A High Dimensional Statistical Model for Adversarial Training: Geometry and Trade-Offs

Kasimir Tanner[1]   **Matteo Vilucchio[1]**   **Bruno Loureiro[2]**   **Florent Krzakala[1]**

[1] Information Learning and Physics Laboratory, École Polytechnique Fédérale de Lausanne (EPFL)
[2] Departement d'Informatique, École Normale Supérieure, PSL & CNRS (ENS)

Abstract

This work investigates adversarial training in the context of margin-based linear classifiers in the high-dimensional regime where the dimension d and the number of data points n diverge with a fixed ratio $\alpha = n/d$. We introduce a tractable mathematical model where the interplay between the data and adversarial attacker geometries can be studied, while capturing the core phenomenology observed in the adversarial robustness literature. Our main theoretical contribution is an exact asymptotic description of the sufficient statistics for the adversarial empirical risk minimiser, under generic convex and non-increasing losses for a Block Feature Model. Our result allow us to precisely characterise which directions in the data are associated with a higher generalisation/robustness trade-off, as defined by a robustness and a usefulness metric. We show that the the presence of multiple different feature types is crucial to the high sample complexity performances of adversarial training. In particular, we unveil the existence of directions which can be defended without penalising accuracy. Finally, we show the advantage of defending non-robust features during training, identifying a uniform protection as an inherently effective defence mechanism.

1   INTRODUCTION

The susceptibility of machine learning models to adversarial attacks — subtle yet strategically crafted data perturbations — has been an ongoing concern for various machine learning models, from linear classifiers to deep neural networks. In particular, this vulnerability is intrinsic to margin-based classifiers (Ilyas et al., 2019). Also in image classification, seemingly innocuous modifications, like tiny stickers on road signs, can dramatically mislead models that otherwise exhibit strong generalisation (Pavlitska et al., 2023).

The problem has been theoretically studied in linear models where solutions could be obtained analytically. Previous studies were focusing on the fundamental limits of the trade-off between adversarial and generalisation errors (Javanmard et al., 2020) or on including a non trivial data covariance and comparing the performances of adversarial training to Bayes optimal errors (Taheri et al., 2023).

Nonetheless a comprehensive understanding of how structure defines feature types — robust, useful, or both — affect model performance is still developing. The interplay of the data structure with attack and training geometries remains a particularly fertile ground for theoretical investigation, with potential implications for enhancing adversarial robustness and developing more effective protection methods.

In this paper, we introduce a structured, high-dimensional model for studying adversarial classification under margin-based classifiers. Our **main contributions** are fourfold:

- We introduce a mathematically tractable model for investigating the interplay between the data, attack and defence geometries. Despite its simplicity, we show our model is rich enough to capture the key phenomenology observed in practical adversarial

---

training setups.

- We show that, in the high dimensional proportional limit (where the number of samples and covariate dimension diverge at fixed ratio), the relevant statistical properties of the adversarial empirical risk minimiser can be exactly characterised by a finite set of sufficient statistics.

- Leveraging our high-dimensional characterisation, we are able to show the importance of distinguishing different features to have different performances in the large sample complexity regime: considering single block models leads to the same performances for any kind of adversarial training. Additionally we also derive specific conditions under which defending non-robust features is beneficial in this prototypical data model.

- Finally, building on our findings we investigate the interplay between data and attack geometry in the effectiveness of adversarial training. In particular, we show how attack geometry's direction can be divided in two groups: directions leading to a trade-off and directions that can be successfully defended without sacrificing accuracy.

This manuscript is organised as follows. Section 2 introduces the data model and the margin-based adversarial training protocol. Section 3 describes our main theoretical results, namely the asymptotic characterisation of the linear adversarial estimator. Section 4 discusses the implications of the main theoretical formulas. Finally, Section 5 reports on the additional experiments. The code used to produce all the figures in this manuscript is in the Supplementary Materials.

Related works

Adversarial attacks —   In the study of neural networks, the vulnerability to adversarial attacks is well-established, with early works like (Szegedy et al., 2013; Goodfellow et al., 2014; Papernot et al., 2016) uncovering this intriguing weakness. Adversarial Training, particularly through methods like projected gradient descent, has emerged as a leading defence strategy, as explored in-depth by Madry et al. (2017).

Understanding adversarial robustness has been a long standing challenge. Schmidt et al. (2018) highlighted the necessity of reaching higher sample complexity in adversarial training to achieve the same generalisation performance as standard training. Ilyas et al. (2019) proposed that datasets contain predictive yet imperceptible features vulnerable to attacks. Additionally, Tsilivis et al. (2022) notices the difficulty of optimising a data-set to improve adversarial robust-

ness. The adversarial setting has also been studied in the neural tangent kernel regime by Tsilivis and Kempe (2022).

The idea of a fundamental trade-off between adversarial robustness and standard accuracy has been noted in Tsipras et al. (2019); Zhang et al. (2019); Suggala et al. (2019). Later, Bhagoji et al. (2019); Dan et al. (2020); Javanmard and Soltanolkotabi (2022) have examined this trade-off in the case of Gaussian Mixture Models. Raghunathan et al. (2020) show that a class of augmented estimators can have a worse generalisation error than the standard estimator. Taheri et al. (2023); Javanmard et al. (2020) contribute to this discourse by examining the impact of adversarial training on the interpolation threshold and double descent in the case of Gaussian data. More recently Roth et al. (2020); Ribeiro et al. (2024) have studied adversarial training as a form of data dependent regularisation and Vilucchio et al. (2024a) considered the interplay between data structure and regularisation geometry. The trade-off between robust and clean generalisation error that we are going to consider in this paper has been found fundamental in the design of algorithms such as TRADES (Zhang et al., 2019) and ARoW Regularisation (Yang et al., 2023).

Exact asymptotics —   Our main theoretical result pertains to an established literature employing techniques from high-dimensional probability (Thrampoulidis et al., 2014, 2015; Sur and Candès, 2019; Dhifallah and Lu, 2020a), random matrix theory (Bean et al., 2013; Mai et al., 2019; Liao et al., 2020; Mei and Montanari, 2022; Xiao et al., 2022; Schröder et al., 2023) and statistical physics (Aubin et al., 2020; Mignacco et al., 2020; Gerace et al., 2021; Bordelon et al., 2020; Vilucchio et al., 2024b; Okajima et al., 2023; Adomaityte et al., 2023a,b) to derive exact asymptotic results of high-dimensional statistical estimation problems. Of particular relevance to our work is Loureiro et al. (2022a), who proved a formula for the sufficient statistics of general Gaussian Covariate models. While our work leverages their results, our formulas are more general, as they account for adversarial attacks and training. Moreover, in our proof we use a mapping to an Approximate Message Passing (AMP) for adversarial training on structured problems, which builds upon Rangan (2011); Javanmard and Montanari (2013); Takahashi and Kabashima (2022); Loureiro et al. (2021, 2022b); Gerbelot and Berthier (2023). Recently Vilucchio et al. (2025) mixed the AMP techniques used in this work with a Gaussian lower bound to study non-convex settings.

Kasimir Tanner, Matteo Vilucchio, Bruno Loureiro, Florent Krzakala

The present work extends these previous analyses in three fundamental aspects. First, unlike the unstructured Gaussian setting of Javanmard et al. (2020), our framework explicitly captures feature-dependent robustness through the Block Feature Model, allowing us to study how different features contribute to adversarial vulnerability. Second, we generalise the attack models of Taheri et al. (2023) by introducing structured attacks and defences through arbitrary positive definite matrices, providing a more complete picture of adversarial training. Third, while Zhang et al. (2019) provided finite-dimensional bounds, our analysis yields exact asymptotic in high dimensions.

