} + b_0\right). \tag{2}$$

Although teacher and student have the same architecture, the teacher's parameters $\mathbf{w}^0$ and b_0 are fixed, while the student parameters $\mathbf{w}$ and b are learned through the minimization of the loss

$$\mathcal{E}(\mathbf{w}, b) = \sum_{\ell=1}^{P} \epsilon\left(g\left(\frac{\mathbf{w}\cdot\mathbf{S}^\ell}{\sqrt{N}} + b\right), g_0^\ell\right), \tag{3}$$

where P is the number of training samples. We use a square loss, $\epsilon(x, y) = \frac{1}{2}(x-y)^2$. The shape of the loss matters little, since in practice errors have cost +2 and correct predictions have cost 0.

Modeling Class Imbalance. We model the ADI by introducing an imbalance parameter ρ_{train}, that fixes the ratio between samples in the majority and minority classes in the training set.

The set of all data-points observed during training is denoted as $\{\mathbf{S}^\ell\}_{\ell=1}^P$ with $\mathbf{S}^\ell = (S_1^\ell, \dots, S_N^\ell) \in \mathbb{R}^N$ and $P = N\alpha$. We fix the ratio α, which represents the abundance of data, to be finite. This is a classical choice in this setup, which ensures that the problem is non trivial, *i.e.* it is neither overconstrained nor underconstrained. Samples are i.i.d., and for each sample $\mathbf{S}$ the components are distributed according to: $S_i \sim DS_i = \frac{\mathrm{d}S_i}{\sqrt{2\pi}} e^{-S_i^2/2}$. We will use the shorthand notation $D\mathbf{S} = \prod_{i=1}^N DS_i$ to denote the measure of probability of the single sample.

In order to have a skewed distribution of samples with a number $N\alpha\rho_{\text{train}}$ of positive (anomalous) samples ($g_0^\ell = +1$) and $N\alpha(1-\rho_{\text{train}})$ of negative (normal) samples ($g_0^\ell = -1$), we define the training set measure as:

$$\mathrm{d}\mu_{\text{train}}(\{\mathbf{S}^\ell\}) = \frac{1}{c_+^{N\alpha\rho_{\text{train}}}} \left(\prod_{\ell=1}^{\alpha N\rho_{\text{train}}} D\mathbf{S}^\ell \Theta\left(\frac{\mathbf{w}^0\cdot\mathbf{S}^\ell}{\sqrt{N}} + b_0 \right) \right) \frac{1}{c_-^{N\alpha(1-\rho_{\text{train}})}} \left(\prod_{\ell'=\alpha N\rho_{\text{train}}+1}^{\alpha N} D\mathbf{S}^{\ell'} \Theta\left(-\frac{\mathbf{w}^0\cdot\mathbf{S}^{\ell'}}{\sqrt{N}} - b_0 \right) \right). \tag{4}$$

The notation $\mathrm{d}\mu(\{\mathbf{S}^\ell\})$ is shorthand for $\mathrm{d}\mu(\mathbf{S}^1, \dots, \mathbf{S}^{N\alpha})$ and denotes the measure of probability over all the training samples. We use the Heaviside (Θ) function to select the samples according to the relevant output sign of the Teacher Perceptron. The constants $c_+ = 1/2\,\mathrm{Erfc}(-b_0/\sqrt{2})$, and $c_- = 1 - c_+$, represent respectively the normalization constant for positive and negative samples. Note that $\mathrm{d}\mu_{\text{train}}$ is in general not a Gaussian measure, since in the direction of $\mathbf{w}^0$, it is a piecewise Gaussian with normalization factors which depend on ρ_{train}: see Fig. 1 for a sketch in two dimensions.

Intrinsic, Train and Test Imbalance. Here, we elaborate on the definition and roles of the imbalance ratios ρ_0, ρ_{train} and ρ_{test}.

The introduction of ρ_{train} in Eq. (4) allows to control the amount of imbalance in the **training set**. This parameter is externally imposed and enables us to explore scenarios where the model is trained with varying levels of imbalance.

Instead, when samples are extracted from the Gaussian measure $\mathrm{d}\mu_{\text{pop}}(\{\mathbf{S}^\ell\}) \propto \prod_{\ell=1}^{\alpha N} D\mathbf{S}^\ell$, an imbalance between classes arises naturally due to the bias parameter b_0. It can be computed as:

$$\rho_0(b_0) = P\left(\frac{\mathbf{w}_0\cdot\mathbf{S}}{\sqrt{N}} + b_0 > 0 \right) = \frac{1}{2}\mathrm{Erfc}\left(-\frac{b_0}{\sqrt{2}} \right) = c_+ \,. \tag{5}$$

As soon as $b_0 \neq 0$, we have $\rho_0 \neq 0.5$, *i.e.* there is intrinsic imbalance. We refer to this ρ_0 as **Intrinsic Imbalance** as it measures the intrinsic imbalance present in the data generation process. It describes the rarity of observing anomalies and in our case, it depends solely on the teacher's bias. It can easily be visualized geometrically: when dealing with the population, all the samples are drawn from a Gaussian distribution centered at the origin, while varying b_0 amounts to translating the Teacher's hyperplane away from the origin of the N-dimensional space. Thus, one of the two classes lies on the tail of the distribution and is less represented. This is depicted in 2D sketches of Fig. 2(a,b,c), where a decreasing b_0 (in magnitude) results in a less biased teacher and a less imbalanced population.

Since the population distribution is rarely available in practice, a **test set** consisting of samples not observed during training is introduced in standard machine learning settings to test the performance of the trained model. We define the test distribution as

$$\mathrm{d}\mu_{\text{test}}(\mathbf{S}) = \frac{\rho_{\text{test}}}{c_+}\Theta\left(\frac{\mathbf{w}^0\cdot\mathbf{S}}{\sqrt{N}} + b_0 \right) D\mathbf{S} + \\ + \frac{1-\rho_{\text{test}}}{c_-}\Theta\left(-\frac{\mathbf{w}^0\cdot\mathbf{S}}{\sqrt{N}} - b_0 \right) D\mathbf{S}\,, \tag{6}$$

where ρ_{test} measures the probability of having an anomaly in the test set. Common choices are $\rho_{\text{test}} = 0.5$ (balanced test set) or $\rho_{\text{test}} = \rho_0(b_0)$ (test set that reflects the intrinsic imbalance). As we will discuss below, while some performance metrics do not explicitly depend on ρ_{test}, others do. The choice of the test imbalance is as important as the train imbalance, and can lead to misleading results if not properly considered, by inducing choices of ρ_{train} which do not allow to properly reconstruct the teacher model.

Informative samples. The teacher bias b_0 also determines how informative the two classes are. When $b_0 \neq 0$, the anomalous samples concentrate on the hyperplane of the teacher (in our convention $b_0 < 0$ and it is the class +1 which is the anomaly and concentrates), while the samples of the other class have a lower probability to lie close to this boundary. In particular, some of the anomaly-class points will be close to the hyperplane but far from the projection of the origin onto the hyperplane, and these are expected to be more in-

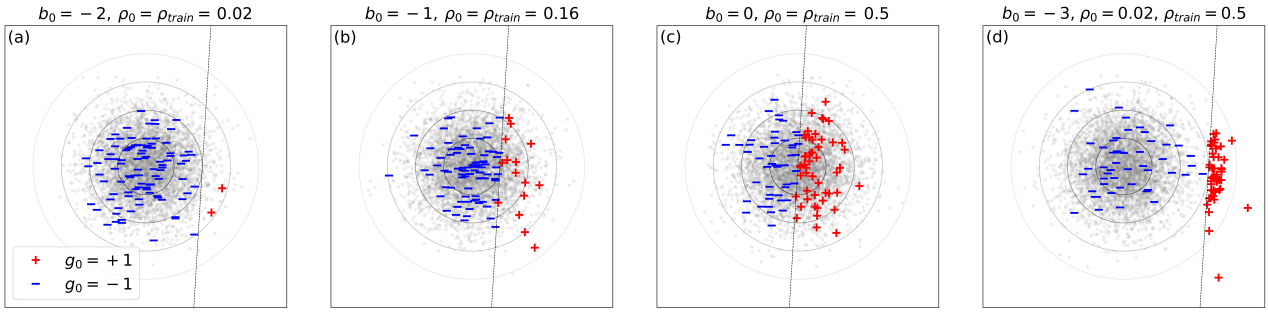

Figure 2: **Plots (a,b,c):** *Intrinsic Imbalance.* Two-dimensional sketches showing the effect of the teacher bias b_0 on the intrinsic imbalance ρ_0. The symbols in the figures are as in Fig. 1. All the examples are extracted from the underlying Gaussian distribution and no specific imbalance ratio is imposed externally. Increasing the magnitude of the teacher bias b_0 ((c) $\rightarrow$ (a)) translates the teacher's hyperplane (black dashed line) away from the origin and makes anomalies more rare, as the Gaussian tail is cut further away. **Plots (c,d):** *Informative samples.* Two cases are compared where the train imbalance is fixed to $\rho_{\text{train}} = 0.5$ while the teacher biased is varied. As $|b_0|$ grows, anomalies become more and more concentrated around the teacher's hyperplane, becoming more informative about the teacher's direction.

formative about its location. Fig. 2(c,d) shows this effect through two-dimensional sketches. In App. B we demonstrate this quantitatively, by calculating the density of each class close to the teacher's hyperplane. As $|b_0|$ grows, points close to this boundary are predominantly from the anomalous class. This effect is dominant when the student has perfect information about the teacher's bias, as discussed in App. E.

3 THEORETICAL ANALYSIS

3.1 Statistical Mechanics Approach

We are interested in finding the configurations of the student $(\mathbf{w}, b)$ that minimize the Loss in Eq. (3) and in computing the properties of these configurations over the distribution of input samples.

Statistical Mechanics (SM) comes in handy for this problem, since it allows us to exactly calculate properties of the typical solutions of our model, in the large-model limit, $N \rightarrow \infty$. This is the relevant limit for the study of our system, provided that the number of constraints is proportional to the number of weights: $\alpha = P/N$ is finite [Roberts et al., 2022]. Since the derivation of our results is long and technical, we here focus on the hypotheses and on the interpretation of the results, and provide the full calculation in App. C, where we calculate all the quantities described in the main paper, using Replica Theory (see *e.g.* [Altieri and Baity-Jesi, 2024] for a thoroughly referenced review).

SM assumes that the model configurations are sampled through Langevin dynamics at a temperature T, which roughly corresponds to the ratio between the learning rate and the batch size [Jastrzebski et al., 2017]. Analyzing the system at $T = 0$ returns the properties of the absolute minima of the loss; doing it for $T > 0$ identifies the typical configurations that are reached with a certain level of noise in the dynamics.

The main quantity is the free energy F, which is the negative log-probability of a weight configuration, so the most probable configurations are those that minimize F. In the $N \rightarrow \infty$ limit, the original high dimensional optimization problem reduces to a system of coupled deterministic equations for some order parameters, $(R, \hat{R}, q, \hat{q}, b)$, which must be solved through recursive iteration (see App. C). These parameters characterize the typical configurations that are reached by the Student at the end of the training, *i.e.* when training is long enough to reach an equilibrium state. This means that, once the values of the order parameters are given, any other quantity derives deterministically from them. The order parameters R and b have a crucial interpretation; R is the Teacher-Student overlap and it's defined as $R = \lim_{N\rightarrow\infty} \frac{\mathbf{w}\cdot\mathbf{w}^0}{N}$. It measures the typical alignment between the teacher's and the student's hyperplane. Instead, b represents the typical bias learned by the student, and should be compared to the teacher's value, b_0.

Calculating Metrics. For any given choice of the hyperparameters $b_0, \rho_{\text{train}}, T$ and α (these are also called control parameters, although b_0, α are intrinsic to the ML task and cannot be controlled in practical settings), we can solve the set of self-consistent equations (Eqs. (71–75) in App. C) numerically, to obtain the order parameters, with particular interest in (R, b).

The generalization error and the other performance metrics on the test-set are all derived from the order

parameters. For any metric $M(\mathbf{w}, b; \mathbf{S})$, its generalization value is:

$$M_g = \langle\langle\langle \mathbb{E}_T[M(\mathbf{w}, b; \mathbf{S})]\rangle\rangle_{\mu_{\text{train}}}\rangle\rangle_{\mu_{\text{test}}}\,, \tag{7}$$

where $\mathbb{E}_T[\dots]$ denotes the average over realizations of the thermal noise, and $\langle\langle\dots\rangle\rangle_\mu$ the average over a chosen dataset measure μ. The idea behind the computation of generalization metrics is to add one sample that was not observed during training and evaluate the performance of the trained student on it. In App. D, we provide the derivation of the expressions of all the quantities shown in the paper.

The performance metrics we are interested in are: Recall (r), Specificity (s), Accuracy (a), Balanced Accuracy (a_{bal}), Positive Predicted Value or Precision (p), F1-Score (F_1), generalization error (ϵ_g):

$$r = \frac{\#\text{True Positives}}{\#\text{Teacher Positives}}\,, \tag{8}$$

$$s = \frac{\#\text{True Negatives}}{\#\text{Teacher Negatives}}\,, \tag{9}$$

$$a = \rho_{\text{test}} r + (1-\rho_{\text{test}})s\,, \tag{10}$$

$$a_{\text{bal}} = \frac{(r+s)}{2}\,, \tag{11}$$

$$p = \frac{r}{r + \frac{1-\rho_{\text{test}}}{\rho_{\text{test}}}(1-s)}\,, \tag{12}$$

$$F_1 = 2\cdot\frac{pr}{p+r}\,, \tag{13}$$

$$\epsilon_g = 1 - a\,. \tag{14}$$

3.2 Theoretical Results

Good and bad models, Energy-Entropy interplay. We now investigate the influence of ρ_{train} on the learned model. Figure 3(a,b), reports the solution of the self-consistent equations [1] for the overlap R and the learned bias b as a function of ρ_{train}, for multiple choices of the teacher bias b_0. Learning under strong imbalance leads to a model that has a strong bias and low alignment with the teacher, this is a *bad* model since it is not able to reproduce b_0 and $\mathbf{w}_0$ correctly. Meanwhile, learning on a more balanced training set, leads to a *good* model that is able to better reproduce the teacher's labeling rule, learning a good overlap R and not being overly biased. This phenomenon can be geometrically interpreted (Fig. 1(a,c)): since the loss counts the number of misclassified examples, a dummy model that has a strong bias and always predicts the majority class pays a small price in terms of loss (for small ρ_{train}). However, such a model is statistically

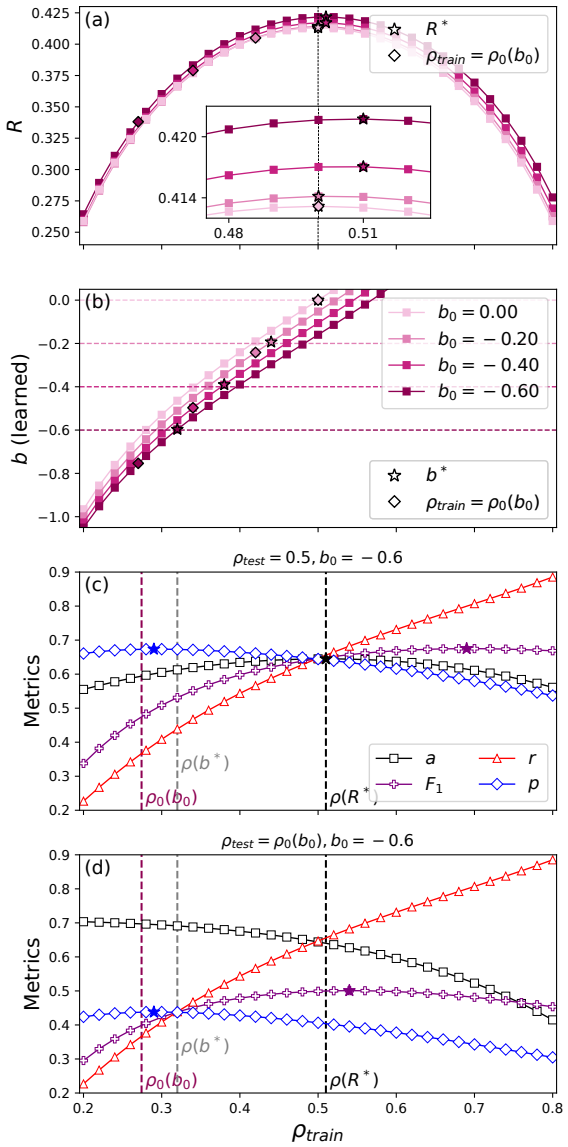

Figure 3: *Performance as function of* ρ_{train}. Analytical results as a function of ρ_{train}, for $\alpha = 1.1$ and $T = 0.5$. **(a):** Student overlap R, for $b_0 = 0, -0.2, -0.4, -0.6$ ($\rho_0 = 0.5, 0.42, 0.34, 0.27$). Stars indicate maximal overlap point, diamonds indicate $\rho_{\text{train}} = \rho_0$. Vertical line indicates $\rho_{\text{train}} = 0.5$. Inset is a zoom. **(b):** As in (a), but for the student bias b. Horizontal lines indicate b_0. Stars now indicate points where bias is learned perfectly ($b = b_0$), diamonds indicate $\rho_{\text{train}} = \rho_0$. **(c):** Metrics for $b_0 = -0.6$, $\rho_{\text{test}} = 0.5$. Stars indicate the peak of each curve. Vertical lines indicate ρ_0 (imbalance $\rho(b^*)$ at which the bias is optimal), and that at which the overlap is optimal, $\rho(R^*)$. **(d):** Same as (c), but for $\rho_{\text{test}} = \rho_0(b_0)$.