The code to reproduce the figures can be found at github.com/IdePHICS/Adversarial-Setting.

2 SETTING SPECIFICATION

Consider a supervised binary classification problem with training data $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \in \mathbb{R}^d \times \{-1, +1\}$. In the following, we focus on a generalised estimation problem, where for each $i = 1, \cdots, n$ we assume the covariates are independently drawn from $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma_x})$ and with labels $y_i \sim \mathbb{P}(y|\boldsymbol{\theta}_0^\top \boldsymbol{x}_i)$ for a fixed parameter $\boldsymbol{\theta}_0 \in \mathbb{R}^d$.[1] Although our theoretical results in Section 3 hold under a generic likelihood $P(y|z)$, for concreteness the discussion in Section 4 will be mostly focused on the probit model $\mathbb{P}(y|z) = 1/2 \operatorname{erfc}(-z/\sqrt{2}\tau)$, where the parameter $\tau > 0$ controls the noise level. In particular, note that for $\tau \to 0$ we have $\mathbb{P}(y|z) = \delta(y - \operatorname{sign}(z))$.

Given the training data $\mathcal{D}$, our goal in the following is to investigate the capacity of margin-based linear classifiers $\hat{y}(\hat{\boldsymbol{\theta}}, \boldsymbol{x}) = \operatorname{sign}(\hat{\boldsymbol{\theta}} \cdot \boldsymbol{x}/\sqrt{d})$ in robustly and efficiently classifying the data under adversarial attacks, where $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathcal{D})$ is an estimator that has been learned from the training data. The setting introduced above is often refereed to as a *teacher-student* setting, and is widespread in the high-dimensional statistics literature (Sur and Candès, 2019; Taheri et al., 2023; Clarté et al., 2023a).

In the case where the data points are not attacked, the metric of interest is the (clean) *generalisation error* defined as

$$E_{\text{gen}} = \mathbb{E}_{y,\boldsymbol{x}}\left[\mathbb{1}(y \neq \hat{y}(\hat{\boldsymbol{\theta}}, \boldsymbol{x}))\right], \quad (1)$$

where the expectation is taken over input-output pairs generated using the same $\boldsymbol{\theta}_0$ as in training.

If the adversary can perturb the input sample $\boldsymbol{x}_i$ its

objective is to find a perturbation $\boldsymbol{v}_i$ which leads to a wrongly classified sample, *i.e.* $y(\boldsymbol{x}_i) \neq \hat{y}(\hat{\boldsymbol{\theta}}, \boldsymbol{x}_i + \boldsymbol{v}_i)$. We focus on allowed perturbations that must have a $\boldsymbol{\Sigma_v}$-induced norm smaller or equal than a fixed attack strength ε_g, beyond which an attack could be identified — $\|\boldsymbol{v}\|_{\boldsymbol{\Sigma_v}^{-1}} \leq \varepsilon_g$. We will call $\boldsymbol{\Sigma_v}$ the *attack matrix*. By normalising $\boldsymbol{\Sigma_v}$ we can interpret ε_g as the global strength of the attack or attack budget and $\boldsymbol{\Sigma_v}$ as the attack geometry.

This model generalises the case considered by Javanmard et al. (2020); Taheri et al. (2023) where they consider the cases of bounded attacks in ℓ_2 norm.

Given the previous adversarial constraints, we introduce the *adversarial generalisation error*. This metric quantifies the student's performance under adversarial attack

$$E_{\text{adv}} = \mathbb{E}_{y,\boldsymbol{x}}\left[\max_{\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma_v}^{-1}} \leq \varepsilon_g} \mathbb{1}(y \neq \hat{y}(\hat{\boldsymbol{\theta}}, \boldsymbol{x} + \boldsymbol{\delta}))\right], \quad (2)$$

where the expectation is taken over pairs of input-output generated with the same $\boldsymbol{\theta}_0$ used during training. Notably, this adversarial generalisation error is an extension of the standard generalisation error, with the latter being the special case where $E_{\text{gen}} = E_{\text{adv}}(\varepsilon_g = 0)$.

As noted by Zhang et al. (2019); Yang et al. (2023) the adversarial generalisation error can be decomposed into a $E_{\text{adv}} = E_{\text{gen}} + E_{\text{bnd}}$, where we will refer to E_{bnd} as *boundary error*. This metric measures the number of samples correctly classified but that are within attack range from the decision boundary. Explicitly one has

$$E_{\text{bnd}} = \mathbb{E}_{y,\boldsymbol{x}}\left[\mathbb{1}(y = \hat{y}(\hat{\boldsymbol{\theta}}; \boldsymbol{x})) \max_{\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma_v}^{-1}} \leq \varepsilon_g} \mathbb{1}(y \neq \hat{y}(\hat{\boldsymbol{\theta}}, \boldsymbol{x} + \boldsymbol{\delta}))\right].$$
$$(3)$$

2.1 Empirical Risk Minimisation and Adversarial Case

To estimate the student vector that achieves the minimal adversarial error the most common way is to consider a convex surrogate empirical version of the adversarial error we are trying to minimise (Bach, 2024). We define the adversarial risk function as

$$\sum_{i=1}^n \max_{\|\boldsymbol{\delta}_i\|_{\boldsymbol{\Sigma_\delta}^{-1}} \leq \varepsilon_t} g\left(y_i \frac{\boldsymbol{\theta}^\top(\boldsymbol{x}_i + \boldsymbol{\delta}_i)}{\sqrt{d}}\right) + r(\boldsymbol{\theta}), \quad (4)$$

where g is a convex loss, $r(\boldsymbol{\theta})$ is a convex regularisation term and $\boldsymbol{\Sigma_\delta}$ is a positive definite matrix.

---

[1]Note our results also hold under the assumption of $\boldsymbol{\theta}_0 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma_\theta})$.

We will call $\boldsymbol{\Sigma_\delta}$ the *defence matrix* and it will be normalised in the same way as $\boldsymbol{\Sigma_v}$.

By considering any decreasing convex loss g, we can simplify the inner maximisation problem leading to an equivalent form of the risk

$$\sum_{i=1}^{n} g\left(y_i \frac{\boldsymbol{\theta}^\top \boldsymbol{x}_i}{\sqrt{d}} - \varepsilon_t \frac{\sqrt{\boldsymbol{\theta}^\top \boldsymbol{\Sigma_\delta} \boldsymbol{\theta}}}{\sqrt{d}}\right) + r(\boldsymbol{\theta}). \quad (5)$$

Minimising the risk function provides an estimate $\hat{\boldsymbol{\theta}}(\mathcal{D})$ for the student weights. While our framework is versatile enough to accommodate various convex regularisation functions for the rest of the paper we set $r(\boldsymbol{\theta}) = \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2$.

Our analysis will be carried out in the *high-dimensional proportional limit*. Specifically, we investigate settings where both the dimension d and the number of training samples n are large $d, n \to \infty$, whilst maintaining a fixed sample complexity $\alpha := n/d$.

2.2 Block Feature Data Model

In Tsipras et al. (2019); Ilyas et al. (2019); Tsilivis and Kempe (2022) the discourse is centred on distinguishing between *useful* and *robust* features. The usefulness of a feature is a measure of how much that specific feature correlates with the output that we want to predict. The robustness is a measure of the same correlation after an attack is performed on the data point. The proposed view is that adversarial vulnerability increases when the classification is based on useful but non-robust features.

In the context of linear models considered in this manuscript we define *usefulness* $\mathcal{U}_{\boldsymbol{\theta}_0}$ and *robustness* $\mathcal{R}_{\boldsymbol{\theta}_0}$ respectively as

$$\mathcal{U}_{\boldsymbol{\theta}_0} = \frac{1}{\sqrt{d}} \mathbb{E}_{\boldsymbol{x},y}[y\boldsymbol{\theta}_0^\top \boldsymbol{x}], \quad (6)$$

$$\mathcal{R}_{\boldsymbol{\theta}_0} = \frac{1}{\sqrt{d}} \mathbb{E}_{\boldsymbol{x},y}\left[\inf_{\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma_v}^{-1}} \leq \varepsilon_g} y\boldsymbol{\theta}_0^\top (\boldsymbol{x} + \boldsymbol{\delta})\right]. \quad (7)$$

These two metrics capture the relationship between the learning task ($\boldsymbol{\theta}_0$) and the data model, focusing on how informative are the features ($\mathcal{U}_{\boldsymbol{\theta}_0}$) and how they remain informative under worst-case input perturbations ($\mathcal{R}_{\boldsymbol{\theta}_0}$).

Following the interpretation of linear models in Hastie et al. (2022, Sec. 1.2) we see the importance of defining different types of features with different properties and introducing features that are more/less easily attack-able.

The Block Feature Model While our theoretical results hold for a wide array of data models, we want to study adversarial robustness in a reductionist spirit. We introduce the *Block Feature Model (BFM)*, which allows to systematically vary feature usefulness and robustness to understand their impact on (adversarial) generalisation metrics.

For a given dimension d we define $k \leq d$ blocks with individual sub-dimensions $\{d_\ell\}_{\ell=1}^k$ satisfying $\sum_{\ell=1}^k d_\ell = d$. This allows us to write the quantities of interest as

$$\begin{aligned}
\boldsymbol{\Sigma_x} &= \text{blockdiag}\left(\psi_1 \mathbb{1}_{d_1}, \dots, \psi_k \mathbb{1}_{d_k}\right), \\
\boldsymbol{\Sigma_\delta} &= \text{blockdiag}\left(\Delta_1 \mathbb{1}_{d_1}, \dots, \Delta_k \mathbb{1}_{d_k}\right), \\
\boldsymbol{\Sigma_v} &= \text{blockdiag}\left(\Upsilon_1 \mathbb{1}_{d_1}, \dots, \Upsilon_k \mathbb{1}_{d_k}\right), \\
\boldsymbol{\Sigma_\theta} &= \text{blockdiag}\left(t_1 \mathbb{1}_{d_1}, \dots, t_k \mathbb{1}_{d_k}\right),
\end{aligned} \quad (8)$$

where each $\psi_\ell, \Delta_\ell, t_\ell$ and Υ_ℓ is greater than zero for each $\ell = 1, \dots k$ to preserve the positive definiteness. Notice, this model can easily be extended to allow for power-law distributions on the eigenvalues by considering $d_\ell = 1$ and $\psi_\ell = \ell^{-\beta_\psi}$ for $\beta_\psi > 1$.

The parameters of the BFM have direct interpretations: ψ_ℓ characterises the variance of features in block ℓ, Δ_ℓ describes how sensitive these features are to attacks, Υ_ℓ determines the defence strategy for the block, and t_ℓ captures the learning task's importance on that given feature block. This parametrisation naturally emerges when considering the training of neural networks in specific regimes (Chizat and Bach, 2019) and additionally the features will have distinct robustness properties (Tsipras et al., 2019).

The BFM allows for the expression of artificial datasets capturing the simplest definitions of structure, whilst capturing the intricacies of realistic power-law data, as can be found in real images (Wainwright and Simoncelli, 1999; Simoncelli and Olshausen, 2001).

Additionally, we define the *Strong Weak Feature Model (SWFM)* as a special case of the BFM. In the SWFM, we only consider two blocks $k = 2$ where the relative sizes are tuned as $d_\ell/d \to \phi_\ell \in (0, 1)$. Under this data model we will also speak of the usefulness and robustness of a single block of features, where we generalise Eqs. (6) and (7) just for a subset of the features.

3 MAIN TECHNICAL RESULTS : EXACT ASYMPTOTICS

The core technical result is a rigorous, closed-form characterisation of the properties of the estimator

for the previously described model, and the corresponding training and generalisation errors in the high-dimensional limit.

Assumptions The first assumption that we consider is that all matrices to have a well defined spectral distribution in the high dimensional limit. We will consider $\boldsymbol{\Sigma_x} = \mathrm{S}^\top \mathrm{diag}(\omega_i)\mathrm{S}$, $\zeta_i = \mathrm{diag}(\mathrm{S}\boldsymbol{\Sigma_\delta}\mathrm{S}^\top)_i$ and $v_i = \mathrm{diag}(\mathrm{S}\boldsymbol{\Sigma_v}\mathrm{S}^\top)_i$. Next we assume that $\boldsymbol{\theta}_0^\top \boldsymbol{\Sigma_x}\boldsymbol{\theta}_0/d$ converges to a given value ρ in the limit and that the entries of $\bar{\boldsymbol{\theta}} = \mathrm{S}\boldsymbol{\Sigma_x}^\top\boldsymbol{\theta}_0/\sqrt{\rho}$ converge as well to a limiting distribution. Finally we assume that in the high dimensional limit the spectral distributions for the matrices and the distributions of the elements of the vectors just defined converge jointly to a p.d.f., *i.e.* $1/d\sum_{i=1}^d \delta(\omega - \omega_i)\delta(\bar{\theta} - \bar{\theta}_i)\delta(\zeta - \zeta_i)\delta(v - v_i) \to \mu(\omega, \bar{\theta}, \zeta, v)$.

These assumptions are standard in high-dimensional statistics and naturally extend previous frameworks to our adversarial setting. The existence of well-defined spectral distributions, appearing in works like Sur and Candès (2019); Mei and Montanari (2022), is necessary to characterise the asymptotic behaviour of the data model and training procedure. The convergence of $\boldsymbol{\theta}_0^\top \boldsymbol{\Sigma_x}\boldsymbol{\theta}_0/d$ and $\bar{\boldsymbol{\theta}}$ ensures the signal strength remains controlled as dimensionality increases, following similar conditions in Loureiro et al. (2022a). Finally, the joint convergence assumption of the spectral distributions, as used in Dhifallah and Lu (2020b), guarantees that correlations between data, attack, and defence geometries are well-behaved in the high-dimensional limit.

Under the previous assumptions and the model of Section 2 we can characterise the behaviour of the clean and adversarial generalisation error of the ERM estimator $\hat{\boldsymbol{\theta}}(\mathcal{D})$ that minimises the risk in Eq. (4).

Theorem 3.1. *For the ERM estimator of the risk function with ℓ_2 regularisation $r(\boldsymbol{\theta}) = \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2$ and $\lambda \geq 0$, under the data model defined in Section 2 and in the high dimensional proportional limit, the generalisation error E_{gen} and the boundary error E_{bnd} defined in Eqs. (1) and (3) concentrate to*

$$E_{\mathrm{gen}} = \frac{1}{\pi}\arccos\left(m/\sqrt{(\rho+\tau^2)q}\right), \quad (9)$$

$$E_{\mathrm{bnd}} = \int_0^{\varepsilon_g \frac{\sqrt{A}}{\sqrt{q}}}\mathrm{erfc}\left(\frac{-\frac{m}{\sqrt{q}}\nu}{\sqrt{2(\rho+\tau^2-m^2/q)}}\right)\frac{e^{-\frac{\nu^2}{2}}}{\sqrt{2\pi}}\,\mathrm{d}\nu, \quad (10)$$

and the adversarial generalisation error concentrates to $E_{\mathrm{adv}} = E_{\mathrm{gen}} + E_{\mathrm{bnd}}$.