[1] Although our results are analytical, each point in the plot is a new solution of the system of self-consistent equations (Eqs. (71–75) in the appendix). For this reason, our analytical results are provided as points instead of a continuous curve.

favored with respect to a model with $b = b_0$, since with $|b| \gg |b_0|$, there is a very large number of weight configurations $\mathbf{w}$ that allow for the same small training error, while with $b = b_0$ the number of weight configurations giving a small error is much lower. This is an example of what is called energy-entropy interplay [Carbone et al., 2020]: solutions with large b have an entropic advantage (there are more of them), at the expense of a few misclassifications (what is sometimes called an *energetic cost*). As ρ_{train} becomes more balanced, the energetic cost increases, eventually overcoming the entropic advantage. We also note that, while the situations in Figs. 1(a,c) are similar with what regards the training errors, they are of course very dissimilar when one looks into the generalization error (accuracy curve in Fig. 3(d)).
Finally, we highlight that training at $\rho_{\text{train}} = \rho_0$ is always suboptimal, both in terms of R and of b.

Performance metrics, which one? When testing a trained model, it is common practice to build a balanced test set with samples not observed during training, *i.e.*, $\rho_{\text{test}} = 0.5$. Another practice is to leave the distribution untouched, $\rho_{\text{test}} = \rho_0(b_0)$. How are different performance metrics affected by the test imbalance, and which metric is best able to identify a *good* model in terms of R and b? Figure 3(c,d) reports different performance metrics for models trained with varying ρ_{train} and tested with $\rho_{\text{test}} = 0.5$ and $\rho_{\text{test}} = \rho_0$.
The recall is trivially maximized for $\rho_{\text{train}} = 1$, since this generates a dummy model which identifies anything as an anomaly. The sensitivity s has the opposite trend (Fig. 9 in App. D), being trivially maximized for $\rho_{\text{train}} = 0$.
The balanced accuracy (shown implicitly as accuracy in Fig. 3(c), since a_{bal} coincides with a when $\rho_{\text{test}} = 0.5$, Eq. (11)) is the quantity that best reproduces $R(\rho_{\text{train}})$, and shares with $R(\rho_{\text{train}})$ the feature of not depending on ρ_{test}.
For $\rho_{\text{test}} = \rho_0$, the accuracy a is maximized at small ρ_{train}, because this generates a model which always guesses the majority class. This can be understood from Eq. (10): if $\rho_{\text{test}} \sim 0$ then the contribution of s dominates.
The trend of the precision p (confirmed in App. F) seems independent of ρ_{test}, though its specific value is.

The precision p peaks between ρ_0 and $\rho(b^*)$, making it the best candidate to identify b_0. Although the precision is a metric that is notoriously dependent on ρ_{test} [Burkhard et al., 2025], the position of the peak does not seem to have a strong dependence on ρ_{test}. While, quantitatively, this observation may vary with α and T, we show in App. F that p is still the metric that peaks closest to $\rho_0(b_0)$.
Finally, the F_1 score, when calculated with $\rho_{\text{test}} = 0.5$, peaks at a value representing a low R and a bias which is underestimated compared to b_0.
Comparing panels (c) and (d) we note that some metrics (a or F_1 score) do not peak at the same location, highlighting the relevance of the choice of ρ_{test}.
In summary, no known metric perfectly matches the optimal overlap (R^*) nor the optimal bias (b^*) over the whole space of control parameters (b_0, α, T) but we identify a_{bal} as the most suitable metric to identify the overlap, due to the qualitative agreement with R and the independence from ρ_{test}, and p as the most suited to track the optimal b.

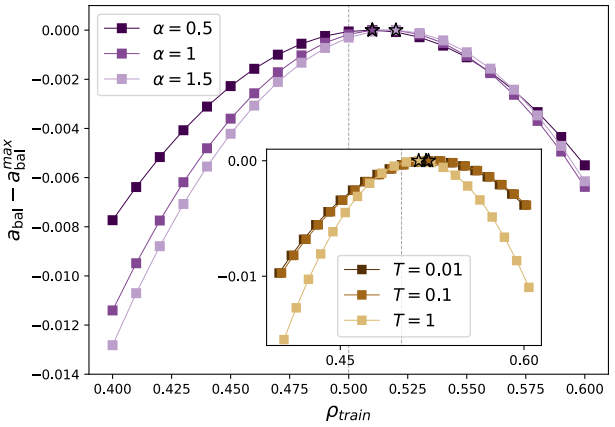

Figure 4: *Dependence on ρ_{train} for different α or T.* The optimal balanced accuracy (a_{bal}). We plot a_{bal} as a function of ρ_{train}, shifted so that all the curves peak at 0. The vertical dotted lines indicate $\rho_{\text{train}} = 0.5$. **Main:** study at $b_0 = -1$ and $T = 0.5$. Varying α changes the position of the peak, as well as how fast the performance decreases when leaving the peak. **Inset:** same b_0 and fixed $\alpha = 1.1$, varying T. The cases $T = 0.01$ and $T = 0.1$ are almost impossible to distinguish because they both correspond to the low-temperature region (Fig. 5). For high T the curvature is larger.

Optimal train imbalance. We already noted that training at $\rho_{\text{train}} = \rho_0$ never gives the best model (in terms of best R nor b). We now turn to $\rho_{\text{train}} = 0.5$, which is commonly believed to lead to optimal generalization performances, and the most common choice in CI reweighting/resampling schemes. We challenge this assumption, showing that the optimal train imbalance, $\rho_{\text{train}}(R^*) = \text{argmax}_{\rho_{\text{train}}}(R)$, is different from 0.5. This is true for the overlap (Fig. 3(a)–inset), and it is also true for the best proxy of the overlap, a_{bal}. Fig. 4 shows that, when $b_0 < 0$ and $\alpha = 1.1$, a_{bal} peaks at $\rho_{\text{train}} > 0.5$, *i.e.* when there are slightly more of the anomalous examples. In App. F we investigate extreme values of $\rho_0(b_0)$ (down to $\rho_0 = 3 \cdot 10^{-7}$) and

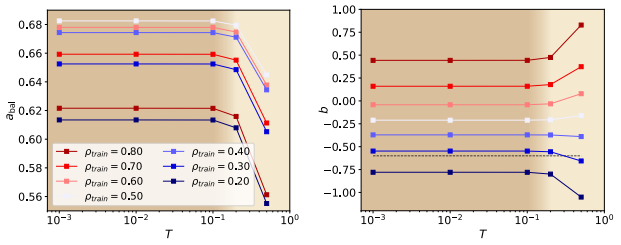

Figure 5: *Performance as function of T.* **Left:** Balanced accuracy as a function of temperature T, for $\alpha = 1.1$. The teacher bias is $b_0 = -0.6$ (dotted horizontal line in the inset, $\rho_0 = 0.27$). **Right:** Same, for the learned bias b.

confirm these findings. This is consistent with previous empirical observations on SVMs and Random Forests, which found that $\rho_{\text{train}} = 0.5$ is not the optimal training ratio [Kamalov et al., 2022].

We also look into the influence of the degree of data abundance (α) and the amount of noise in the dynamics (T) on the value of $\rho_{\text{train}}(R^*)$: for the values considered in Fig. 4, increasing α and decreasing T make the curves more tilted, shifting the optimal train imbalance, and the curves more peaked, thus increasing the penalty for choosing $\rho_{\text{train}} \neq \rho_{\text{train}}(R^*)$.

In App. F and in particular in Fig. 10 we further investigate these effects, showing that $\rho_{\text{train}}(R^*)$ depends on α, T and b_0; that the effect is non-monotonic; and it can shift $\rho_{\text{train}}(R^*)$ both to values > 0.5 (as in Fig. 4) and < 0.5.
We argue that this is the result of two competing effects. On one side, as depicted in Fig. 2(d), minority class examples are more informative, so it is more convenient to train with more of those (*i.e.* increase ρ_{train}). On the other side, if $|b_0|$ is large enough, there is a large region $\mathcal{R}$ between typical negatives and (informative) positives that is empty of points, thus allowing for many possible hyperplane directions $\mathbf{w}$ that separate the training set. This means that a large fraction of student models with $|b| < |b_0|$ will result in a small error. Since there are many more weight configurations allowing for $|b| < |b_0|$ than weight configurations allowing $b = b_0$, configurations with a wrong b and $\mathbf{w}$ are entropically favored. One way to decrease this entropic contribution is to fill the region $\mathcal{R}$ with negatives, *i.e.* increase the proportion of negatives (decrease ρ_{train}).

Interplay between noise and CI. We investigate the impact of the amount of noise in the dynamics (T) on the quality of the learned model. Fig. 5 shows a crossover around a temperature T^*. Below T^* the performance is optimal, and largely unaffected by the noise level. Above T^*, the performance degrades and becomes sensitive to any additional amount of noise. By comparing with Fig. 4–inset, we see that the low performance in the high-noise regime is connected to its lower tolerance to non-optimal values of ρ_{train} (more peaked shape).

A similar effect has already been observed in the teacher-student perceptron, from a dynamical perspective, by studying the regimes of Stochastic Gradient Descent (SGD) as shown in [Sclocchi and Wyart, 2024]. They identify a Gradient Descent-like regime with low noise and optimal performances and a noise-dominated regime where performance deteriorates as noise increases. Here, we elucidate the interplay between the noise level and the train imbalance. In particular, we observe that ρ_{train} determines the evolution of the student's bias with T: increasing T favors the entropic contribution discussed earlier in this section (seen in Fig. 1) and this pushes towards overly-biased models, because there are more of such solution: *i.e.* dummy models that classify most points as positives (for large ρ_{train}) or as negatives (for smaller ρ_{train}), despite making a number of (training) errors. This is a typical energy-entropy interplay. To summarize, it is the difficulty to guess the correct bias (or guess the corresponding ρ_0) that explains the noise dependence.

3.3 Experiments

Most of the findings discussed above emerge from the study of the static properties of the loss landscape. The dynamics (which in our calculations is Langevin) enters the discussion only through the parameter T, which measures the amount of stochastic uncorrelated noise. Since in practical experiments the dynamics used is SGD, our experiments are with SGD dynamics too. In App. G we provide empirical evidence which is consistent with our analytical findings, in three cases: (i) in the case of a Teacher-Student Perceptron on Gaussian data with SGD dynamics; (ii) in the case of MLPs or pretrained ResNet backbone, trained on AD CIFAR-10; (iii) in the case of a pretrained ResNet backbone followed by PCA and Perceptron, on a smaller anomaly dataset, BTAD [Mishra et al., 2021].

4 DISCUSSION

We analyzed the effect of ADI on learning, through exact analytical calculations, which are compatible with experiments in realistic settings. In addition to the train and test imbalance, ρ_{train} and ρ_{test}, in ADI we can identify an intrinsic imbalance, ρ_0, over which practitioners have no control. If data generation is un-

biased and no rebalancing of the class distributions is performed and the dataset faithfully represents the deployment distribution, then one has $\rho_0 = \rho_{\text{train}} = \rho_{\text{test}}$. Although our analysis stresses the importance of three kinds of imbalance, we highlight that more sources might arise depending on the situation. For example, it was recently highlighted that, in common architectures, further phenomena akin to class imbalance can occur at the beginning of learning [Francazi et al., 2024].

Varying ρ_{train} corresponds to rebalancing the training distribution. Since our results are in the asymptotic data limit, they equally represent the effect of both class reweighting and resampling. Note, however, that these two rebalancing strategies influence SGD differently [Francazi et al., 2023]. Our work shows that the value of ρ_{train} which maximizes the overlap between teacher and student is generally not 0.5. This is consistent with previous empirical work [Kamalov et al., 2022], where on different kind of architectures it was shown that re-sampling using some $\rho_{\text{train}} < 0.5$ was consistently optimal over a broad range of tasks and models. The case $\rho_{\text{train}} > 0.5$ was however not explored. A trend was observed: as data is initially more abundant (corresponding to larger α for us), more resampling can be done. Our results show that the picture can in general be more complex. In fact, this deviation from $\rho^*_{\text{train}} = 0.5$ depends non-linearly on ρ_0 and on α. While α indicates how much data is available in comparison with the model size, in our linear classifier it also indicates the dimensionality of the input space. Therefore, we cannot disentangle whether this effect is due to model size or to input dimensionality. We also find that the importance of this deviation is amplified in dynamics with a strong noise (*e.g.* large learning rate), with small-noise dynamics leading to better solutions than larger-noise ones, with a clear separation between two regimes.
This asymmetry is at least in part a consequence of the fact that, in ADI, examples from different classes are intrinsically not equally informative. While this asymmetry was, to our knowledge, not observed in previous work on MGI, we believe that similar deviations from $\rho_{\text{train}}(R^*) = 0.5$ can also be observed, in cases where different classes inform differently on the classification boundary (*e.g.* a class having smaller variance). In particular, in MG classification, we conjecture that this asymmetry could also be observed in the absence of imbalance. In fact, while in ADI, $\rho_{\text{train}}(R^*) \neq 0.5$ is a feature of the imbalance, in MG classification it can be a feature of the data structure. A consequence is that, under any kind of imbalance, even ρ_{train} itself may also be considered as a hyper-parameter.

From the point of view of what happens to the training landscape when varying ρ_{train}, we see that it causes smooth variations in the solutions, with no abrupt changes even when values such as $\rho_{\text{train}} = \rho_0$ or 0.5 are crossed. We notice such an absence of phase transitions also when tuning ρ_0 and ρ_{test}.
Varying ρ_{train} and ρ_{test}, and the evaluation metrics, informs us on what each metric reproduces. The balanced accuracy seems the best proxy for the teacher overlap R, while the quantity that best reflects the bias is the precision p.

The choice of ρ_{test} also has an influence on the *way* data is split between training and test set. While it is common practice to split according to the metadata, the data is sometimes stratified in terms of output value [Zubrod et al., 2023]. Because minority class examples are more informative than majority class examples, the choice of ρ_{test} can influence not only the meaning of the optimum, but also the sheer value of the measured metrics. For example, the F1 score, despite being macro-averaged, has higher values when calculated with $\rho_{\text{test}} = 0.5$ than with $\rho_{\text{test}} = \rho_0$.

5 CONCLUSION

Our analysis of ADI within an exactly solvable model offers both conceptual insights—such as distinguishing ADI from MGI—and practical implications. In particular we challenge the common knowledge of a balanced training set being optimal, and we show that smaller learning rates are less sensitive to imbalance.

Acknowledgments

We thank E. Loffredo, B. Loureiro, M. Pastore for insightful conversations. This work was supported by the Swiss National Foundation, SNF grant # 196902. This work is supported by a public grant overseen by the French National Research Agency (ANR) through the program UDOPIA, project funded by the ANR-20-THIA-0013-01. This work is supported by the French government under the France 2030 program (PhOM - Graduate School of Physics) with reference ANR-11-IDEX-0003.

References

[Almeida et al., 2011] Almeida, T. A., Hidalgo, J. M. G., and Yamakami, A. (2011). Contributions to the study of sms spam filtering: new collection and results. In *Proceedings of the 11th ACM symposium on Document engineering*, pages 259–262.

[Altieri and Baity-Jesi, 2024] Altieri, A. and Baity-Jesi, M. (2024). An introduction to the theory of spin glasses. In Chakraborty and Tapash, editors,

Encyclopedia of Condensed Matter Physics, 2e, volume 2, pages 361–370. Elsevier, Oxford.

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

[Ando and Huang, 2017] Ando, S. and Huang, C.-Y. (2017). Deep Over-sampling Framework for Classifying Imbalanced Data. arXiv:1704.07515 [cs, stat].

[Behnia et al., 2023] Behnia, T., Kini, G. R., Vakilian, V., and Thrampoulidis, C. (2023). On the Implicit Geometry of Cross-Entropy Parameterizations for Label-Imbalanced Data.

[Bergmann et al., 2019] Bergmann, P., Fauser, M., Sattlegger, D., and Steger, C. (2019). MVTec AD — A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9584–9592, Long Beach, CA, USA. IEEE.

[Burkhard et al., 2025] Burkhard, Y., Francazi, E., Lavender, E., Volpi, M., Volpi, M., Baity-Jesi, M., and Moor, H. (2025). Single species identification in camera trap images: architecture choice, training strategies, and the meaning of performance metrics. In preparation.

[Carbone et al., 2020] Carbone, M. R., Astuti, V., and Baity-Jesi, M. (2020). Effective traplike activated dynamics in a continuous landscape. *Phys. Rev. E*, 101:052304.