The values of m and q are the solutions of a system of eight self-consistent equations for the unknowns $(m, q, V, P, \hat{m}, \hat{q}, \hat{V}, \hat{P})$. The first four equations are

dependant on the loss function g and the adversarial training strength ε_t and read

$$\begin{cases} \hat{m} = \alpha\mathbb{E}_\xi\left[\int_\mathbb{R}\mathrm{d}y\,\partial_\omega\mathcal{Z}_0 f_g(y, \sqrt{q}\xi, P)\right] \\ \hat{q} = \alpha\mathbb{E}_\xi\left[\int_\mathbb{R}\mathrm{d}y\,\mathcal{Z}_0 f_g^2(y, \sqrt{q}\xi, P)\right] \\ \hat{V} = -\alpha\mathbb{E}_\xi\left[\int_\mathbb{R}\mathrm{d}y\,\mathcal{Z}_0\partial_\omega f_g(y, \sqrt{q}\xi, P)\right] \\ \hat{P} = -\frac{\varepsilon_t}{2\sqrt{P}}\alpha\mathbb{E}_\xi\left[\int_\mathbb{R}\mathrm{d}y\,y\mathcal{Z}_0 f_g(y, \sqrt{q}\xi, P)\right] \end{cases}, \quad (11)$$

where $\xi \sim \mathcal{N}(0,1)$ and $\mathcal{Z}_0 = 1/2\,\mathrm{erfc}(-y\omega/\sqrt{2(V+\tau^2)})$ and $f_g(y, \omega, V, P) = (\mathcal{P}(\omega) - \omega)/V$, where $\mathcal{P}$ is the following proximal operator

$$\mathcal{P}(\omega) = \min_x\left[\frac{(x-\omega)^2}{2V} + g(yx - \varepsilon_t\sqrt{P})\right]. \quad (12)$$

The second set of equation depend on the spectral distribution of the matrices $\boldsymbol{\Sigma_x}, \boldsymbol{\Sigma_\delta}$ and on the limiting distribution of the elements of $\bar{\boldsymbol{\theta}}$. The equations read

$$\begin{cases} m = \mathbb{E}_\mu\left[\frac{\hat{m}\bar{\theta}^2}{\lambda+\hat{V}\omega+\hat{P}\delta}\right] \\ q = \mathbb{E}_\mu\left[\frac{\hat{m}^2\bar{\theta}^2\omega+\hat{q}\omega^2}{(\lambda+\hat{V}\omega+\hat{P}\delta)^2}\right] \\ V = \mathbb{E}_\mu\left[\frac{\omega}{\lambda+\hat{V}\omega+\hat{P}\delta}\right] \\ P = \mathbb{E}_\mu\left[\zeta\frac{\hat{m}^2\bar{\theta}^2+\hat{q}\omega^2}{(\lambda+\hat{V}\omega+\hat{P}\delta)^2}\right] \end{cases}. \quad (13)$$

The value of A can be obtained from the solution of the same system of self consistent equations as

$$A = \mathbb{E}_\mu\left[v\frac{\hat{m}^2\bar{\theta}^2\omega+\hat{q}\omega^2}{(\lambda+\hat{V}\omega+\hat{P}\delta)^2}\right]. \quad (14)$$

The proof is based on rephrasing the minimisation of the risk as a constrained optimisation problem for which we develop a GAMP algorithm. This algorithm, upon convergence, minimises the risk and has the advantage that his performance can be asymptotically described by a low dimensional description called state evolution. The details are provided in Appendices A and B.

The separation of the equations into two parts, one depending on loss function and the other one on regularisation is common to many high-dimensional analyses of convex estimation problems. The difference is that the adversarial setting introduces two additional parameter $P, \hat{P}$ without changing this structure.

The parameters m, q, P and A are the values to which the following quantities concentrate in high-dimension

$$m = \mathbb{E}_\mathcal{D}\left[\frac{1}{d}\boldsymbol{\theta}_0^\top\boldsymbol{\Sigma_x}\hat{\boldsymbol{\theta}}\right], \quad q = \mathbb{E}_\mathcal{D}\left[\frac{1}{d}\hat{\boldsymbol{\theta}}^\top\boldsymbol{\Sigma_x}\hat{\boldsymbol{\theta}}\right],$$

$$P = \mathbb{E}_\mathcal{D}\left[\frac{1}{d}\hat{\boldsymbol{\theta}}^\top\boldsymbol{\Sigma_\delta}\hat{\boldsymbol{\theta}}\right], \quad A = \mathbb{E}_\mathcal{D}\left[\frac{1}{d}\hat{\boldsymbol{\theta}}^\top\boldsymbol{\Sigma_v}\hat{\boldsymbol{\theta}}\right]. \quad (15)$$

These order parameters have the following interpretations m measures the alignment between the estimator and the true parameter in the geometry of

the data, q quantifies the magnitude of the estimator in the same geometry, while P and A measure the defence and attack strengths respectively in their corresponding geometries.

The previous quantities, because of the average over the dataset of $\hat{\boldsymbol{\theta}}$, are not directly accessible if not after a d dimensional minimisation. We rephrased the problem into finding the solution of a low-dimensional system of coupled equations which can be solved efficiently.

4 TRADE-OFFS IN THE LARGE SAMPLE COMPLEXITY REGIME

Introduction: Adversarial Trade-Off From the decomposition of the adversarial error we can simplify its form to get

$$
E_{\mathrm{adv}} = E_{\mathrm{gen}}(\vartheta, \mathcal{U}_{\boldsymbol{\theta}_0}) + \int_0^{\varepsilon_g \varkappa} f(\xi; \vartheta, \mathcal{U}_{\boldsymbol{\theta}_0}) \, \mathrm{d}\xi \,, \quad (16)
$$

where we introduce the variable $\vartheta = m/\sqrt{\rho q}$ and $\varkappa = \sqrt{A}/\sqrt{q}$. ϑ is the cosine of the angle between the teacher weights $\boldsymbol{\theta}_0$ and the student estimate $\hat{\boldsymbol{\theta}}$ in the geometry of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ and $\varkappa$ is the norm of $\hat{\boldsymbol{\theta}}$ under the attack matrix. The function $f(\xi; \vartheta)$ is positive $\forall \vartheta, \forall \xi \in [0, +\infty)$ and it is strictly increasing in ϑ for any fixed $\xi \in [0, +\infty)$.

On the one hand, E_{gen} is a monotonically decreasing function of ϑ. In other words, to improve generalisation error, it is best to align well with $\boldsymbol{\theta}_0$. On the other hand, E_{bnd} is an increasing function of ϑ. For a fixed attack strength the boundary error decreases choosing a $\hat{\boldsymbol{\theta}}$ that aligns less with $\boldsymbol{\theta}_0$ and more in the directions where the attack is weak. As a result, when we optimise the student vectors $\hat{\boldsymbol{\theta}}$ to a lower generalisation error, we increase the boundary error, and vice versa. To minimise the adversarial error E_{adv}, we must find a balance between these competing objectives. This behaviour is common for boundary based classifiers (Tanay and Griffin, 2016).