[Charbonneau et al., 2023] Charbonneau, P., Marinari, E., Parisi, G., Ricci-tersenghi, F., Sicuro, G., Zamponi, F., and Mezard, M. (2023). *Spin Glass Theory and Far Beyond: Replica Symmetry Breaking after 40 Years*. World Scientific.

[Chawla et al., 2002] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357.

[Fontanari and Meir, 1993] Fontanari, J. F. and Meir, R. (1993). The statistical mechanics of the Ising perceptron. *Journal of Physics A: Mathematical and General*, 26(5):1077–1089.

[Francazi et al., 2023] Francazi, E., Baity-Jesi, M., and Lucchi, A. (2023). A theoretical analysis of the learning dynamics under class imbalance. In *International Conference on Machine Learning*, volume 202, pages 10285–10322. PMLR.

[Francazi et al., 2024] Francazi, E., Lucchi, A., and Baity-Jesi, M. (2024). Initial guessing bias: How untrained networks favor some classes. In *International Conference on Machine Learning*, volume 235, pages 13783–13839. PMLR.

[Gardner and Derrida, 1989] Gardner, E. and Derrida, B. (1989). Three unfinished works on the optimal storage capacity of networks. *Journal of Physics A: Mathematical and General*, 22(12):1983.

[Györgyi, 1990] Györgyi, G. (1990). First-order transition to perfect generalization in a neural network with binary synapses. *Physical Review A*, 41(12):7097–7100.

[He et al., 2015] He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

[Huang, 1987] Huang, K. (1987). *Statistical Mechanics*. John Wiley and Sons, Hoboken, NJ, second edition.

[Japkowicz and Stephen, 2002] Japkowicz, N. and Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5):429–449.

[Jastrzebski et al., 2017] Jastrzebski, S., Kenton, Z., Arpit, D., Ballas, N., Fischer, A., Bengio, Y., and Storkey, A. (2017). Three factors influencing minima in sgd. *arXiv:1711.04623*.

[Kamalov et al., 2022] Kamalov, F., Atiya, A. F., and Elreedy, D. (2022). Partial Resampling of Imbalanced Data. arXiv:2207.04631 [cs].

[Kini et al., 2021] Kini, G. R., Paraskevas, O., Oymak, S., and Thrampoulidis, C. (2021). Label-Imbalanced and Group-Sensitive Classification under Overparameterization.

[Kunstner et al., 2024] Kunstner, F., Yadav, R., Milligan, A., Schmidt, M., and Bietti, A. (2024). Heavy-tailed class imbalance and why adam outperforms gradient descent on language models. *arXiv:2402.19449*.

[Kyathanahally et al., 2021] Kyathanahally, S. P., Hardeman, T., Merz, E., Bulas, T., Reyes, M., Isles, P., Pomati, F., and Baity-Jesi, M. (2021). Deep learning classification of lake zooplankton. *Frontiers in microbiology*, page 3226.

[Loffredo et al., 2024] Loffredo, E., Pastore, M., Cocco, S., and Monasson, R. (2024). Restoring balance: principled under/oversampling of data for optimal classification. arXiv:2405.09535 [cond-mat].

[Mannelli et al., 2023] Mannelli, S. S., Gerace, F., Rostamzadeh, N., and Saglietti, L. (2023). Unfair geometries: exactly solvable data model with fairness implications. arXiv:2205.15935 [cond-mat, stat].

[Menon et al., 2021] Menon, A. K., Jayasumana, S., Rawat, A. S., Jain, H., Veit, A., and Kumar, S. (2021). Long-tail learning via logit adjustment. arXiv:2007.07314 [cs, stat].

[Mézard et al., 1987] Mézard, M., Parisi, G., and Virasoro, M. (1987). *Spin-Glass Theory and Beyond*. World Scientific, Singapore.

[Mezard et al., 1987] Mezard, M., Parisi, G., and Virasoro, M. (1987). *Spin Glass Theory And Beyond: An Introduction To The Replica Method And Its Applications*. World Scientific Lecture Notes In Physics. World Scientific Publishing Company.

[Mignacco et al., 2020] Mignacco, F., Krzakala, F., Lu, Y., Urbani, P., and Zdeborova, L. (2020). The role of regularization in classification of high-dimensional noisy gaussian mixture. In *International conference on machine learning*, pages 6874–6883. PMLR.

[Mishra et al., 2021] Mishra, P., Verk, R., Fornasier, D., Piciarelli, C., and Foresti, G. L. (2021). VT-ADL: A vision transformer network for image anomaly detection and localization. *CoRR*, abs/2104.10036.

[Nishimori, 2001] Nishimori, H. (2001). *Statistical physics of spin glasses and information processing: an introduction*. Number 111. Clarendon Press.

[Reiss et al., 2020] Reiss, T., Cohen, N., Bergman, L., and Hoshen, Y. (2020). PANDA - adapting pretrained features for anomaly detection. *CoRR*, abs/2010.05903.

[Roberts et al., 2022] Roberts, D. A., Yaida, S., and Hanin, B. (2022). *The principles of deep learning theory*, volume 46. Cambridge University Press Cambridge, MA, USA.

[Schür et al., 2023] Schür, C., Gasser, L., Perez-Cruz, F., Schirmer, K., and Baity-Jesi, M. (2023). A benchmark dataset for machine learning in ecotoxicology. *Scientific Data*, 10(1):718.

[Sclocchi and Wyart, 2024] Sclocchi, A. and Wyart, M. (2024). On the different regimes of Stochastic Gradient Descent. *Proceedings of the National Academy of Sciences*, 121(9):e2316301121. arXiv:2309.10688 [cond-mat, stat].

[Seung et al., 1992] Seung, H. S., Sompolinsky, H., and Tishby, N. (1992). Statistical mechanics of learning from examples. *Physical Review A*, 45(8):6056–6091.

[Tang et al., 2020] Tang, K., Huang, J., and Zhang, H. (2020). Long-tailed classification by keeping the good and removing the bad momentum causal effect. *Advances in neural information processing systems*, 33:1513–1524.

[Thrampoulidis et al., 2022] Thrampoulidis, C., Kini, G. R., Vakilian, V., and Behnia, T. (2022). Imbalance Trouble: Revisiting Neural-Collapse Geometry.

[Xie and Manski, 1989] Xie, Y. and Manski, C. F. (1989). The Logit Model and Response-Based Samples. *Sociological Methods & Research*, 17(3):283–302.

[Yamanishi et al., 2000] Yamanishi, K., Takeuchi, J.-I., Williams, G., and Milne, P. (2000). On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 320–324.

[Ye et al., 2021] Ye, H.-J., Zhan, D.-C., and Chao, W.-L. (2021). Procrustean training for imbalanced deep learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 92–102.

[Zhang and Mani, 2003] Zhang, J. and Mani, I. (2003). kNN Approach to Unbalanced Data Distributions: A Case Study involving Information Extraction.

[Zhang et al., 2019] Zhang, L., Shen, X., Zhang, F., Ren, M., Ge, B., and Li, B. (2019). Anomaly detection for power grid based on time series model. In *2019 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*, pages 188–192. IEEE.

[Zubrod et al., 2023] Zubrod, J. P., Galic, N., Vaugeois, M., and Dreier, D. A. (2023). Physiological variables in machine learning qsars allow for both cross-chemical and cross-species predictions. *Ecotoxicology and Environmental Safety*, 263:115250.

Checklist

The checklist follows the references. For each question, choose your answer from the three possible options: Yes, No, Not Applicable. You are encouraged to include a justification to your answer, either by referencing the appropriate section of your paper or providing a brief inline description (1-2 sentences). Please do not modify the questions. Note that the Checklist section does not count towards the page limit. Not including the checklist in the first submission won't result in desk rejection, although in such case we will ask you to upload it during the author response period and include it in camera ready (if accepted).

In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [**Yes**/No/Not Applicable]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes/No/**Not Applicable**] **We are not proposing new numerical methods.**
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/No/Not Applicable] **This is a mostly theoretical paper.**

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [**Yes**/No/Not Applicable] All assumptions are provided.
   (b) Complete proofs of all theoretical results. [**Yes**/No/Not Applicable]
   (c) Clear explanations of any assumptions. [Yes/No/Not Applicable] **An intuition around any assumptions is provided.**

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [**Yes**/No/Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [**Yes**/No/Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [**Yes**/No/Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [**Yes**/No/Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes/No/**Not Applicable**]
   (b) The license information of the assets, if applicable. [Yes/No/**Not Applicable**]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes/No/**Not Applicable**]
   (d) Information about consent from data providers/curators. [Yes/No/**Not Applicable**]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes/No/**Not Applicable**]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Yes/No/**Not Applicable**]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Yes/No/**Not Applicable**]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Yes/No/**Not Applicable**]

Appendix to "Class Imbalance in Anomaly Detection: Learning from an Exactly Solvable Model"

A TEACHER-STUDENT SETUP

The teacher-student setup has been widely studied in learning theory since the seminal work [Gardner and Derrida, 1989] on the perceptron model. Figure 6 illustrates this setup, along with the notation used in this work for the teacher and student weights and bias parameters.

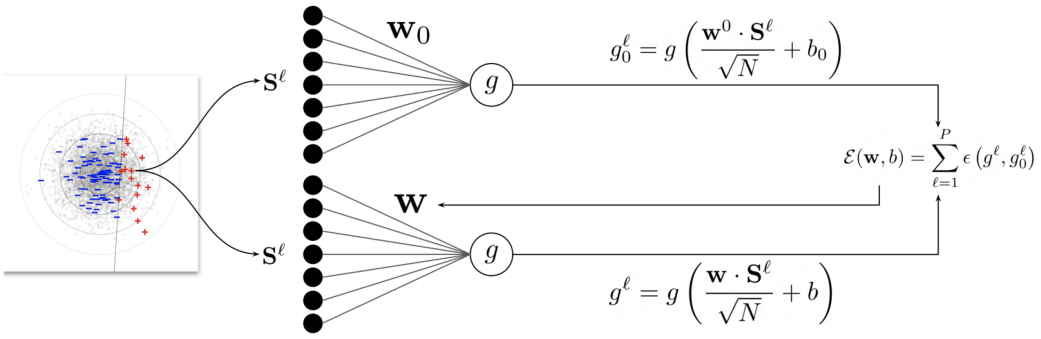

Figure 6: Schematic representation of the perceptron model within the teacher-student setup. Labels for the data $\mathbf{S}^\ell$ are generated by a teacher network with weights $\mathbf{w}_0$, bias b_0 and activation function g. The student network (bottom) has the same activation as the teacher (top), and learns its weights $\mathbf{w}$ and bias b by minimizing a loss $\mathcal{E}$ summed over all data, where ϵ measures a distance between the label of the data and the student outcome.

B INFORMATIVE SAMPLES

Here we show that, in the ADI setting, minority class examples are much closer to the teacher's hyperplane (decision boundary) than majority class examples are. We do this by evaluating the probability of a sample lying on the boundary. Let us consider a sample $\mathbf{S}$ extracted from an N-dimensional Gaussian distribution centered at the origin,

$$P(\mathbf{S}) = \frac{1}{(2\pi)^{N/2}} e^{-1/2 \sum_{i=1}^{N} S_i^2} . \tag{15}$$

As we did in the main paper, we study the case $b_0 < 0$ (the positives are the anomalies). We want to compute the probability $P^{+}_{boundary}(b_0)$ that the sample lies between the teacher's hyperplane and the parallel one at a distance δx, conditioned to $g_0 > 0$. Namely,

$$P^{+}_{boundary}(b_0) \equiv P\left(0 < \frac{\mathbf{w}_0 \cdot \mathbf{S}}{\sqrt{N}} + b_0 < \delta x \middle| \frac{\mathbf{w}_0 \cdot \mathbf{S}}{\sqrt{N}} + b_0 > 0\right) = \frac{P\left(0 < \frac{\mathbf{w}_0 \cdot \mathbf{S}}{\sqrt{N}} + b_0 < \delta x\right)}{P\left(\frac{\mathbf{w}_0 \cdot \mathbf{S}}{\sqrt{N}} + b_0 > 0\right)} . \tag{16}$$

The denominator is equal to

$$P\left(\frac{\mathbf{w}_0 \cdot \mathbf{S}}{\sqrt{N}} + b_0 > 0\right) = \int D\mathbf{S} \Theta\left(\frac{\mathbf{w}_0 \cdot \mathbf{S}}{\sqrt{N}} + b_0\right) \tag{17}$$

$$= \int \mathrm{d}y \Theta(y + b_0) \int D\mathbf{S} \delta\left(y - \frac{\mathbf{w}_0 \cdot \mathbf{S}}{\sqrt{N}}\right) . \tag{18}$$

By exploiting the Fourier representation of the Dirac δ distribution one gets

$$P\left(\frac{\mathbf{w}_0 \cdot \mathbf{S}}{\sqrt{N}} + b_0 > 0\right) = \int \mathrm{d}y \Theta(y + b_0) \int \frac{\mathrm{d}\hat{y}}{2\pi} e^{i\hat{y}y} \int D\mathbf{S} e^{-i\hat{y}\frac{\mathbf{w}_0 \cdot \mathbf{S}}{\sqrt{N}}} \tag{19}$$

$$= \int \mathrm{d}y \Theta(y + b_0) \int \frac{\mathrm{d}\hat{y}}{2\pi} e^{i\hat{y}y - 1/2\hat{y}^2} \tag{20}$$

$$= \int \Theta(y + b_0) Dy = \frac{1}{2} \mathrm{Erfc}\left(\frac{-b_0}{\sqrt{2}}\right) . \tag{21}$$

As for the numerator, the computation follows the same lines and one obtains

$$P\left(0 < \frac{\mathbf{w}_0 \cdot \mathbf{S}}{\sqrt{N}} + b_0 < \delta x\right) = \int_{-b_0}^{-b_0 + \delta x} Dy \xrightarrow{\delta x \to 0} \frac{1}{\sqrt{2\pi}} e^{-b_0^2/2} \delta x . \tag{22}$$

Similarly, one can compute the probability $P^{-}_{boundary}(b_0)$ for samples with $g_0 < 0$, and finally get

$$P^{+}_{boundary}(b_0) = \frac{\frac{1}{\sqrt{2\pi}} e^{-b_0^2/2} \delta x}{\frac{1}{2} \mathrm{Erfc}\left(\frac{-b_0}{\sqrt{2}}\right)} , \tag{23}$$

$$P^{-}_{boundary}(b_0) = \frac{\frac{1}{\sqrt{2\pi}} e^{-b_0^2/2} \delta x}{1 - \frac{1}{2} \mathrm{Erfc}\left(\frac{-b_0}{\sqrt{2}}\right)} . \tag{24}$$

These two quantities are shown in Fig. 7 as a function of $b_0 < 0$ (in order to eliminate the dependence on δx, we may use the corresponding densities; and to set the scale we normalize these quantities by their value at $b_0 = 0$, which has the property $P^{-}_{boundary}(0) = P^{+}_{boundary}(0)$). It is clear that positive samples have a higher probability to lie on the boundary (hyperplane) as long as the teacher bias is negative, so they represent the minority class in the population distribution. It is noteworthy that these quantities are actually independent of N.

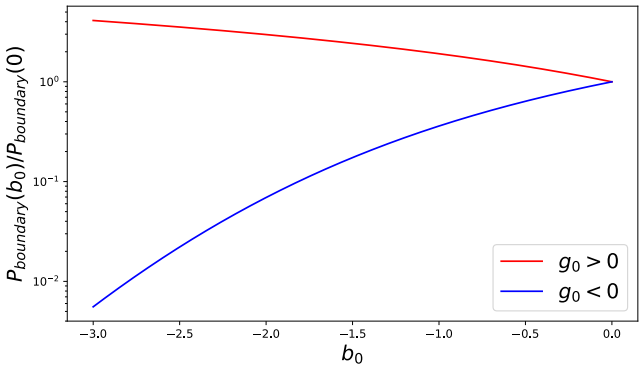

Figure 7: Probability of being close to the boundary (hyperplane), for the minority ($g_0 > 0$) and for the majority class ($g_0 < 0$). As a function of b_0, we plot the quantities $P^-_{boundary}(b_0)$ and $P^+_{boundary}(b_0)$, divided by their value at $b_0 = 0$.

C STATISTICAL MECHANICS AND REPLICA CALCULATION

C.1 Statistical mechanics setting

Statistical mechanics allows to infer macroscopic properties from a large number of interacting agents. In our case, the agents are the model weights, and the properties are quantities such as the performance metrics.

The main quantity we want to calculate is the free energy, F, which is a function of the model weights. The minima of F indicate the typical configurations that are assumed by the trained model [Huang, 1987].

To calculate the free energy, one usually first calculates the partition function,

$$Z = \int \mathrm{d}\mu(\mathbf{w}) \int \mathrm{d}\mu(b) e^{-\beta \mathcal{E}(\mathbf{w},b)} . \tag{25}$$

From the partition function, one can obtain the free energy through:

$$F = -\frac{1}{\beta} \langle\!\langle \log Z \rangle\!\rangle_{\mu_{\text{train}}} , \tag{26}$$

where $\beta = 1/T$ is the inverse temperature.

Eq. (26) is formally simple, but it involves calculating an integral of the logarithm of a non-trivial function, which is a hard-to-tract problem. Replica theory allows us to calculate F by calculating the average of Z^n (where n is a parameter), instead of the average of $\log Z$.

C.2 Replica Calculation

The goal of the Replica Calculation is to compute the *Quenched* Free-Energy of the system [Mézard et al., 1987],

$$-\beta F = \langle\!\langle \log Z \rangle\!\rangle_{\mu_{\text{train}}} = \langle\!\langle \log \int \mathrm{d}\mu(\mathbf{w}) \int \mathrm{d}\mu(b) e^{-\beta \mathcal{E}(\mathbf{w},b)} \rangle\!\rangle_{\mu_{\text{train}}} , \tag{27}$$

where Z is the partition function, and $\langle\!\langle \ldots \rangle\!\rangle_{\mu_{\text{train}}} = \int \mathrm{d}\mu_{\text{train}}(\{\mathbf{S}^\ell\})$ denotes the average over the distribution of the training data. $\mathcal{E}$ is the training loss as defined in the main text,

$$\mathcal{E}(\mathbf{w}, b) = \sum_{\ell=1}^{P} \epsilon \left(g\left(\frac{\mathbf{w} \cdot \mathbf{S}^\ell}{\sqrt{N}} + b \right), g_0^\ell \right) , \tag{28}$$

with $\epsilon(x, y) = \frac{1}{2}(x - y)^2$, square loss and P number of training samples. We will also use another shorthand notation for the loss, $\mathcal{E}(\mathbf{w}, b) = \sum_{\ell=1}^{N\alpha} \epsilon(\mathbf{w}, b; \mathbf{S}^\ell)$, to denote the dependence of term ℓ in the sum on the student

weights and on the sample ℓ. The quantity $\mathrm{d}\mu(\mathbf{w})$ denotes the integration measure over the student weights and it enforces the spherical constraint:

$$\mathrm{d}\mu(\mathbf{w}) = \prod_{i=1}^{N} \frac{\mathrm{d}w_i}{\sqrt{2\pi e}} \delta(\mathbf{w}\cdot\mathbf{w} - N)\,. \tag{29}$$

The differential $\mathrm{d}\mu(b) = \mathrm{d}b$ is the integration measure over the student bias.

We recall the shape of the training data distribution,[2]

$$\mathrm{d}\mu_{\text{train}}(\{\mathbf{S}^\ell\}) = \frac{1}{\mathcal{N}_{\text{bias}}}\binom{N\alpha}{N\alpha\rho}\left(\prod_{\ell=1}^{\alpha N\rho} D\mathbf{S}^\ell \Theta\left(\frac{\mathbf{w}^0\cdot\mathbf{S}^\ell}{\sqrt{N}} + b_0\right)\right)\left(\prod_{\ell'=\alpha N\rho+1}^{\alpha N} D\mathbf{S}^{\ell'} \Theta\left(-\frac{\mathbf{w}^0\cdot\mathbf{S}^{\ell'}}{\sqrt{N}} - b_0\right)\right), \tag{30}$$

with $\mathcal{N}_{\text{bias}} = \binom{N\alpha}{N\alpha\rho}\left(\frac{1}{2}\mathrm{erfc}\left(\frac{-b_0}{\sqrt{2}}\right)\right)^{\alpha N\rho}\left(1-\frac{1}{2}\mathrm{erfc}\left(\frac{-b_0}{\sqrt{2}}\right)\right)^{\alpha N(1-\rho)} = \binom{N\alpha}{N\alpha\rho} c_+^{\alpha N\rho} c_-^{\alpha N(1-\rho)}$ normalization constant.

The core of the Replica Approach [Mezard et al., 1987] [Seung et al., 1992] lies in exploiting the identity

$$\langle\!\langle \log Z \rangle\!\rangle = \lim_{n\to 0} \frac{1}{n} \log \langle\!\langle Z^n \rangle\!\rangle\,, \tag{31}$$

where n is promoted to an integer and Z^n is computed by replicating the partition function n times (*i.e.* considering n independent copies of the original system). Finally one takes the limit $n \to 0$ to recover $\langle\!\langle \log Z \rangle\!\rangle$. This procedure provides exact asympotic results, and was widely used both in machine learning and in several other disordered systems [Charbonneau et al., 2023].

We begin with computing

$$\langle\!\langle Z^n \rangle\!\rangle = \int \mathrm{d}\mu_{\text{train}}(\{\mathbf{S}^\ell\}) \int \left(\prod_{\sigma=1}^{n} \mathrm{d}\mu(b_\sigma)\,\mathrm{d}\mu(\mathbf{w}^\sigma)\right) e^{-\beta\sum_\sigma\sum_\ell \epsilon(\mathbf{w}^\sigma, b_\sigma; \mathbf{S}^\ell)}\,. \tag{32}$$

By expanding the integration measure over the training samples, and collecting respectively positive and negative terms we get,

$$\begin{aligned}\langle\!\langle Z^n \rangle\!\rangle = \int \prod_{\sigma=1}^{n} \mathrm{d}\mu(b_\sigma)\,\mathrm{d}\mu(\mathbf{w}^\sigma) \left[\frac{1}{c_+}\int D\mathbf{S}\Theta\left(\frac{\mathbf{w}^0\cdot\mathbf{S}}{\sqrt{N}} + b_0\right) e^{-\beta\sum_\sigma \epsilon(\mathbf{w}^\sigma, b_\sigma; \mathbf{S})}\right]^{N\alpha\rho} \cdot \\ \cdot \left[\frac{1}{c_-}\int D\mathbf{S}\Theta\left(-\frac{\mathbf{w}^0\cdot\mathbf{S}}{\sqrt{N}} - b_0\right) e^{-\beta\sum_\sigma \epsilon(\mathbf{w}^\sigma, b_\sigma; \mathbf{S})}\right]^{N\alpha(1-\rho)}\,.\end{aligned} \tag{33}$$

Now we define

$$\mathcal{G}_r^{\pm}(\{\mathbf{w}^\sigma, b_\sigma\}) = -\log \frac{1}{c_\pm}\int D\mathbf{S}\Theta\left(\pm\frac{\mathbf{w}^0\cdot\mathbf{S}}{\sqrt{N}} \pm b_0\right) e^{-\beta\sum_\sigma \epsilon(\mathbf{w}^\sigma, b_\sigma; \mathbf{S})}\,, \tag{34}$$

where $\{\mathbf{w}^\sigma, b_\sigma\}$ denotes the dependence of $\mathcal{G}_r^\pm$ on the whole set of n replicated students. Then one can rewrite the replicated partition function as:

$$\langle\!\langle Z^n \rangle\!\rangle = \int \prod_{\sigma=1}^{n} \mathrm{d}\mu(b_\sigma)\,\mathrm{d}\mu(\mathbf{w}^\sigma) e^{-N\alpha\rho\mathcal{G}_r^+(\{\mathbf{w}^\sigma, b_\sigma\}) - N\alpha(1-\rho)\mathcal{G}_r^-(\{\mathbf{w}^\sigma, b_\sigma\})}\,. \tag{35}$$

We introduce the auxiliary variables x_σ and y in order to remove $\mathbf{S}$ from the argument of g function in the expression of $\mathcal{G}_r^\pm$:

$$e^{-\mathcal{G}_r^+} = \frac{1}{c_+}\int \prod_{\sigma=1}^{n} \mathrm{d}x_\sigma \int \mathrm{d}y\ \Theta(y+b_0) e^{-\frac{\beta}{2}\sum_\sigma [g(x_\sigma+b_\sigma) - g(y+b_0)]^2} \int D\mathbf{S} \prod_{\sigma=1}^{n} \delta\left(x_\sigma - \frac{\mathbf{w}^\sigma\cdot\mathbf{S}}{\sqrt{N}}\right)\delta\left(y - \frac{\mathbf{w}^0\cdot\mathbf{S}}{\sqrt{N}}\right) \tag{36}$$

$$= \frac{1}{c_+}\int \prod_{\sigma=1}^{n} \frac{\mathrm{d}x_\sigma\,\mathrm{d}\hat{x}_\sigma}{2\pi} \int \frac{\mathrm{d}y\,\mathrm{d}\hat{y}}{2\pi}\Theta(y+b_0) e^{\left(-\frac{\beta}{2}\sum_\sigma [g(x_\sigma+b_\sigma) - g(y+b_0)]^2 + i\sum_\sigma x_\sigma\hat{x}_\sigma + iy\hat{y}\right)} \int D\mathbf{S} e^{-i\left(\sum_\sigma \mathbf{w}^\sigma\hat{x}_\sigma + \mathbf{w}^0\hat{y}\right)\cdot\frac{\mathbf{S}}{\sqrt{N}}}\,, \tag{37}$$

[2] **Note**: To make the notation lighter, we abbreviate ρ_{train} as ρ for the whole derivation, since we are now only focusing on the training set.

where we exploited the Fourier representation of the δ-function. The last integral in Eq. (37) is Gaussian, and it yields: $e^{-\frac{1}{2N}(\sum_\sigma \mathbf{w}^\sigma \hat{x}_\sigma + \mathbf{w}^0 \hat{y})^2}$. The function $\mathcal{G}_r^+$ depends on the vectors $\mathbf{w}^\sigma, \mathbf{w}^0$ only through the overlaps

$$Q_{\sigma\sigma'} = \frac{\mathbf{w}^\sigma \cdot \mathbf{w}^{\sigma'}}{N}, \quad R_\sigma = \frac{\mathbf{w}^\sigma \cdot \mathbf{w}^0}{N}, \tag{38}$$

which are emergent order parameters of the theory. In terms of these functions, $\mathcal{G}_r^+$ can be written as

$$e^{-\mathcal{G}_r^+} = \frac{1}{c_+} \int \prod_{\sigma=1}^n \frac{\mathrm{d}x_\sigma \, \mathrm{d}\hat{x}_\sigma}{2\pi} \int \frac{\mathrm{d}y \, \mathrm{d}\hat{y}}{2\pi} \Theta(y + b_0) e^{-\frac{\beta}{2}\sum_\sigma [g(x_\sigma + b_\sigma) - g(y + b_0)]^2 + i \sum_\sigma x_\sigma \hat{x}_\sigma + iy\hat{y}} e^{-\frac{1}{2}\sum_{\sigma,\sigma'} \hat{x}_\sigma \hat{x}_{\sigma'} Q_{\sigma\sigma'} - \hat{y}\sum_\sigma \hat{x}_\sigma R_\sigma - \frac{1}{2}\hat{y}^2}, \tag{39}$$

and similarly

$$e^{-\mathcal{G}_r^-} = \frac{1}{c_-} \int \prod_{\sigma=1}^n \frac{\mathrm{d}x_\sigma \, \mathrm{d}\hat{x}_\sigma}{2\pi} \int \frac{\mathrm{d}y \, \mathrm{d}\hat{y}}{2\pi} \Theta(-y - b_0) e^{-\frac{\beta}{2}\sum_\sigma [g(x_\sigma + b_\sigma) - g(y + b_0)]^2 + i \sum_\sigma x_\sigma \hat{x}_\sigma + iy\hat{y}} e^{-\frac{1}{2}\sum_{\sigma,\sigma'} \hat{x}_\sigma \hat{x}_{\sigma'} Q_{\sigma\sigma'} - \hat{y}\sum_\sigma \hat{x}_\sigma R_\sigma - \frac{1}{2}\hat{y}^2}. \tag{40}$$

The replicated partition function can thus be written as an integral over the order parameters:

$$\begin{aligned} \langle\!\langle Z^n \rangle\!\rangle &= \int \prod_{\sigma=1}^n \mathrm{d}\mu(b_\sigma) \, \mathrm{d}\mu(\mathbf{w}^\sigma) e^{-N\alpha\rho\mathcal{G}_r^+(\{\mathbf{w}^\sigma, b_\sigma\}) - N\alpha(1-\rho)\mathcal{G}_r^-(\{\mathbf{w}^\sigma, b_\sigma\})} \times \\ &\quad \times \int \prod_{\sigma>\sigma'} \mathrm{d}Q_{\sigma\sigma'} \int \prod_\sigma \mathrm{d}R\sigma \prod_{\sigma>\sigma'} \delta(\mathbf{w}^\sigma \cdot \mathbf{w}^{\sigma'} - NQ_{\sigma\sigma'}) \prod_\sigma \delta(\mathbf{w}^\sigma \cdot \mathbf{w}^0 - NR_\sigma) \qquad (41) \\ &= \int \prod_\sigma \mathrm{d}b_\sigma \int \prod_{\sigma>\sigma'} \frac{\mathrm{d}Q_{\sigma\sigma'} \, \mathrm{d}\hat{Q}_{\sigma\sigma'}}{2\pi i} \int \prod_\sigma \frac{\mathrm{d}R_\sigma \, \mathrm{d}\hat{R}_\sigma}{2\pi i} e^{-N\alpha\rho\mathcal{G}_r^+(\{Q_{\sigma\sigma'}, R_\sigma, b_\sigma\}) - N\alpha(1-\rho)\mathcal{G}_r^-(\{Q_{\sigma\sigma'}, R_\sigma, b_\sigma\})} \times \\ &\quad \times e^{N(-\sum_{\sigma>\sigma'} Q_{\sigma\sigma'}\hat{Q}_{\sigma\sigma'} - \sum_\sigma \hat{R}_\sigma R_\sigma)} \int \prod_{\sigma=1}^n \mathrm{d}\mu(\mathbf{w}^\sigma) e^{\sum_{\sigma>\sigma'} \hat{Q}_{\sigma\sigma'} \mathbf{w}^\sigma \mathbf{w}^{\sigma'} + \sum_\sigma \hat{R}_\sigma \mathbf{w}^\sigma \mathbf{w}^0}, \qquad (42) \end{aligned}$$

with $\hat{Q}_{\sigma\sigma'}$ and $\hat{R}_\sigma$ conjugates of the overlaps. Then, the replicated partition function can be rewritten as

$$\langle\!\langle Z^n \rangle\!\rangle = \int \prod_\sigma \mathrm{d}b_\sigma \int \prod_{\sigma>\sigma'} \frac{\mathrm{d}Q_{\sigma\sigma'} \, \mathrm{d}\hat{Q}_{\sigma\sigma'}}{2\pi i} \int \prod_\sigma \frac{\mathrm{d}R_\sigma \, \mathrm{d}\hat{R}_\sigma}{2\pi i} e^{-N\mathcal{A}_r(\{Q_{\sigma\sigma'}, \hat{Q}_{\sigma\sigma'}, R_\sigma, \hat{R}_\sigma\})}, \tag{43}$$

where we defined

$$\mathcal{A}_r(\{Q_{\sigma\sigma'}, \hat{Q}_{\sigma\sigma'}, R_\sigma, \hat{R}_\sigma\}) = \alpha\rho\mathcal{G}_r^+(\{Q_{\sigma\sigma'}, R_\sigma, b_\sigma\}) + \alpha(1-\rho)\mathcal{G}_r^-(\{Q_{\sigma\sigma'}, R_\sigma, b_\sigma\}) - \mathcal{G}_0(\{Q_{\sigma\sigma'}, \hat{Q}_{\sigma\sigma'}, R_\sigma, \hat{R}_\sigma\}), \tag{44}$$

and

$$\mathcal{G}_0(\{Q_{\sigma\sigma'}, \hat{Q}_{\sigma\sigma'}, R_\sigma, \hat{R}_\sigma\}) = -\sum_{\sigma>\sigma'} Q_{\sigma\sigma'}\hat{Q}_{\sigma\sigma'} - \sum_\sigma \hat{R}_\sigma R_\sigma + \frac{1}{N} \log \int \prod_{\sigma=1}^n \mathrm{d}\mu(\mathbf{w}^\sigma) e^{\sum_{\sigma>\sigma'} \hat{Q}_{\sigma\sigma'} \mathbf{w}^\sigma \mathbf{w}^{\sigma'} + \sum_\sigma \hat{R}_\sigma \mathbf{w}^\sigma \mathbf{w}^0}. \tag{45}$$

Replica Symmetric Ansatz According to Eq. (31), the Free Energy density in the thermodynamic limit ($N \to \infty$) reads:

$$-\beta\mathcal{F} = \lim_{N\to\infty} \lim_{n\to 0} \frac{1}{nN} \log \langle\!\langle Z^n \rangle\!\rangle. \tag{46}$$

In order to evaluate the integral in Eq. (43) by saddle-point we switch the order of the two limits as prescribed within replica theory, getting,

$$-\beta\mathcal{F} = \lim_{n\to 0} \frac{1}{n} \min_{\substack{Q_{\sigma\sigma'}, R_\sigma, \\ \hat{Q}_{\sigma\sigma'}, \hat{R}_\sigma, b_\sigma}} \left\{ \mathcal{G}_0(\{Q_{\sigma\sigma'}, \hat{Q}_{\sigma\sigma'}, R_\sigma, \hat{R}_\sigma\}) - \alpha\rho\mathcal{G}_r^+(\{Q_{\sigma\sigma'}, R_\sigma, b_\sigma\}) - \alpha(1-\rho)\mathcal{G}_r^-(\{Q_{\sigma\sigma'}, R_\sigma, b_\sigma\}) \right\}. \tag{47}$$

To carry on the computation one has to find a parametrization of the order parameters and express Eq. (47) as a function of the elements of these multi-dimensional arrays and the number of replicas n. We adopt the Replica

Symmetric (RS) ansatz [Mézard et al., 1987], where one assumes that the replicated students are symmetric, *i.e.* they all have the same overlap with the teacher and among them:

$$Q_{\sigma\sigma'} = \delta_{\sigma\sigma'} + (1-\delta_{\sigma\sigma'})q\,, \tag{48}$$
$$\hat{Q}_{\sigma\sigma'} = \delta_{\sigma\sigma'} + (1-\delta_{\sigma\sigma'})\hat{q}\,, \tag{49}$$
$$R_\sigma = R\,, \tag{50}$$
$$\hat{R}_\sigma = \hat{R}\,, \tag{51}$$
$$b_\sigma = b\,. \tag{52}$$

We stress that in our computation the student bias b is treated as an order parameter and its value is fixed by saddle point as for the other parameters. We now define:

$$G_r^\pm = \lim_{n\to 0} \frac{\mathcal{G}_r^\pm}{n}, \qquad G_0 = \lim_{n\to 0} \frac{\mathcal{G}_0}{n}\,. \tag{53}$$

At this point, the optimization problem to solve in order to find the equilibrium values of the order parameters becomes the following:

$$-\beta\mathcal{F} = \min_{\substack{q,\hat{R},\\ \hat{q},\hat{R},b}} \left\{ G_0(q,\hat{q},R,\hat{R}) - \alpha\rho G_r^+(q,R,b) - \alpha(1-\rho)G_r^-(q,R,b) \right\}. \tag{54}$$

We will refer to $G_r^\pm$ as the energetic terms as they represent the energetic contribution to the free-energy of positive and negative class samples. G_0 is the entropic or volume term and quantifies the number of student configurations that correspond to a given choice of the order parameters.