4.1 Building Non-Robust, but Useful Features

We notice that the values for E_{gen} and E_{bnd} change by varying the usefulness and robustness of the features for fixed types of attacks. Intuitively, we have that the more usefulness one has the less generalisation error one makes, indeed we can write a lower

bound for the generalisation error

$$
E_{\mathrm{gen}} \geq \frac{1}{\pi} \arccos\left(\sqrt{\frac{\pi}{2\rho}} \mathcal{U}_{\boldsymbol{\theta}_0} \right). \quad (17)
$$

We note that ρ and $\mathcal{U}_{\boldsymbol{\theta}_0}$ only depend on $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ and $\boldsymbol{\theta}_0$.

Robustness only affects the boundary error. High robustness implies less sensibility to adversarial attacks: robust features have less samples within an attack range of the student decision boundary. The highest value that the boundary error can achieve is limited by both the robustness and the usefulness as

$$
E_{\mathrm{bnd}} \leq 2\mathrm{T}\left(\varepsilon_g \mathcal{A}\mathcal{B}, \mathcal{A}^{-1}\right) - \frac{1}{\pi} \arctan\left(\mathcal{A}^{-1}\right)
$$
$$
- \frac{1}{\pi} \mathrm{erf}\left(\frac{\varepsilon_g \mathcal{B}}{\sqrt{2}}\right) \mathrm{erfc}\left(\frac{\varepsilon_g \mathcal{A}\mathcal{B}}{\sqrt{2}}\right), \quad (18)
$$

where $\mathcal{B} = \max_i \sqrt{(\boldsymbol{\Sigma}_{\boldsymbol{v}})_{ii}/(\boldsymbol{\Sigma}_{\boldsymbol{x}})_{ii}}$, $\mathcal{A} = \sqrt{\pi}\mathcal{U}_{\boldsymbol{\theta}_0}/\sqrt{2\rho}$ and T is the Owen function. This previous bound is a decreasing function of the robustness.

We are particularly interested in studying the effects of the interplay between E_{gen} and E_{bnd} in the large sample complexity regime. We show that the adversarial error always goes to a constant in the high α regime, *i.e.* $\lim_{\alpha \to \infty} \partial_\alpha E_{\mathrm{adv}} = 0$ where the limit $\alpha \to \infty$ is taken after the high-dimensional proportional limit. Additionally, we show that the constant to which E_{adv} converges, can be zero only if $\tau = \varepsilon_g = 0$. The details are provided in Appendix C.2.

One could expect that by performing a correct cross validation of the hyper parameters ε_t and $\boldsymbol{\Sigma}_{\boldsymbol{\delta}}$, one could achieve an improvement in the value of E_{adv} in the regime of data abundance. We show that this is not the case for too simple data models.

Proposition 4.1. *Under the same setting as Theorem 3.1 and considering a BFM with a single type of feature, i.e. $k = 1$ one has that $\forall \varepsilon_g, \varepsilon_t \geq 0$ for α big enough exist two positive numbers M_1, M_2 such that*

$$
|E_{\mathrm{adv}}(\varepsilon_g, \varepsilon_t) - E_{\mathrm{adv}}(\varepsilon_g, \varepsilon_t = 0)| < M_1/\alpha \,,
$$
$$
|E_{\mathrm{gen}}(\varepsilon_t) - E_{\mathrm{gen}}(\varepsilon_t = 0)| < M_2/\alpha \,, \quad (19)
$$

where $E_{\mathrm{adv}}(\varepsilon_g, \varepsilon_t)$ and $E_{\mathrm{gen}}(\varepsilon_t)$ define the adversarial and generalisation error of $\hat{\boldsymbol{\theta}}$ trained with ε_t and evaluated for ε_g.

The proof is based on the asymptotic expansion of the result of Theorem 3.1 and it is presented in Appendix C.3. Thus we proved that in the large sample complexity regime, there is no benefit in adversarial training and indeed the values of $E_{\mathrm{adv}}, E_{\mathrm{gen}}$ are universal for any ε_t chosen in training. Note that
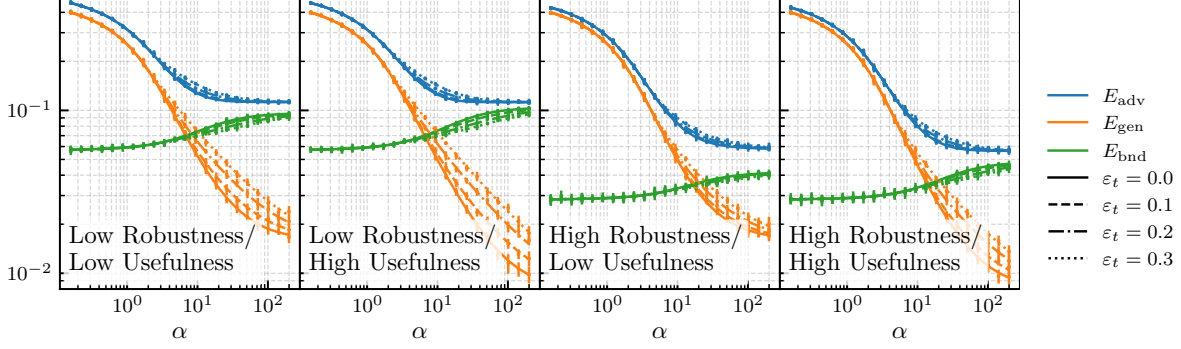
Figure 1: Error metrics as a function of the sample complexity α for different combinations of high/low robustness and high/low usefulness. We see a good agreement between the theory (lines) and finite size simulations (error bars) already for $d = 1000$. Settings in Appendix H.1.

the previous proposition also covers the setting of Javanmard et al. (2020).

Figure 1 shows the dependency of E_{gen}, E_{bnd} and E_{adv} as a function of the sample complexity α for the different combination of usefulness and robustness. We see that the plateau's value of the generalisation error mainly depends on usefulness (higher usefulness/lower plateau) while the one of boundary error mainly depends on robustness (higher robustness/lower plateau). Additionally, we also see that the effect of adversarial training (distance between curves) gets smaller as α increases.

4.2 Directional Defences & Structured Data

After the analysis of previous section, we proceed to study the effects that defending different features through the choice of $\mathbf{\Sigma_\delta}$ has on adversarial performances. We consider the following three defence strategies: *defending the robust features, uniformly defend all the features* and *defend the non-robust features*.

Figure 2 (Left) presents the comparison between the three different defence matrices for a SWFM with one robust block of features and one not. The different normalised defence matrices $\mathbf{\Sigma_\delta}$ change the relative protection between the two blocks of features, while $\mathbf{\Sigma_v} = \mathbb{1}$ for all cases. In the high α region of the curves we see that the more we protect the non robust features the more E_{gen} increases for the same value of ε_t. Simultaneously we have that the more we protect non-robust features the more E_{bnd} decreases. The value of E_{adv} does not always decrease or increase by protecting more the non-robust features, hinting at the trade off between E_{gen} and E_{bnd}.

A similar behaviour can be found for datasets like

CIFAR10 (Alex, 2009) and FashionMNIST (Xiao et al., 2017) as we explore in Appendix G. We find a choice of $\mathbf{\Sigma_\delta}$ with entries proportional to the inverse of the eigenvalue and uniform entries effective.

We are able to provide an analytical description of this phenomena under the data model considered.

Proposition 4.2. *Consider the SWFM defined in Section 2.2 where the defence matrix is $\mathbf{\Sigma_\delta} = \mathrm{blockdiag}\,((\Delta_1 + \delta_1 \varrho)\mathbb{1}_{d_1}, (\Delta_2 + \delta_2 \varrho)\mathbb{1}_{d_2})$, with ϱ the parameter that makes the defence matrix change. Assume also that $\psi_1 > \psi_2$, $\Delta_2\psi_1 \geq \Delta_1\psi_2$ and $\Upsilon_i = 1$.[2] In the $\alpha \to \infty$ (taken after the $n, d \to \infty$) there exists $\kappa > 0$ such that $\forall \delta_1 > \kappa$, $\delta_2 = -\delta_1$ one has that*

$$
\begin{aligned}
E_{\mathrm{bnd}}(\varrho) &= E_{\mathrm{bnd}}^0 + E_{\mathrm{bnd}}^1 \varrho + \mathcal{O}(\varrho^2)\,, \\
E_{\mathrm{gen}}(\varrho) &= E_{\mathrm{gen}}^0 + E_{\mathrm{gen}}^1 \varrho + \mathcal{O}(\varrho^2)\,,
\end{aligned}
\tag{20}
$$

where $E_{\mathrm{gen}}^1 > 0$, $E_{\mathrm{bnd}}^1 < 0$ and $E_{\mathrm{bnd}}^0, E_{\mathrm{gen}}^0$ are the errors when $\varrho = 0$. Additionally, this leads to an improved value of E_{adv} at order ϱ iff the following condition is satisfied

$$
\frac{\varepsilon_g}{\sqrt{2}} \operatorname{erfc}\left(-\frac{\vartheta_0 u_0 \varepsilon_g}{\sqrt{2 - 2\vartheta_0^2}}\right) < \frac{e^{-\frac{\vartheta_0^2 u_0^2 \varepsilon_g^2}{2(1-\vartheta_0^2)}}}{\sqrt{\pi}\sqrt{1-\vartheta_0^2}}\,,
\tag{21}
$$

where $\vartheta_0 = m_0/\sqrt{\rho q_0}$ and $u_0 = \sqrt{A_0}/\sqrt{q_0}$ the solution of the problem with $\varrho = 0$. Notice that for ε_g small enough this condition is always verified.

We prove this proposition in Appendix C.4 by expanding in ϱ the large α equations of Theorem 3.1.

A similar phenomenology of protection of weak features can be seen for a power-law BFM in Fig. 2

---

[2]This assumption corresponds to saying that the first set of component is more robust than the second and that the defence's effect is greater in the more robust subspace or at least equal between the two. Additionally the attacker does not distinguish the two subspaces.
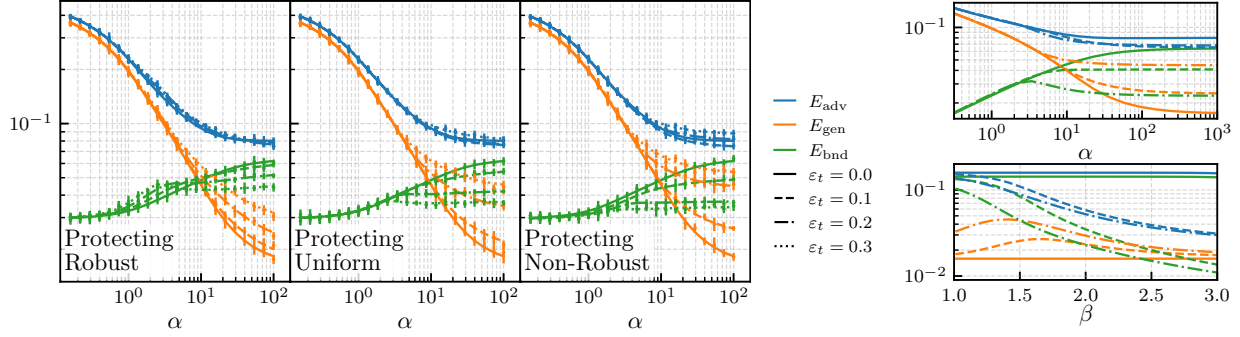
Figure 2: **(Left)** Error metric as a function of sample complexity for different choices of $\boldsymbol{\Sigma_\delta}$. The model considered is a SWFM: both blocks have the same usefulness and they differ in robustness. Each defence strategy leads to a different asymptotic values in the large sample complexity regime for the same values of ε_t. **(Right, Top)** Error metric as a function of α for a power-law BFM with $\boldsymbol{\Sigma_v} = \boldsymbol{\Sigma_\delta} = \mathbf{1}$. **(Right, Bottom)** Error metrics as a function of the power law exponent of the data β, i.e. $(\boldsymbol{\Sigma_x})_{ii} = i^{-\beta}$ and for fixed $\boldsymbol{\Sigma_v} = \boldsymbol{\Sigma_\delta} = \mathbf{1}$. We fixed the same usefulness for all values of β. A higher β increases the percentage of non-robust features increasing the number of features protected by $\boldsymbol{\Sigma_\delta}$. Settings in Appendix H.2.

(Right, Top). Adversarial training ($\varepsilon_t > 0$) decreases E_{bnd} and increases E_{gen} in the high sample complexity regime, while the adversarial error does not always decrease.

In Fig. 2 (Right, Bottom) we present the behaviour of $(E_{\text{adv}}, E_{\text{gen}}, E_{\text{bnd}})$ as a function of the power law exponent β of $\boldsymbol{\Sigma_x}$. Heavier tails (lower β) increase the total number of robust features and thus more features are not sensible to the attack. Conversely, weaker tails (bigger β) reduce the number of robust features.

To summarise, we see that the uniform defence strategy ($\boldsymbol{\Sigma_\delta} = \mathbb{1}$) performs well as it successfully defends the non-robust features, which are naturally prioritised when acting on all of them equally. The robust features are less affected by the perturbation as they have a bigger margin.

5 ADDITIONAL EXPLORATIONS

Tradeoff directions and innocuous directions We now investigate the effect that different types of geometries have on the trade-off between E_{gen} and E_{bnd}. Depending on the attack geometry $\boldsymbol{\Sigma_v}$ one can choose different defence geometries $\boldsymbol{\Sigma_\delta}$ and ask if and for which, protection without trade-off is possible. Any attack matrix $\boldsymbol{\Sigma_v}$ eigenvalues can be split into directions orthogonal to the teacher and directions aligned with the teacher. Figure 3 (Left) considers the effect of the adversarial training strength ε_t on the errors for different choices of matrices $\boldsymbol{\Sigma_\delta} = \boldsymbol{\Sigma_v}$. In the the top we consider matrices whose biggest eigenvalues are orthogonal to the teacher vector and

in the bottom one matrices where there is a leading eigenvector in the direction of the teacher.

We start by noticing that there is a qualitative difference between the two cases. If the attack focuses on features not important for the learning task (orthogonal to the teacher vector) the effect of the attack can be neutralised by choosing an ε_t big enough. In this case, even if the features are just orthogonal on average, we see that E_{adv} is decreasing as ε_t decreases and for ε_t big enough we have that $E_{\text{adv}} = E_{\text{gen}}$ with $E_{\text{bnd}} = 0$. The features that are important for the learning task (aligned with $\boldsymbol{\theta}_0$) are the ones causing the trade-off. In this case there exists a single value of ε_t such that is minimal and thus cross validation over this hyperparmeter is necessary to obtain optimal performances.