In the following, we show the detailed computations for G_r^+, as the derivation for G_r^- follows the same lines. The plan is to substitute the RS ansatz in the expression of $\mathcal{G}_r^+$ and integrate over the conjugate variables $\hat{y}, \hat{x}_\sigma$. The integral in $\hat{y}$ is a Gaussian integral and yields:

$$e^{-\mathcal{G}_r^+} = \frac{1}{c_+}\int \prod_{\sigma=1}^n \frac{\mathrm{d}x_\sigma\,\mathrm{d}\hat{x}_\sigma}{2\pi} \int \frac{\mathrm{d}y}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}\Theta(y+b_0) \\ e^{-\frac{\beta}{2}\sum_\sigma [g(x_\sigma+b)-g(y+b_0)]^2} e^{-\frac{1}{2}(1-q)\sum_\sigma \hat{x}_\sigma^2 + \frac{1}{2}(R^2-q)\sum_{\sigma,\sigma'}\hat{x}_\sigma\hat{x}_{\sigma'} + i\sum_\sigma \hat{x}_\sigma(x_\sigma - yR)}\,. \tag{55}$$

In order to integrate out the $\hat{x}_\sigma$ we need to decouple the term $\hat{x}_\sigma\hat{x}_{\sigma'}$ through Hubbard-Stratonovich transform,

$$e^{-\frac{1}{2}(q-R^2)\sum_{\sigma,\sigma'}\hat{x}_\sigma\hat{x}_{\sigma'}} = \int Dt e^{(i\sqrt{q-R^2}\sum_\sigma \hat{x}_\sigma)t}\,, \tag{56}$$

where we recall that $Dt = \frac{\mathrm{d}t}{\sqrt{2\pi}}e^{-t^2/2}$.

Integrals over $\hat{x}_\sigma$ are now Gaussian and yield:

$$e^{-\mathcal{G}_r^+} = \frac{1}{c_+}\int Dy \int Dt\Theta(y+b_0)\prod_{\sigma=1}^n \int \frac{\mathrm{d}x_\sigma}{\sqrt{2\pi(1-q)}} e^{-\frac{(x_\sigma - yR + t\sqrt{q-R^2})^2}{2(1-q)}} e^{-\frac{\beta}{2}[g(x_\sigma+b)-g(y+b_0)]^2}\,. \tag{57}$$

Performing the shift and re-scaling $x_\sigma \to x_\sigma\sqrt{1-q} + yR - t\sqrt{q-R^2}$ one gets:

$$e^{-\mathcal{G}_r^+} = \frac{1}{c_+}\int Dy\Theta(y+b_0)\int Dt\left[\int Dx e^{-\frac{\beta}{2}\left[g(x\sqrt{1-q}+yR-t\sqrt{q-R^2}+b)-g(y+b_0)\right]^2}\right]^n\,. \tag{58}$$

Now we can compute the $n\to 0$ limit. We call $A = \int Dx e^{-\frac{\beta}{2}\left[g(x\sqrt{1-q}+yR-t\sqrt{q-R^2}+b)-g(y+b_0)\right]^2}$, and exploit the identity $A^n \sim 1 + n\log A$, which is valid in the $n\to 0$ limit:

$$G_r^+ = \lim_{n\to 0} -\frac{1}{n}\log\left(1 + n\frac{1}{c_+}\int Dy\Theta(y+b_0)\int Dt\log A\right). \tag{59}$$

Since n is small, we can expand the first logarithm around 1:

$$G_r^+ = -\frac{1}{c_+}\int Dy\Theta(y+b_0)\int Dt\log A\,. \tag{60}$$

We now recall our choice of the activation function $g(x) = \text{sign}(x)$. The square-loss per sample in this case reads $\epsilon(\mathbf{w};\mathbf{S}) = 2\Theta(-(\mathbf{w}\cdot\mathbf{S}+b)(\mathbf{w}^0\cdot\mathbf{S}+b_0))$. We re-define it without the the factor 2 in order to count the number of mis-classified samples. G_r^+ becomes

$$G_r^+ = -\frac{1}{c_+}\int_{-b_0}^{\infty} Dy\int_{-\infty}^{\infty} Dt\log\left(\int_{-\infty}^{\infty} Dx e^{-\beta\Theta\left(-(x\sqrt{1-q}+yR-t\sqrt{q-R^2}+b)(y+b_0)\right)}\right). \tag{61}$$

We define $u \equiv \frac{t\sqrt{q-R^2}-yR-b}{\sqrt{1-q}}$ and $H(x) \equiv \int_x^\infty Dt = \frac{1}{2}\text{erfc}\left(\frac{x}{\sqrt{2}}\right)$. The final form for $G_r^\pm$ reads

$$G_r^+ = -\frac{1}{c_+}\int_{-b_0}^{\infty} Dy\int_{-\infty}^{\infty} Dt\log\left(e^{-\beta}+(1-e^{-\beta})H(u)\right), \tag{62}$$

$$G_r^- = -\frac{1}{c_-}\int_{-\infty}^{-b_0} Dy\int_{-\infty}^{\infty} Dt\log\left((e^{-\beta}-1)H(u)+1\right). \tag{63}$$

We now show the detailed computation for the entropic term G_0. Starting from Eq. (45), and substituting the RS ansatz one gets

$$\mathcal{G}_0 = -\frac{1}{2}n(n-1)q\hat{q} - n\hat{R}R + \frac{1}{N}\log\int\prod_{\sigma=1}^{n}\mathrm{d}\mu(\mathbf{w}^\sigma)e^{\hat{q}\sum_{\sigma>\sigma'}\mathbf{w}^\sigma\mathbf{w}^{\sigma'}+\hat{R}\sum_\sigma\mathbf{w}^\sigma\mathbf{w}^0}. \tag{64}$$

We decouple $\mathbf{w}^\sigma\mathbf{w}^{\sigma'}$ through Hubbard-Stratonovich:

$$e^{\hat{q}\sum_{\sigma>\sigma'}\mathbf{w}^\sigma\mathbf{w}^{\sigma'}} = e^{\frac{1}{2}\hat{q}\sum_{\sigma,\sigma'}\mathbf{w}^\sigma\mathbf{w}^{\sigma'}-\frac{1}{2}\hat{q}\sum_\sigma\mathbf{w}^\sigma\mathbf{w}^\sigma} = e^{-\frac{1}{2}\hat{q}\sum_\sigma\mathbf{w}^\sigma\mathbf{w}^\sigma}\int D\mathbf{z}e^{\sqrt{\hat{q}}\sum_\sigma\mathbf{w}^\sigma\mathbf{z}}. \tag{65}$$

This allows us to rewrite $\mathcal{G}_0$ as

$$\mathcal{G}_0 = -\frac{1}{2}n(n-1)q\hat{q} - n\hat{R}R + \frac{1}{N}\log\int D\mathbf{z}\left(\int\mathrm{d}\mu(\mathbf{w})e^{\mathbf{w}(\hat{R}\mathbf{w}^0+\sqrt{\hat{q}}\mathbf{z}-\frac{1}{2}\hat{q}\mathbf{w})}\right)^n. \tag{66}$$

Now we can take the $n\to 0$ limit:

$$G_0 = \lim_{n\to 0}\frac{\mathcal{G}_0}{n} = -\hat{R}R + \frac{1}{2}q\hat{q} + \frac{1}{N}\int D\mathbf{z}\log\int\mathrm{d}\mu(\mathbf{w})e^{\mathbf{w}(\hat{R}\mathbf{w}^0+\sqrt{\hat{q}}\mathbf{z}-\frac{1}{2}\hat{q}\mathbf{w})}. \tag{67}$$

The last integral on Eq. (67) is equal to:

$$\int D\mathbf{z}\log\int\frac{\mathrm{d}\lambda}{4\pi i}e^{\frac{N\lambda}{2}}e^{-\frac{N}{2}\log[e(\lambda+\hat{q})]}e^{\frac{N}{2(\lambda+\hat{q})}(\hat{R}^2+\hat{q}\frac{\sum_i z_i^2}{N}+2\sqrt{\hat{q}}\hat{R}\frac{\sum_i\omega_i^0 z_i}{N})}. \tag{68}$$

Computing the integral over λ with a saddle point approximation we reduce the double integral to

$$N\left[\frac{\lambda}{2}-\frac{1}{2}\log[e(\lambda+\hat{q})]+\frac{\hat{R}^2}{2(\lambda+\hat{q})}+\frac{1}{2(\lambda+\hat{q})}\int D\mathbf{z}\left(\hat{q}\frac{\sum_i z_i^2}{N}+2\sqrt{\hat{q}}\hat{R}\frac{\sum_i\omega_i^0 z_i}{N}\right)\right], \tag{69}$$

where now λ denotes the saddle point value. The Gaussian integral over $\mathbf{z}$ is easily computed, and one gets as a result:

$$G_0 = -\hat{R}R + \frac{1}{2}q\hat{q} + \frac{\lambda}{2} - \frac{1}{2}\log(\lambda+\hat{q}) + \frac{1}{2}\frac{\hat{R}^2+\hat{q}}{(\lambda+\hat{q})} - \frac{1}{2}. \tag{70}$$

Saddle Point Equations We look for a stationary point of the variational Free-Energy to fix the value of the order parameters at equilibrium. We then set to 0 the derivatives of the variational Free-Energy with respect to the order parameters $(q, \hat{q}, R, \hat{R}, b)$ and the additional Lagrange multiplier λ that we introduced to enforce the spherical constraint for the students' weights. By doing so, we obtain the following system of coupled equations:

$$R = \hat{R}(1-q)\,, \tag{71}$$

$$q = (\hat{q} + \hat{R}^2)(1-q)^2\,, \tag{72}$$

$$\hat{R} = \alpha \frac{e^{-\frac{b_0^2 q}{2(q-R^2)}}}{2\pi\sqrt{1-q}} \left\{ \frac{\rho}{c_+} \int_{-\infty}^{\infty} Dt \frac{e^{-v^2/2 + \frac{b_0 R}{\sqrt{q-R^2}} t}}{(e^{\beta}-1)^{-1} + H(v)} - \frac{1-\rho}{c_-} \int_{-\infty}^{\infty} Dt \frac{e^{-v^2/2 + \frac{b_0 R}{\sqrt{q-R^2}} t}}{(e^{-\beta}-1)^{-1} + H(v)} \right\}, \tag{73}$$

$$\hat{q} = \frac{\alpha}{2\pi(1-q)} \left\{ \frac{\rho}{c_+} \int_{-b_0}^{\infty} Dy \int_{-\infty}^{\infty} Dt \frac{e^{-u^2}}{[(e^{\beta}-1)^{-1} + H(u)]^2} + \frac{1-\rho}{c_-} \int_{-\infty}^{-b_0} Dy \int_{-\infty}^{\infty} Dt \frac{e^{-u^2}}{[(e^{-\beta}-1)^{-1} + H(u)]^2} \right\}, \tag{74}$$

$$0 = \frac{\rho}{c_+} \int_{-b_0}^{\infty} Dy \int_{-\infty}^{\infty} Dt \frac{e^{-u^2/2}}{(e^{\beta}-1)^{-1} + H(u)} + \frac{1-\rho}{c_-} \int_{-\infty}^{-b_0} Dy \int_{-\infty}^{\infty} Dt \frac{e^{-u^2/2}}{(e^{-\beta}-1)^{-1} + H(u)}\,. \tag{75}$$

In the following and in the main manuscript, when we talk of self-consistently solving the saddle-point equations, we refer to solving Eqs. (71–75).

D TRAIN AND GENERALIZATION METRICS

Analytical expression for the Metrics Train and Generalization metrics can be expressed in terms of order parameters evaluated at equilibrium (solutions of the saddle-point equations). Here we derive their expressions.

The average **train error** per sample can be evaluated as:

$$\epsilon_t = \frac{1}{P} \langle\!\langle \mathbb{E}_T[\mathcal{E}(\mathbf{w}, b)] \rangle\!\rangle_{\mu_{\text{train}}}\,, \tag{76}$$

where $\mathbb{E}_T[\dots]$ denotes the average over the realizations of the thermal noise. Explicitly, one has

$$\epsilon_t = \frac{1}{N\alpha} \langle\!\langle \frac{1}{Z} \int \mathrm{d}\mu(\mathbf{w}) \int \mathrm{d}\mu(b) \mathcal{E}(\mathbf{w}, b) e^{-\beta \mathcal{E}(\mathbf{w}, b)} \rangle\!\rangle_{\mu_{\text{train}}} = -\frac{1}{N\alpha} \langle\!\langle \frac{\partial}{\partial \beta} \log Z \rangle\!\rangle = \frac{1}{N\alpha} \frac{\partial(\beta F)}{\partial \beta}\,, \tag{77}$$

$$\epsilon_t = -\frac{1}{\alpha} \frac{\partial}{\partial \beta} \left\{ G_0(q, \hat{q}, R, \hat{R}) - \alpha \rho_{\text{train}} G_r^+(q, R, b) - \alpha(1-\rho_{\text{train}}) G_r^-(q, R, b) \right\}\,, \tag{78}$$

evaluated at the saddle point. The volume term G_0 does not depend on the temperature. We get

$$\epsilon_t = \rho_{\text{train}} \frac{\partial G_r^+(q, R, b)}{\partial \beta} + (1-\rho_{\text{train}}) \frac{\partial G_r^-(q, R, b)}{\partial \beta} \tag{79}$$

$$= \frac{\rho_{\text{train}}}{c_+} \int_{-b_0}^{\infty} Dy \int_{-\infty}^{\infty} Dt \frac{1 - H(u)}{1 + (e^{\beta}-1) H(u)} + \frac{1-\rho_{\text{train}}}{c_-} \int_{-\infty}^{-b_0} Dy \int_{-\infty}^{\infty} Dt \frac{H(u)}{e^{\beta} + (1-e^{\beta}) H(u)}\,. \tag{80}$$

To compute the generalization metrics we introduce the test-set distribution

$$\mathrm{d}\mu_{\text{test}}(\mathbf{S}) = \frac{\rho_{\text{test}}}{c_+} \Theta\left(\frac{\mathbf{w}^0 \cdot \mathbf{S}}{\sqrt{N}} + b_0 \right) D\mathbf{S} + \frac{1-\rho_{\text{test}}}{c_-} \Theta\left(-\frac{\mathbf{w}^0 \cdot \mathbf{S}}{\sqrt{N}} - b_0 \right) D\mathbf{S}\,, \tag{81}$$

where ρ_{test} determines the probability of having a positive sample in the test-set. The idea behind the computation of generalization metrics is to add one sample that was not observed during training and evaluate the performance of the trained student on it. In practice, we can define the **generalization error** as

$$\epsilon_g = \langle\!\langle \langle\!\langle \mathbb{E}_T[\epsilon(\mathbf{w}, b; \mathbf{S})] \rangle\!\rangle_{\mu_{\text{train}}} \rangle\!\rangle_{\mu_{\text{test}}}\,. \tag{82}$$

The first average is on the train-set and it yields the saddle-point equations shown in the previous section. The second average, on the test-set is needed to evaluate the trained student on the new, unseen sample. Explicitly, we get

$$\epsilon_g = \langle\!\langle \frac{1}{Z}\int \mathrm{d}\mu(\mathbf{w}) \int \mathrm{d}\mu(b)\epsilon(\mathbf{w},b;\mathbf{S})e^{-\beta\mathcal{E}(\mathbf{w},b)}\rangle\!\rangle_{\mu_{\text{train}}}\rangle_{\mu_{\text{test}}} \tag{83}$$

$$= \lim_{n\to 0}\langle\!\langle Z^{n-1}\int \mathrm{d}\mu(\mathbf{w}) \int \mathrm{d}\mu(b)\epsilon(\mathbf{w},b;\mathbf{S})e^{-\beta\sum_{\ell=1}^{N\alpha}\epsilon(\mathbf{w},b;\mathbf{S}^\ell)}\rangle\!\rangle_{\mu_{\text{train}}}\rangle_{\mu_{\text{test}}} \tag{84}$$

$$= \lim_{n\to 0}\int \mathrm{d}\mu_{\text{test}}(\mathbf{S}) \int \mathrm{d}\mu_{\text{bias}}(\{\mathbf{S}^\ell\}) \int \prod_{\sigma=1}^{n}\mathrm{d}\mu(b_\sigma)\,\mathrm{d}\mu(\mathbf{w}^\sigma)\epsilon(\mathbf{w}^1,b_1;\mathbf{S})e^{-\beta\sum_\ell\sum_\sigma\epsilon(\mathbf{w}^\sigma,b_\sigma;\mathbf{S}^\ell)} \tag{85}$$

$$= \lim_{n\to 0}\int \mathrm{d}\mu_{\text{test}}(\mathbf{S}) \int \prod_{\sigma=1}^{n}\mathrm{d}\mu(b_\sigma)\,\mathrm{d}\mu(\mathbf{w}^\sigma)\epsilon(\mathbf{w}^1,b_1;\mathbf{S})e^{-N\alpha\rho\mathcal{G}_r^+(\{\mathbf{w}^\sigma,b_\sigma\})-N\alpha(1-\rho)\mathcal{G}_r^-(\{\mathbf{w}^\sigma,b_\sigma\})}\,. \tag{86}$$

One can evaluate:

$$\int \mathrm{d}\mu_{\text{test}}(\mathbf{S})\epsilon(\mathbf{w}^1,b_1;\mathbf{S}) = \frac{\rho_{\text{test}}}{c_+}\mathrm{I}_+ + \frac{1-\rho_{\text{test}}}{c_-}\mathrm{I}_-\,, \tag{87}$$

where we have defined:

$$\mathrm{I}_\pm = \int D\mathbf{S}\Theta\left(\pm\frac{\mathbf{w}^0\cdot\mathbf{S}}{\sqrt{N}}\pm b_0\right)\epsilon(\mathbf{w}^1,b_1;\mathbf{S})\,. \tag{88}$$

In the following we show the computation for I_+, the one for I_- follows the same lines:

$$\mathrm{I}_+ = \int D\mathbf{S}\Theta\left(\frac{\mathbf{w}^0\cdot\mathbf{S}}{\sqrt{N}}+b_0\right)\epsilon(\mathbf{w}^1,b_1;\mathbf{S}) \tag{89}$$

$$= \int\frac{\mathrm{d}x\,\mathrm{d}\hat{x}}{2\pi}\int\frac{\mathrm{d}y\,\mathrm{d}\hat{y}}{2\pi}e^{ix\hat{x}+iy\hat{y}}\frac{1}{2}[g(x+b_1)-g(y+b_0)]^2\Theta(y+b)\int D\mathbf{S}e^{-\frac{i}{\sqrt{N}}(\mathbf{w}^1\hat{x}+\mathbf{w}^0\hat{y})\cdot\mathbf{S}}\,. \tag{90}$$

Integrating over $\mathbf{S},\hat{x},\hat{y}$ one gets

$$\mathrm{I}_+ = \int Dx\int_{-b_0}^{\infty}Dy\frac{1}{2}[g(x\sqrt{1-R_1^2}+yR_1+b_1)-g(y+b_0)]^2 = \tag{91}$$

$$= \int Dx\int_{-b_0}^{\infty}Dy\Theta(-(x\sqrt{1-R_1^2}+yR_1+b_1)(y+b_0)) = \tag{92}$$

$$= \int_{-b_0}^{\infty}Dy\int_{-\infty}^{u'}Dx = \int_{-b_0}^{\infty}Dy(1-H(u'))\,, \tag{93}$$

with $u' = \frac{-yR_1-b_1}{\sqrt{1-R_1^2}}$.

Computing also I_- we get:

$$\epsilon(R_1,b_1) = \frac{\rho_{\text{test}}}{\frac{1}{2}\mathrm{erfc}\left(\frac{-b_0}{\sqrt{2}}\right)}\int_{-b_0}^{\infty}Dy\int_{-\infty}^{u'}Dx + \frac{1-\rho_{\text{test}}}{1-\frac{1}{2}\mathrm{erfc}\left(\frac{-b_0}{\sqrt{2}}\right)}\int_{-\infty}^{-b_0}Dy\int_{u'}^{\infty}Dx\,. \tag{94}$$

Thus we can rewrite the generalization error:

$$\epsilon_g = \lim_{n\to 0}\int\prod_\sigma \mathrm{d}b_\sigma\int\prod_{\sigma>\sigma'}\frac{\mathrm{d}Q_{\sigma,\sigma'}\,\mathrm{d}\hat{Q}_{\sigma,\sigma'}}{2\pi i}\int\prod_\sigma\frac{\mathrm{d}R_\sigma\,\mathrm{d}\hat{R}_\sigma}{2\pi i}\times \tag{95}$$

$$\times\,\epsilon(R_1,b_1)e^{-N\alpha\rho\mathcal{G}_r^+(\{Q_{\sigma,\sigma'},R_\sigma,b_\sigma\})-N\alpha(1-\rho)\mathcal{G}_r^-(\{Q_{\sigma,\sigma'},R_\sigma,b_\sigma\})+N\mathcal{G}_0(\{Q_{\sigma,\sigma'},\hat{Q}_{\sigma,\sigma'},R_\sigma,\hat{R}_\sigma\})}\,. \tag{96}$$

Following the same lines of the Replica Calculation performed in Sec. C, one gets

$$\epsilon_g = \epsilon(R,b), \tag{97}$$

with (R, b) parameters at equilibrium *i.e.* the ones that solve the saddle-point equations.

As introduced in the main manuscript, all the **generalization metrics** that we investigate in the manuscript can be expressed in terms of **True Positive Rate** (Recall, r) and the **True Negative Rate** (Specificity, s). Here we show the derivation for these two metrics:

$$r = \frac{\left\langle\!\left\langle \mathbb{E}_T\left[\left[1-\Theta\left(-\left(\frac{\mathbf{w}\cdot\mathbf{S}}{\sqrt{N}}+b\right)\left(\frac{\mathbf{w}^0\cdot\mathbf{S}}{\sqrt{N}}+b_0\right)\right)\right]\Theta\left(\frac{\mathbf{w}^0\cdot\mathbf{S}}{\sqrt{N}}+b_0\right)\right]\right\rangle\!\right\rangle_{\mu_{\text{train}}}\Big\rangle_{\mu_{\text{test}}}}{\rho_{test}} \tag{98}$$

$$= \frac{1}{c_+}\int D\mathbf{S}\Theta\left(\frac{\mathbf{w}^0\cdot\mathbf{S}}{\sqrt{N}}+b_0\right)\left[1-\Theta\left(-\left(\frac{\mathbf{w}\cdot\mathbf{S}}{\sqrt{N}}+b\right)\left(\frac{\mathbf{w}^0\cdot\mathbf{S}}{\sqrt{N}}+b_0\right)\right)\right]\Big|_{S.P.} \tag{99}$$

$$= 1-\frac{1}{c_+}\int D\mathbf{S}\Theta\left(\frac{\mathbf{w}^0\cdot\mathbf{S}}{\sqrt{N}}+b_0\right)\Theta\left(-\left(\frac{\mathbf{w}\cdot\mathbf{S}}{\sqrt{N}}+b\right)\left(\frac{\mathbf{w}^0\cdot\mathbf{S}}{\sqrt{N}}+b_0\right)\right)\Big|_{S.P.} \tag{100}$$

$$= 1-\frac{1}{c_+}\mathrm{I}_+(R,b,b_0)\Big|_{S.P.}\,. \tag{101}$$

The derivation follows the same lines of the one for the Generalization Error. We stress that metric is evaluated at the saddle point for the order parameters,

$$s = \frac{\left\langle\!\left\langle \mathbb{E}_T\left[\left[1-\Theta\left(-\left(\frac{\mathbf{w}\cdot\mathbf{S}}{\sqrt{N}}+b\right)\left(\frac{\mathbf{w}^0\cdot\mathbf{S}}{\sqrt{N}}+b_0\right)\right)\right]\Theta\left(-\frac{\mathbf{w}^0\cdot\mathbf{S}}{\sqrt{N}}-b_0\right)\right]\right\rangle\!\right\rangle_{\mu_{\text{train}}}\Big\rangle_{\mu_{\text{test}}}}{1-\rho_{test}} \tag{102}$$

$$= 1-\frac{1}{c_-}\mathrm{I}_-(R,b,b_0)\Big|_{S.P.}\,. \tag{103}$$

The other metrics shown in the main paper can be inferred from r and s through Eqs. (8–14) in the main paper. Additionally, we here provide the expressions of the precision for negative samples $p(-)$, and of the negative-class F1-score, $F_1(-)$.

$$p(-) = \frac{s}{s+\frac{\rho_{\text{test}}}{1-\rho_{\text{test}}}(1-r)}\,, \tag{104}$$

$$F_1(-) = 2\cdot\frac{p(-)\cdot s}{p(-)+s}\,. \tag{105}$$

E WHEN THE TEACHER BIAS IS KNOWN

Here we consider the simpler scenario where the student's bias is not learned but fixed at $b = b_0$, corresponding to the situation in which the student has prior knowledge of the teacher's bias. This case is particularly insightful because it reveals some underlying symmetries of the problem and clarifies the concept of *informative samples* introduced in the main text.

Optimal training. In this setup, the self-consistent Eqs. (71–75) presented in App. C simplify, reducing to just the first four equations, since the student's bias b is fixed and does not need to be fixed self-consistently. The relevant information is thus captured entirely by the teacher-student overlap R. A key consequence of this fixed bias is that **translations of the student's hyperplane are not permitted**. Therefore, the alignment between teacher and student depends solely on the density of samples near the teacher's hyperplane, regardless of their class labels. When $b_0 \neq 0$, one class becomes more informative about the teacher's direction, meaning that samples closer to the teacher's hyperplane provide more information about its position. This becomes particularly evident here: when $b_0 \neq 0$, one class is inherently more informative, and to maximize the overlap R, **the optimal training set would ideally consist solely of samples from the minority class**. This phenomenon is illustrated in Fig. 8–(left), where the resulting R from training is plotted against ρ_{train} for various values of the teacher's bias b_0. We observe that as $|b_0|$ increases, this effect becomes more pronounced, as the minority class samples become increasingly concentrated near the teacher's hyperplane.

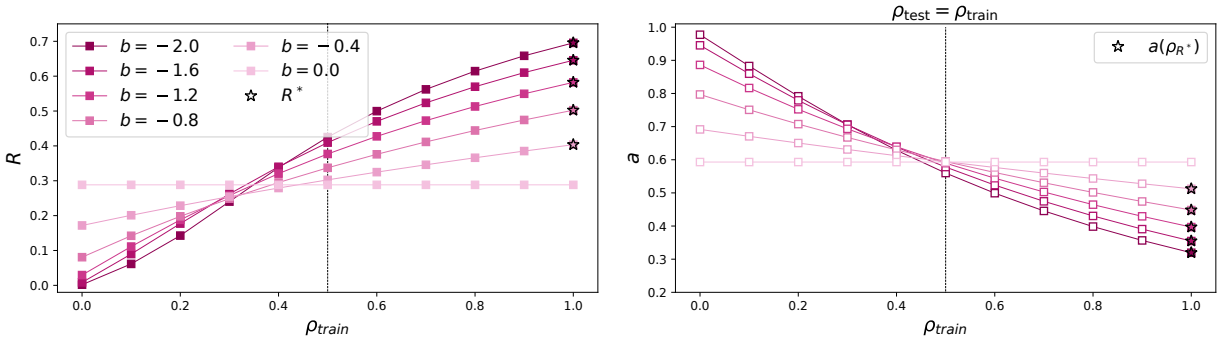

Figure 8: Overlap and accuracy on the spherical teacher-student perceptron, with the constraint $b = b_0$. **Left:** Teacher-student overlap R as a function of ρ_{train}, for $\alpha = 0.7$ and $T = 0.5$. Stars indicate the point where the overlap is maximized. The vertical line indicates $\rho_{\text{train}} = 0.5$. **Right:** Test-set accuracy a with $\rho_{\text{test}} = \rho_{\text{train}}$ as a function of ρ_{train}, for $\alpha = 0.7$ and $T = 0.5$. Stars indicate the point where the overlap is maximized which correspond to a low accuracy on the test set.

Invariance under sample reflection. The case $b = b_0 = 0$ is particularly interesting because it highlights a symmetry in the problem. Here, both classes contribute equally to informativeness, meaning there is no advantage in having more samples from one class over the other. This symmetry is reflected in the flat curve in Fig. 8–(left). It also manifests in the Free-Energy function, where the two energetic terms $\mathcal{G}_r^{\pm}$, defined in Eq. (34), become equal when $b_\sigma = b_0 = 0$. In fact, by applying the change of variables $\mathbf{S} \to -\mathbf{S}$ in the integral, we recover $\mathcal{G}_r^-$ from $\mathcal{G}_r^+$ and vice versa. This symmetry arises because the Boltzmann weight of each sample is identical regardless of its label, meaning that as long as the total number of samples remains fixed, the free-energy remains the same. In essence, when $b = 0$, there exists a bijection between flipping the labels and flipping the samples $\mathbf{S}$. Thus, in the second integral of Eq. (33), imposing a label flip also imposes a reflection in the data space, leading to the problem's invariance under sample reflection.

Test accuracy. Another important insight from this simplified case is that evaluating simple accuracy a on a test set with the same imbalance as the training set ($\rho_{\text{test}} = \rho_{\text{train}}$) can be misleading. We observe that the value of ρ_{train} which maximizes accuracy often corresponds to a lower overlap R. This discrepancy arises because a higher density of samples near the hyperplane increases the likelihood of misclassification, lowering accuracy even when the alignment between teacher and student is quite strong. This effect is demonstrated in Fig. 8–(right), where the accuracy on the test set is plotted against ρ_{train}.

F THE OPTIMAL TRAIN IMBALANCE

In this section, we show that the results provided through Fig. 3 are robust to hyperparameter changes. Moreover, we show that the shift in the optimal training imbalance $\rho(R^*)$ depends non-monotonously on the hyperparameters, and that this shift can be both positive and negative (meaning that $\rho(R^*)$ can be both larger and smaller than 0.5).

Fig. 9 displays the same quantities as Fig. 3, but shows what happens for more extreme b_0 values (both panels) and for a much larger α (right panel). As in the main text, these results are obtained by self-consistently solving the set of equations [Eqs. (71–75)] obtained for the order parameters, and deriving the values of the performance metrics values, as explained in App. D.

More data, better performance. A first obvious and unsurprising fact is that when data is more abundant ($\alpha = 8$), performance is overall better for all relevant metrics (a, p, F_1). In the following we are interested in relative performance and trends rather than on sheer performance.

Position of the peak $\rho(R^*)$. The most important and striking point is that for $\alpha = 8$ (panel (a')), the ρ_{train} at which R peaks ($\rho(R^*)$) is clearly away from 0.5, around 0.45, confirming that the optimal ρ_{train} is in general not 0.5, can be far away from it, and that it can be either larger or smaller than this commonly used value.

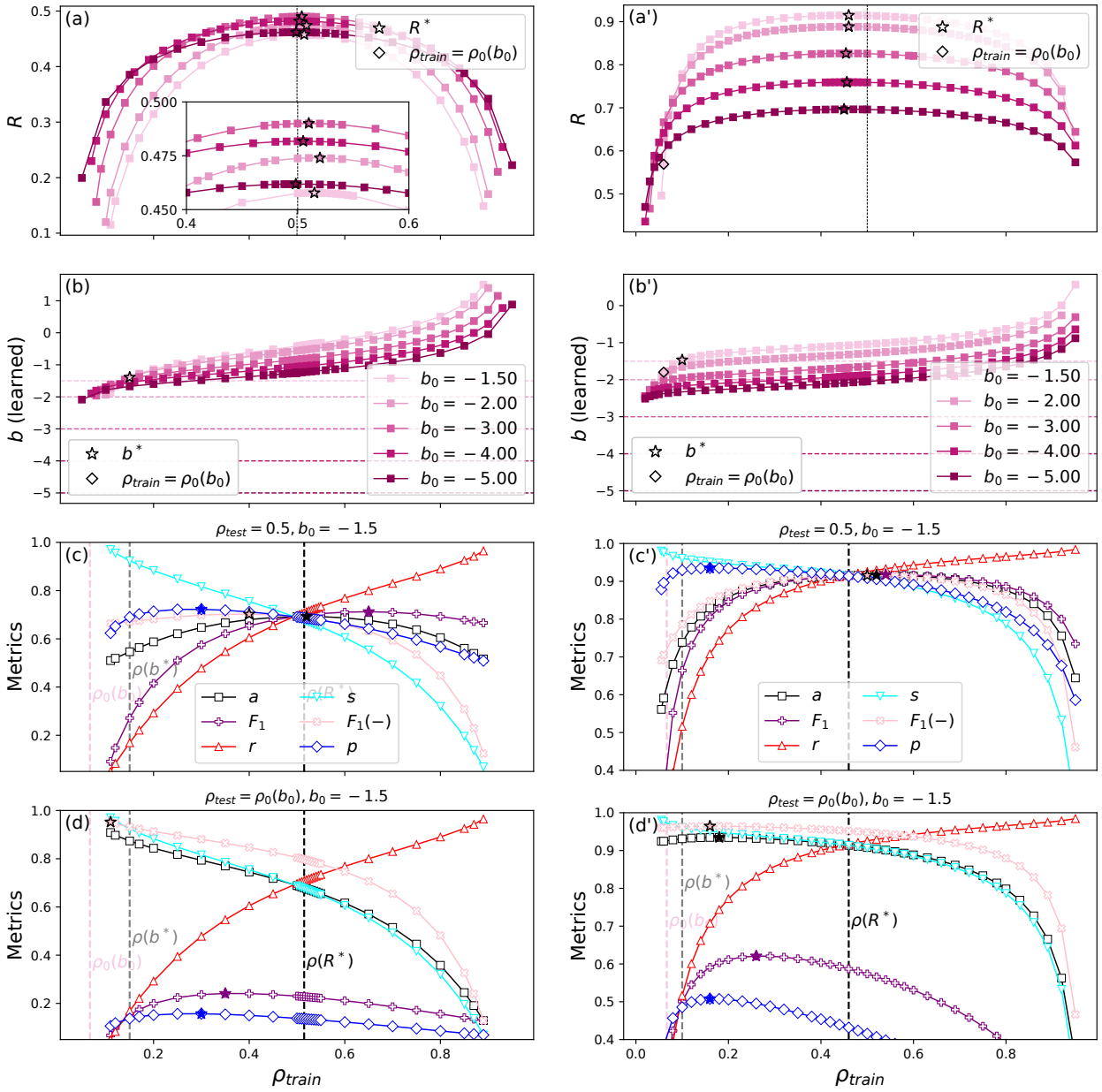

Figure 9: *Performance as function of* ρ_{train}, for $\alpha = 1.1$ (left) and $\alpha = 8$ (right). Analytical results for $T = 0.5$. The bias takes values $b_0 = -1.5, -2, -3, -4, -5$, which correspond to $\rho_0(b_0) = 7{\cdot}10^{-2}, 2{\cdot}10^{-2}, 6{\cdot}10^{-3}, 1{\cdot}10^{-3}, 3{\cdot}10^{-5}, 3{\cdot}10^{-7}$. **(a):** Student overlap R. Stars indicate the point where the overlap is maximized. Diamonds, which as in Fig. 3 indicate the performance at $\rho_{\text{train}} = \rho_0$, were computed only for $\alpha = 8, b_0 = -1.5$. The vertical line indicates $\rho_{\text{train}} = 0.5$. The inset is a zoom. **(b):** Same as (a), but for the student bias b. The horizontal lines indicate b_0. Now the stars indicate the points where bias is learnt perfectly (*i.e.* when $b = b_0$ is reached), and the diamonds that indicate the b learned under $\rho_{\text{train}} = \rho_0$ (shown when computed). **(c):** Accuracy, recall, sensitivity, precision, F1 score, and F1 score of the negative class, for $b_0 = -0.6$, $\rho_{\text{test}} = 0.5$. The stars indicate the peak of each curve. The vertical lines indicate ρ_0, the imbalance $\rho(b^*)$ at which the bias is optimal, and that at which the overlap is optimal, $\rho(R^*)$. **(d):** Same as (c), but for $\rho_{\text{test}} = \rho_0(b_0)$.

We also note (panel (c')) that while $\rho(a^*)$ is also away from 0.5, it is around 0.55, *i.e.* on the other side of the $\rho_{\text{train}} = 0.5$ reference. This indicates that, although we have identified the balanced accuracy as the metric which best reproduces the overlap, there is still a qualitative divide between a_{bal} and R.

Strong biases. In Fig. 3(b), the learned bias b could be smaller or larger than the teacher's bias b_0, depending on ρ_{train}. Here (Fig. 9(b,b')), under strong intrinsic imbalance (extreme ρ_0 values), we note that the bias learnt is essentially always under-estimated (in absolute value) by the student; when it is not the case, performance is extremely bad (for all metrics). This is consistent with our argument about the region $\mathcal{R}$ being quite empty ($\mathcal{R}$ is the region between the positive samples – squished against the teacher's hyperplane – and the typical location of the negative samples). As $|b_0|$ grows, the positive samples are increasingly squished against the teacher's hyperplane while the negative samples remain around the origin. Thus the region $\mathcal{R}$ grows in size and it is thus energetically favorable for the student to locate the hyperplane in this region, as it is easy to get 0 error there, even if the alignment R is not perfect.

Non-monotonicity of $\rho_{\text{train}}(R^*)$. As mentioned in the main text, $\rho_{\text{train}}(R^*)$ does not need to be a monotonic function of b_0, nor of α. This can be seen more clearly in Fig. 10. In panel (a), we see that $\rho_{\text{train}}(R^*)$ is a non-monotonic function of α. In panels (b,c) of Fig. 10 we note that the dependency in ρ_0 (itself controlled monotonously by b_0) is non trivial and itself depends on α. This is challenging to interpret in simple terms, as one could expect a priori that more intrinsic imbalance would have a straightforward effect on ρ_{train}.

Trend of $R^*(b_0)$. In Fig. 11(a), we show that the optimal reconstruction of the teacher (R^*) is an increasing function of the amount of data α. In Fig. 11(b,c), we show that R^* can be non-monotonous, especially in the case $\alpha = 1.1$.

Relative importance of intrinsic and train imbalance. In Fig. 9(a) we note that R mildly depends on ρ_0 but strongly depends on ρ_{train}. This trend is reversed when $\alpha = 8$ (Fig. 9(e)): $R(\rho_{\text{train}})$ curves are relatively more flat and spaced between them, indicating a strong dependence on ρ_0 and relatively weaker one on ρ_{train}. This could suggest that, if one has access to only small amounts of data, the choice of ρ_{train} would become crucial. On the contrary, with large amounts of data, ρ_0 plays a more relevant role than ρ_{train}. Our exploration of the hyperparameter space is however not extensive enough to confirm this conjecture.

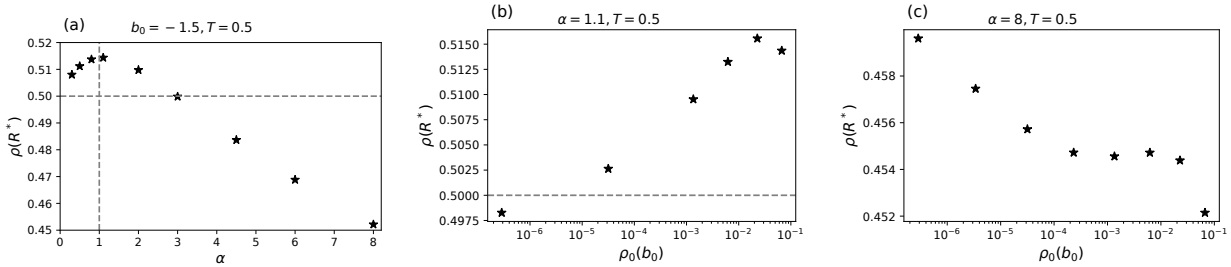

Figure 10: *Optimal ρ_{train} as function of the control parameters α and $\rho_0(b_0)$.* **(a):** $\rho(R^*)$ as function of data abundance α. **(b):** $\rho(R^*)$ as function of the intrinsic imbalance ρ_0 (controlled by b_0), for $\alpha = 1.1$. **(c):** Same as (b), for $\alpha = 8$. Dashed grey lines highlight the values $\rho_{\text{train}} = 0.5$ or $\alpha = 1$.

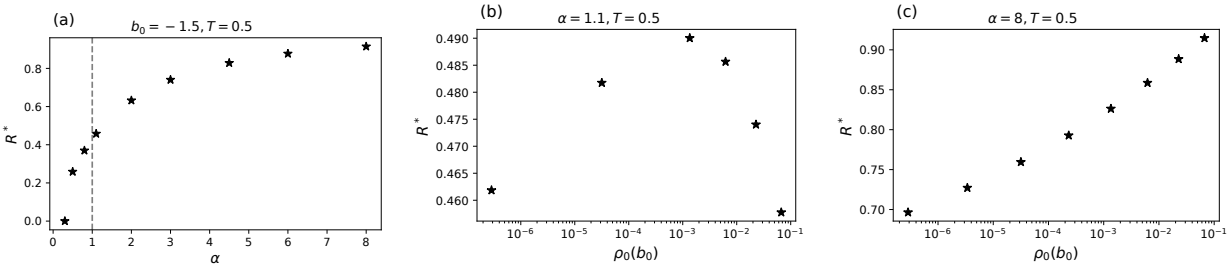

Figure 11: *Optimal overlap, R^*, found by fine-tuning ρ_{train}, as function of the control parameters α and $\rho_0(b_0)$* **(a):** R^* as function of data abundance α. **(b):** R^* as function of the intrinsic imbalance ρ_0 (controlled by b_0), for $\alpha = 1.1$. **(c):** Same as (b), for $\alpha = 8$.

Correlation of Metrics with R In Fig. 12, we show how each metric correlates with the overlap, when tuning ρ_{train}, and with $\rho_{\text{test}} = 0.5$.

The best metric is identified through two indicators:

1. how close the rightmost point (maximal overlap R, what we aim for) is to the topmost point (maximal metric, what we can infer, highlighted with a star).

2. how much the metric is almost in bijection with the overlap R, so that when the peak of the metric is passed, the optimal overlap is passed too.

From Fig. 12, one can see that no metric is perfectly representative of R, and that the balanced accuracy a_{bal} (the curves are for $\rho_{\text{test}} = 0.5$, so $a = a_{\text{bal}}$) is the one that correlates best, since the two branches of the curve are close together and well-aligned to a 1:1 constant slope. This is particularly clear in the bottom-right plot, where we see that a_{bal} peaks most to the right.

We stress that these results indicate which is the best metric to identify which ρ_{train} should be used, and not whether the metric is a good metric for testing.

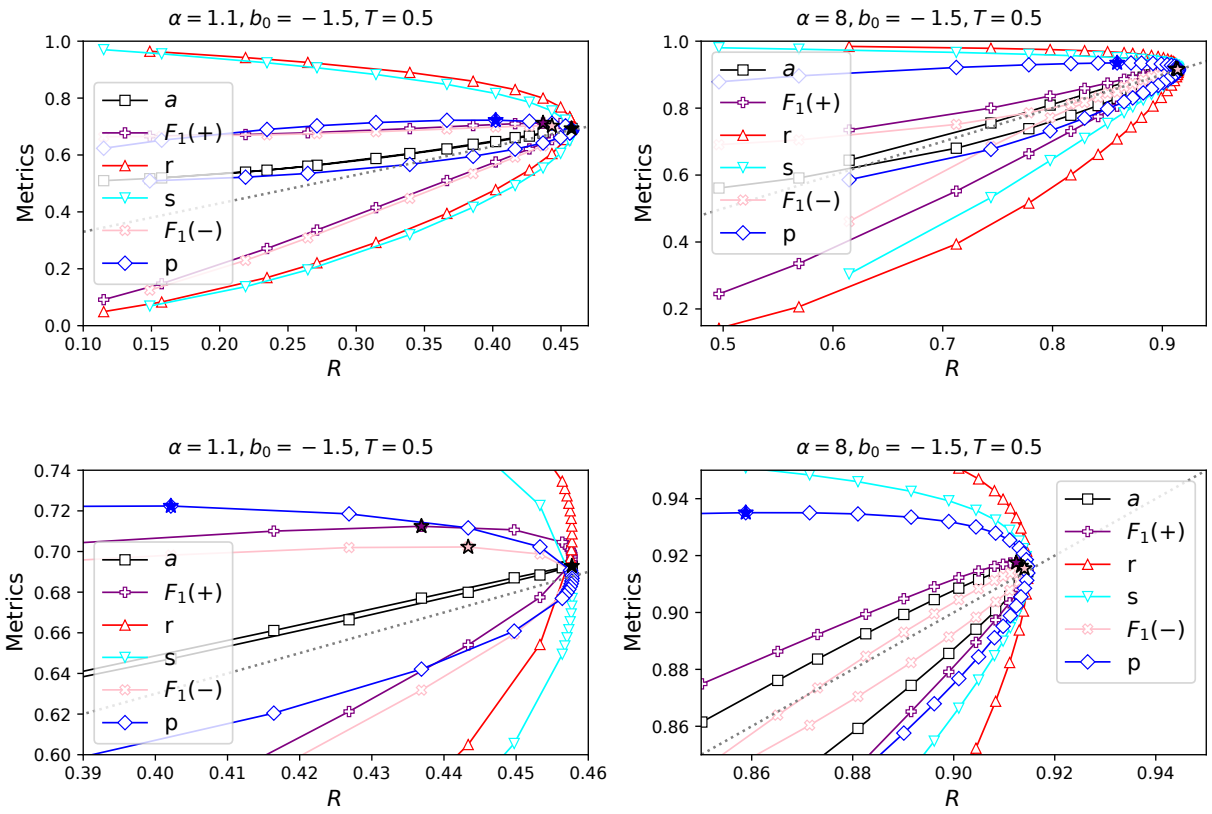

Figure 12: *Metrics as a function of R,* for $\rho_{\text{test}} = 0.5$. The stars indicate the maximum value of each curve when it's non trivial. **Left:** $\alpha = 1.1$. **Right:** $\alpha = 8$. The bottom figures are zooms of the top figures. The grey dotted lines are parallel to the diagonal $y = x$ line, as guide to the eye.

Metrics Dependence on ρ_{test} In Fig. 13 we show how metrics depend on ρ_{test}, as this is a more usual setup, since the ρ_{test} can easily be changed (while tuning ρ_{train} implies re-training the model for each new value of ρ_{train}). An interesting point is that the accuracy a (not balanced since ρ_{test} is varying), which is a weighted sum of recall r and specificity s which here have very similar values, is very robust against changes in ρ_{test} (although not formally constant), for both α values. This is a result of using $\rho_{\text{train}} \approx 0.5$: for different values, this feature is

lost. However one is usually interested in getting the best performance (higher a or other Metric at some ρ_{test}), not in getting equal r and s.

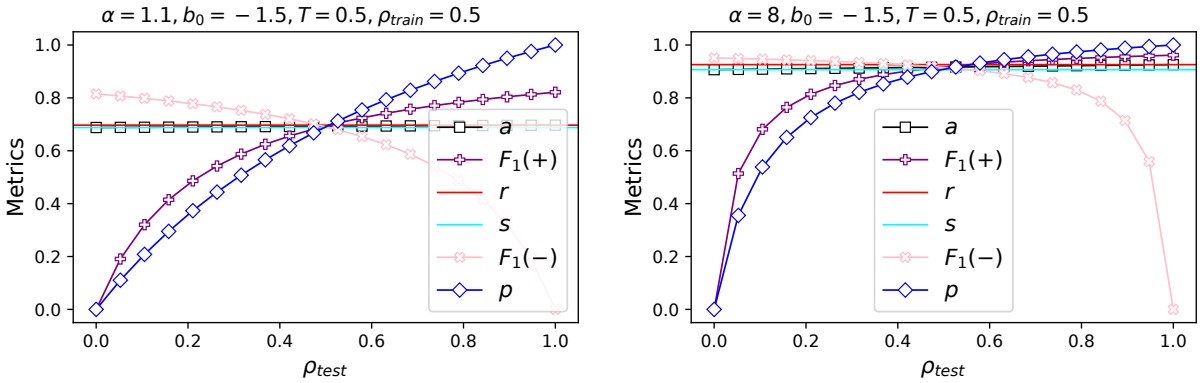

Figure 13: *Metrics as a function of* ρ_{test}, for $\rho_{\text{train}} = 0.5$.

G EXPERIMENTS

We consider three experimental setups, which allow to compare our theoretical model with more realistic setups. More precisely, our experiments allow to answer several questions, respectively: (i) what is the influence of the learning dynamics and of the read-out (activation) function on our results, within a controlled scenario that mirrors theoretical computations; (ii) what are the effects of dataset characteristics and model choice in a scenario more akin to practical machine learning applications; (iii) how realistic is the ADI assumption on data in yet more realistic AD tasks, and whether qualitative results are robust to this change of input data.

G.1 Perceptron Teacher Student on Gaussian data

In this setup, we use a Spherical Perceptron within a Teacher-Student framework. The model is linear, with the weights' norm constrained to be $O(1)$ and initialized according to a normal distribution; a sigmoid function is chosen for the activation. The model is trained through the minimization of the L2 loss. This configuration closely mirrors the theoretical setup, with two key distinctions: the use of a sigmoid activation function that outputs continuous values within the $[0, 1]$ range, and SGD learning dynamics, similar to real machine learning practices. It is important to note that the choice of a continuous activation function is essential to enable gradient descent dynamics, as a discontinuous function, such as the sign function g in Eq. 1, would result in gradients that are almost always zero. To produce binary labels, teacher outputs are discretized, assigning a label of 1 for values above the threshold thr = 0.5 and 0 otherwise. The SGD dynamic shares some characteristics with the Langevin dynamics used in the theoretical derivation, as both implement gradient descent on the training loss with added stochastic noise. The key difference lies in the correlated nature of the SGD noise, which arises from the repeated use of the mini-batch gradient estimates.

To reasonably approximate the conditions described in our theory, we set the data dimension to $N = 5000$. We observe that, by further increasing N, our results remain stable, suggesting that this setting is close enough to the $N \to \infty$ limit. We train the student model by performing multiple passes on the whole training set (epochs) until the training loss has converged. This corresponds to the "end of training" (equilibrium) regime assumed in theoretical computations. The noise level in SGD is governed by the learning rate (lr) and batch size (BS), which can be approximately related to the temperature introduced in the main text as $T \sim \text{lr}/\text{BS}$ [Jastrzebski et al., 2017].

For each combination of control parameters $(b_0, \rho_{\text{train}}, T, \alpha)$, we perform multiple runs, resampling both the dataset and teacher weights to compute the quenched average over the data distribution. Results on a balanced test set ($\rho_{\text{test}} = 0.5$) are illustrated in Fig. 14–(left). We observe trends that are qualitatively compatible with theoretical results, and most importantly, we find a non-trivial maximum of the metrics at $\rho^* \neq 0.5$. In experiments, we also evaluate the AUC metric since a threshold is needed to discretize the output of the learned

perceptron. We observe that it is rather insensitive to imbalance, confirming the findings of [Loffredo et al., 2024]. Figure 15 reports the trends of balanced accuracy and student's bias versus effective temperature. The experimental trends align qualitatively with the theoretical ones, identifying the low-noise and high-noise regions separated by T^* as introduced in Sec. 3.2.

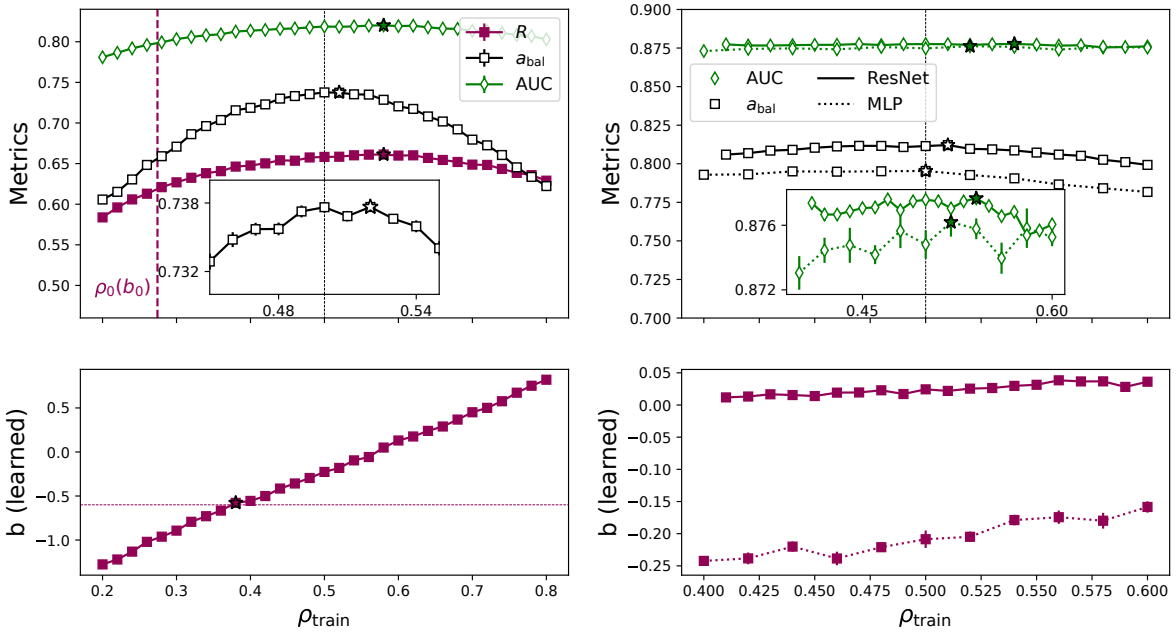

Figure 14: **Left:** Perceptron TS. $b_0 = -0.6$, $\alpha = 2.0$. Effective temperature $T = \frac{\text{lr}}{\text{BS}} = \frac{0.5}{20} = 2.5 \cdot 10^{-2}$. Each point represents the average over 40 re-samplings of the data and the error-bar represents its relative standard error. **Right:** MLP and ResNet34 on AD CIFAR-10. SGD optimizer, with momentum = 0.02 and weight decay 0.01. Each point represents the average over 10 re-samplings of the data and the error-bar represents its relative standard error.

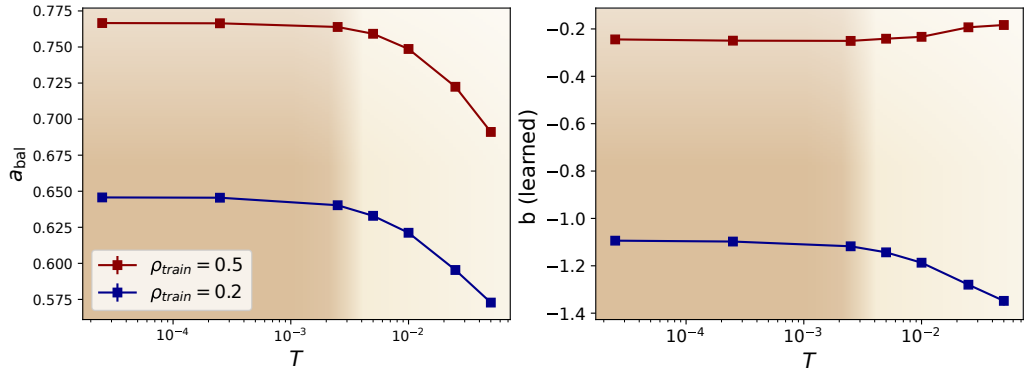

Figure 15: Perceptron TS vs T. Temperature is varied in SGD experiments by tweaking the mini-batch size. The learning rate is fixed to be lr = 0.05 and the mini-batch size varies $BS = \{2000, 200, 20, 10, 5, 2, 1\}$.

G.2 MLP and ResNet on AD CIFAR-10

In this setup, we employ a real-world anomaly detection dataset: Anomaly Detection CIFAR-10, a standard benchmark for anomaly detection tasks (see *e.g.* [Reiss et al., 2020]). This dataset involves re-labeling the original CIFAR-10 classes such that the first class (Airplanes) is designated as the anomaly (label +1), while all other classes are labeled as normal samples (label 0). CIFAR-10 consists of 60,000 samples, with 6,000 samples per

class, structured and non-independent by design. Defining one class as the anomaly sets the intrinsic imbalance to $\rho_0 = 0.1$. To explore various values of ρ_{train}, we perform sub-sampling on the dataset according to the desired level of imbalance, keeping the total number of training samples fixed at $N_{\text{train}} = 6000$.

We evaluate two representative models: an MLP with one hidden layer of 16 neurons, and an imageNet pre-trained ResNet34 [He et al., 2015] in which only the final linear layer is fine-tuned on the anomaly detection task, while all other layers remain frozen. Training is conducted via L2 loss minimization, using SGD as a learning dynamics.

In this experimental setup, not all theoretical hyper-parameters can be controlled. For instance, ρ_0 is fixed by the dataset, and defining b_0 is not feasible. Additionally, tuning α is challenging for both theoretical and practical reasons. Theoretically, the definition of α differs significantly from that in our analytical model, as the data dimensionality and model parameter count are no longer in one-to-one correspondence. Practically, varying N_{train} or adjusting the MLP's hidden layer size is constrained by limited data availability and the risk of overfitting. Nonetheless, for each parameter configuration, we re-train the models multiple times, re-sampling data from the original CIFAR-10.

Results on a balanced test set, shown in Fig. 14–(right), reveal a phenomenology qualitatively consistent with theoretical predictions. However, the effect strength and available statistics limit the conclusiveness of these findings.

G.3 Pretrained ResNet with PCA on BTAD and MVTec

G.3.1 Data distribution

In our study the input data is assumed to be Gaussian with i.i.d. components, and most importantly we assume to be in the ADI setup, *i.e.* we assume the distribution of the data of both classes to be very similar (and not split in two distinct groups, as in the MGI setup). This assumption may be questioned.

Here we show the relevance of this hypothesis by considering two real-world images datasets corresponding to anomaly detection problems (MVTec [Bergmann et al., 2019] and BTAD [Mishra et al., 2021]). We use an ImageNet-pretrained ResNet50 as feature map and observe the distribution of the first 30 Principal Components, conditioned on the class (see Fig. 16 and Fig. 17). We note that most features have a strongly overlapping distribution, *i.e.* the distribution between classes is very similar. This is indeed the case we are addressing in our modeling. In conclusion, for these datasets the ADI modeling is much more appropriate than the MGI setup.

G.3.2 Dependence on ρ_{train}.

Here we further investigate the impact of using realistic data on our results, by measuring again the changes in performance as a function of ρ_{train}.

For the BATD dataset [Mishra et al., 2021], we again use an imageNet-pretrained ResNet50 as backbone (with 2048 features) followed by a PCA down to 30 components. We can then train a Perceptron model (single layer, similar to our theoretical model) and vary ρ_{train} from 0 to 1. As we have little train and test data ($N_{train} = 300$ with 250 normal samples and 50 anomalous ones, and $N_{test} = 100$ with balanced classes), for each ρ_{train} we shuffle the train data 30 times, and sub-sample from this shuffle, to obtain the various ρ_{train} we are interested in (keeping the total N_{train} fixed). We average over these 30 re-samplings to estimate a mean performance (train and test balanced accuracies). In Fig. 18 (left), it is clear that the optimal ρ_{train} is not 0.5, but rather in the range $\rho^*_{\text{train}} \in [0.1, 0.3]$ (searching for a good test accuracy while keeping overfitting moderate). The key point for us is that ρ^*_{train} is non trivial (it is not 0.5).

As an aside, we note in Fig. 18 (right) that using re-weighting of the loss to correct for the effect of $\rho_{\text{train}} \neq 0.5$ is actually detrimental (in terms of test accuracy, all other things being equal). To be precise, except for the (very overfitting, hard-to-trust) point at $\rho_{\text{train}} = 0.01$ (corresponding to a single example in the anomaly class), and for overly large values of ρ_{train} (where performance is poor anyways), in this case it is always better to not re-weight the classes.

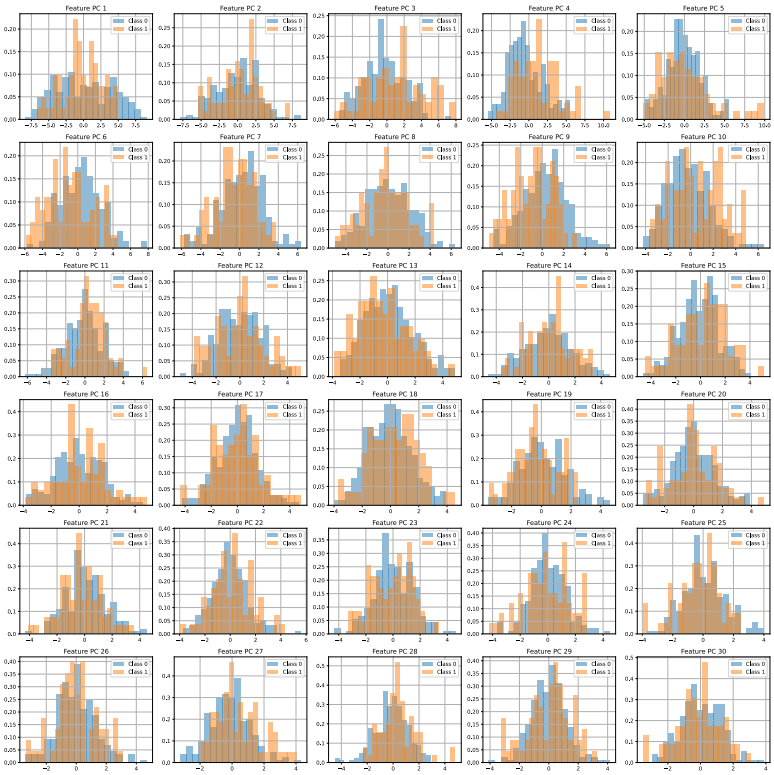

Figure 16: MVTec data distribution, after a pretrained ResNet50 and a PCA.

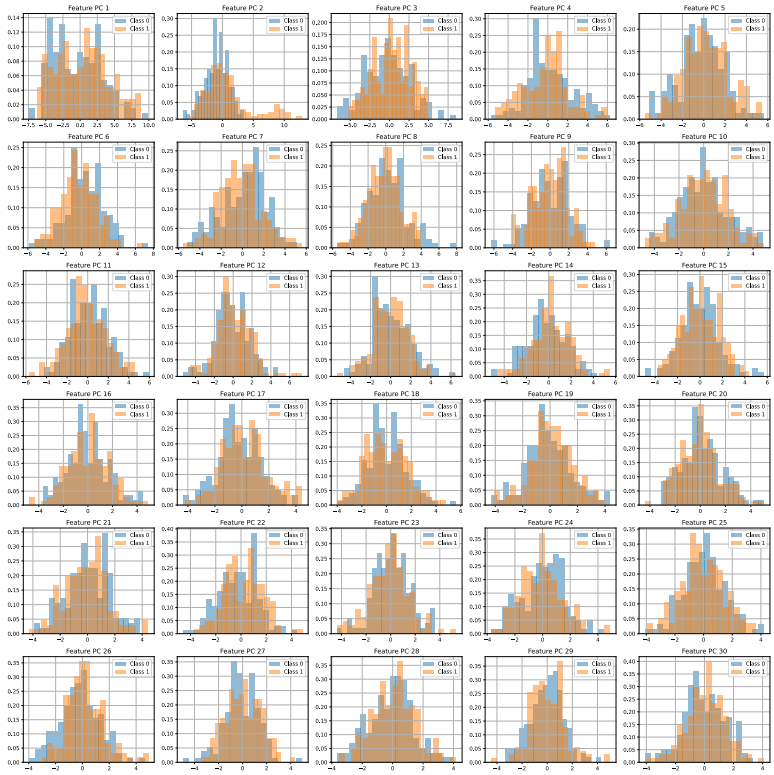

Figure 17: BTAD data distribution, after a pretrained ResNet50 and a PCA.

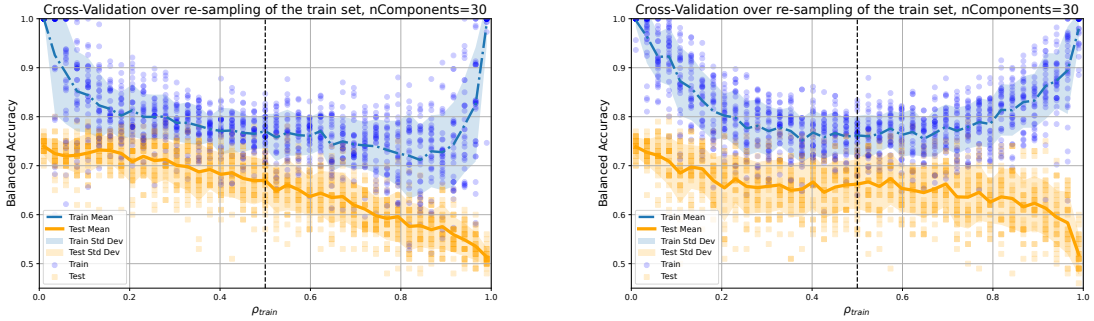

Figure 18: Balanced Accuracy as a function of ρ_{train} for a realistic AD task.
(Left): no re-balancing of the loss. The ρ_{train} values that strike a good balance between good test performance and moderate overfitting are in the range $\rho_{\text{train}} \in [0.1, 0.3]$.
(Right): re-balancing of the loss to correct for the effect of $\rho_{\text{train}} \neq 0.5$. Balanced accuracy is lower (in the relevant range).