Interpreting Adversarial Training as a Data-Dependent Regularisation Recently, Ribeiro et al. (2024) show that for certain problems adversarial training can be exactly expressed as a data dependent regularisation. We investigate this claim within the framework of our model. In Fig. 3 (Center) we show the generalisation error as a function of sample complexity for a BFM where each matrix is chosen as the identity and λ as been optimised to obtain minimal E_{gen}. We see that the performance still depend on the value ε_t used during training. In this sense there is a difference between two $\hat{\boldsymbol{\theta}}$ obtained for different values of ε_t that cannot be eliminated by a careful choice of the regularisation strength.
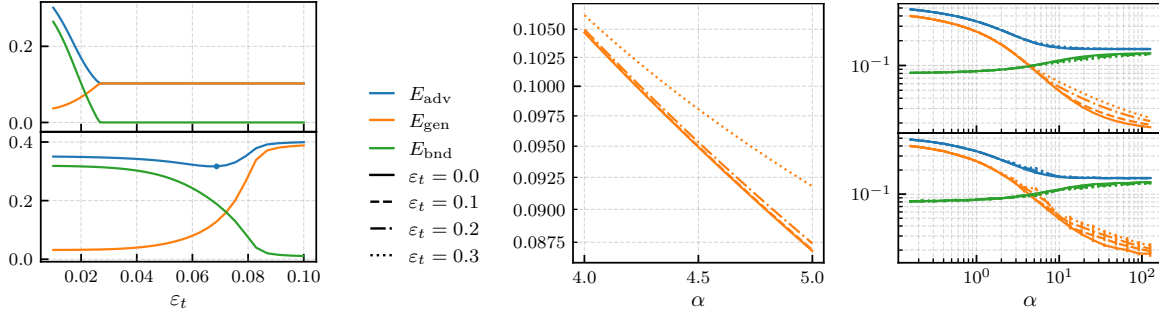
In Appendix D we show how, from a careful series ex-

Figure 3: **(Left)** We show the adversarial errors as a function of the adversarial defence strength in the case $\alpha \gg 1$. *(Top)* Defending features that are on average orthogonal to the teacher vector. *(Bottom)* Defending the features that are on average parallel to the teacher vector. **(Center)** Adversarial training is not just an ℓ_2-regularisation, we show the value of the generalisation error E_{gen} as a function of the sample complexity where the parameter λ is chosen such that it minimises the generalisation error. **(Right)** Learning curves for *(Top)* adversarial training in Eq. (5) and *(Bottom)* the equivalent problem in Eq. (22). Settings in Appendix H.3.

pansion around the minimiser, one can approximate the minimisation problem in Eq. (5) with another minimisation problem without adversarial attacks but with explicit regularisation. The form for the approximate loss in the case of small ε_t is

$$\sum_{i=1}^n g\left(y_i \frac{\boldsymbol{\theta}^\top \boldsymbol{x}_i}{\sqrt{d}}\right) + \tilde{\lambda}_1 \sqrt{\boldsymbol{\theta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\delta}} \boldsymbol{\theta}} + \tilde{\lambda}_2 \boldsymbol{\theta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\delta}} \boldsymbol{\theta} \quad (22)$$

where $\tilde{\lambda}_1$ and $\tilde{\lambda}_2$ depend on the model's parameters and perturbed margins of the points that shift sign under perturbation. In Fig. 3 (Right) we compare the generalisation performances obtained by minimising the true adversarial problem at the top with the performances of minimising the approximate problem. We see a qualitative match between the two even if the problem in Eq. (22) is less numerically stable.

The effective regularisation is equivalent to a directional $\sqrt{\ell_2} + \ell_2$ regularisation. The $\sqrt{\ell_2}$ term indicates a regularisation linearly proportional to the norm of the student vector but that does not favour sparsity as ℓ_1 regularisation.

Acknowledgements

References

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial Examples Are Not Bugs, They Are Features. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/e2c420d928d4bf8ce0ff2ec19b371514-Abstract.html.

Svetlana Pavlitska, Nico Lambing, and J. Marius Zöllner. Adversarial attacks on traffic sign recognition: A survey. In *2023 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME), Tenerife, Spain, 19-21 July 2023*, page 1–6. Institute of Electrical and Electronics Engineers (IEEE), 2023. ISBN 979-83-503-2297-2. doi: 10.1109/ICECCME57830.2023.10252727. 46.23.03; LK 01.

Adel Javanmard, Mahdi Soltanolkotabi, and Hamed Hassani. Precise tradeoffs in adversarial training for linear regression. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, vol-

ume 125 of *Proceedings of Machine Learning Research*, pages 2034–2078. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/javanmard20a.html.

Hossein Taheri, Ramtin Pedarsani, and Christos Thrampoulidis. Asymptotic behavior of adversarial training in binary linear classification. *IEEE Trans. Neural Netw. Learn. Syst.*, PP, July 2023.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, pages 372–387. IEEE, 2016.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially Robust Generalization Requires More Data. *Advances in neural information processing systems*, 31, 2018.

Nikolaos Tsilivis, Jingtong Su, and Julia Kempe. Can we achieve robustness from data alone? *arXiv preprint arXiv:2207.11727*, 2022.

Nikolaos Tsilivis and Julia Kempe. What can the neural tangent kernel tell us about adversarial robustness? *Advances in Neural Information Processing Systems*, 35:18116–18130, 2022.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SyxAb30cY7.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019.

Arun Sai Suggala, Adarsh Prasad, Vaishnavh Nagarajan, and Pradeep Ravikumar. Revisiting adversarial risk. In *The 22nd International Conference on*

Artificial Intelligence and Statistics*, pages 2331–2339. PMLR, 2019.

Arjun Nitin Bhagoji, Daniel Cullina, and Prateek Mittal. Lower bounds on adversarial robustness from optimal transport. *Advances in Neural Information Processing Systems*, 32, 2019.

Chen Dan, Yuting Wei, and Pradeep Ravikumar. Sharp statistical guaratees for adversarially robust gaussian classification. In *International Conference on Machine Learning*, pages 2345–2355. PMLR, 2020.

Adel Javanmard and Mahdi Soltanolkotabi. Precise statistical analysis of classification accuracies for adversarial training. *The Annals of Statistics*, 50 (4):2127–2156, 2022.

Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John Duchi, and Percy Liang. Understanding and mitigating the tradeoff between robustness and accuracy. proceedings of machine learning research. International Conference on Machine Learning, PMLR, 2020.

Kevin Roth, Yannic Kilcher, and Thomas Hofmann. Adversarial training is a form of data-dependent operator norm regularization. *Advances in Neural Information Processing Systems*, 33:14973–14985, 2020.

Antonio Ribeiro, Dave Zachariah, Francis Bach, and Thomas Schön. Regularization properties of adversarially-trained linear regression. *Advances in Neural Information Processing Systems*, 36, 2024.

Matteo Vilucchio, Nikolaos Tsilivis, Bruno Loureiro, and Julia Kempe. On the geometry of regularization in adversarial training: High-dimensional asymptotics and generalization bounds. *arXiv preprint arXiv:2410.16073*, 2024a.

Dongyoon Yang, Insung Kong, and Yongdai Kim. Improving adversarial robustness by putting more regularizations on less robust samples. In *International Conference on Machine Learning*, pages 39331–39348. PMLR, 2023.

Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. The gaussian min-max theorem in the presence of convexity. *arXiv preprint arXiv:1408.4837*, 2014.

Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. Regularized Linear Regression: A precise analysis of the estimation error. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 1683–1709, Paris, France, 03–06 Jul

2015. PMLR. URL https://proceedings.mlr.press/v40/Thrampoulidis15.html.

Pragya Sur and Emmanuel J. Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525, 2019. doi: 10.1073/pnas.1810420116. URL https://www.pnas.org/doi/abs/10.1073/pnas.1810420116.

Oussama Dhifallah and Yue M Lu. A precise performance analysis of learning with random features. *arXiv preprint arXiv:2008.11904*, 2020a.

Derek Bean, Peter J. Bickel, Noureddine El Karoui, and Bin Yu. Optimal m-estimation in high-dimensional regression. *Proceedings of the National Academy of Sciences*, 110(36):14563–14568, 2013. doi: 10.1073/pnas.1307845110. URL https://www.pnas.org/doi/abs/10.1073/pnas.1307845110.

Xiaoyi Mai, Zhenyu Liao, and Romain Couillet. A large scale analysis of logistic regression: Asymptotic performance and new insights. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3357–3361, 2019. doi: 10.1109/ICASSP.2019.8683376.

Zhenyu Liao, Romain Couillet, and Michael W Mahoney. A random matrix analysis of random fourier features: beyond the gaussian kernel, a precise phase transition, and the corresponding double descent. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 13939–13950. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/a03fa30821986dff10fc66647c84c9c3-Paper.pdf.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75 (4):667–766, 2022.

Lechao Xiao, Hong Hu, Theodor Misiakiewicz, Yue Lu, and Jeffrey Pennington. Precise learning curves and higher-order scalings for dot-product kernel regression. *Advances in Neural Information Processing Systems*, 35:4558–4570, 2022.

Dominik Schröder, Hugo Cui, Daniil Dmitriev, and Bruno Loureiro. Deterministic equivalent and error universality of deep random features learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 30285–30320. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/schroder23a.html.

Benjamin Aubin, Florent Krzakala, Lu Yue, and Lenka Zdeborová. Generalization error in high-dimensional perceptrons: Approaching bayes error with convex optimization. volume 33 of *Advances in Neural Information Processing Systems. 33*, pages 12199–12210. Curran Associates, Inc., 2020. URL http://infoscience.epfl.ch/record/283744.

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

Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Generalisation error in learning with random features and the hidden manifold model. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124013, dec 2021. doi: 10.1088/1742-5468/ac3ae6. URL https://dx.doi.org/10.1088/1742-5468/ac3ae6.

Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1024–1034. PMLR, 13–18 Jul 2020.

Matteo Vilucchio, Emanuele Troiani, Vittorio Erba, and Florent Krzakala. Asymptotic characterisation of the performance of robust linear regression in the presence of outliers. In *International Conference on Artificial Intelligence and Statistics*, pages 811–819. PMLR, 2024b.

Koki Okajima, Xiangming Meng, Takashi Takahashi, and Yoshiyuki Kabashima. Average case analysis of lasso under ultra sparse conditions. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 11317–11330. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/okajima23a.html.

Urte Adomaityte, Gabriele Sicuro, and Pierpaolo

Vivo. Classification of Heavy-tailed Features in High Dimensions: a Superstatistical Approach, October 2023a. URL http://arxiv.org/abs/2304.02912. arXiv:2304.02912 [cond-mat, stat].

Urte Adomaityte, Leonardo Defilippis, Bruno Loureiro, and Gabriele Sicuro. High-dimensional robust regression under heavy-tailed data: Asymptotics and universality. *arXiv preprint arXiv:2309.16476*, 2023b.

Bruno Loureiro, Cédric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Learning curves of generic features maps for realistic datasets with a teacher-student model. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(11): 114001, nov 2022a. doi: 10.1088/1742-5468/ac9825. URL https://dx.doi.org/10.1088/1742-5468/ac9825.

Sundeep Rangan. Generalized approximate message passing for estimation with random linear mixing. In *2011 IEEE International Symposium on Information Theory Proceedings*, pages 2168–2172, 2011. doi: 10.1109/ISIT.2011.6033942.

Adel Javanmard and Andrea Montanari. State evolution for general approximate message passing algorithms, with applications to spatial coupling. *Information and Inference: A Journal of the IMA*, 2(2):115–144, 12 2013. ISSN 2049-8764. doi: 10.1093/imaiai/iat004. URL https://doi.org/10.1093/imaiai/iat004.

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

Bruno Loureiro, Gabriele Sicuro, Cedric Gerbelot, Alessandro Pacco, Florent Krzakala, and Lenka Zdeborová. Learning gaussian mixtures with generalized linear models: Precise asymptotics in high-dimensions. Advances in Neural Information Processing Systems. 34, 2021. URL http://infoscience.epfl.ch/record/291929.

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

2022b. URL https://proceedings.mlr.press/v162/loureiro22a.html.

Cédric Gerbelot and Raphaël Berthier. Graph-based approximate message passing iteration. *Information and Inference: A Journal of the IMA*, 12 (4):2562–2628, 09 2023. ISSN 2049-8772. doi: 10.1093/imaiai/iaad020. URL https://doi.org/10.1093/imaiai/iaad020.

Matteo Vilucchio, Yatin Dandi, Cedric Gerbelot, and Florent Krzakala. Asymptotics of non-convex generalized linear models in high-dimensions: A proof of the replica formula. *arXiv preprint arXiv:2502.20003*, 2025.

Lucas Andry Clarté, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Theoretical characterization of uncertainty in high-dimensional linear classification. *Machine Learning: Science and Technology*, 4(2):025029, jun 2023a. doi: 10.1088/2632-2153/acd749. URL https://dx.doi.org/10.1088/2632-2153/acd749.

Francis Bach. *Learning theory from first principles*. MIT press, 2024.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of statistics*, 50(2):949, 2022.

Lenaic Chizat and Francis Bach. Lazy training in differentiable programming. *Advances in Neural Information Processing Systems*, 32, 2019.

Martin J Wainwright and Eero Simoncelli. Scale Mixtures of Gaussians and the Statistics of Natural Images. In *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999. URL https://proceedings.neurips.cc/paper_files/paper/1999/hash/6a5dfac4be1502501489fc0f5a24b667-Abstract.html.

Eero P Simoncelli and Bruno A Olshausen. Natural image statistics and neural representation. *Annual review of neuroscience*, 24(1):1193–1216, 2001.

Oussama Dhifallah and Yue M Lu. A precise performance analysis of learning with random features. *arXiv preprint arXiv:2008.11904*, 2020b.

Thomas Tanay and Lewis Griffin. A boundary tilting persepective on the phenomenon of adversarial examples. *arXiv preprint arXiv:1608.07690*, 2016.

Krizhevsky Alex. Learning multiple layers of features from tiny images. *https://www. cs. toronto. edu/kriz/learning-features-2009-TR. pdf*, 2009.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

David L. Donoho, Arian Maleki, and Andrea Montanari. Message passing algorithms for compressed sensing: I. motivation and construction. In *2010 IEEE Information Theory Workshop on Information Theory (ITW 2010, Cairo)*, pages 1–5, 2010. doi: 10.1109/ITWKSPS.2010.5503193.

Florent Krzakala, Marc Mézard, Francois Sausset, Yifan Sun, and Lenka Zdeborová. Probabilistic reconstruction in compressed sensing: algorithms, phase diagrams, and threshold achieving matrices. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(08):P08009, August 2012. ISSN 1742-5468. doi: 10.1088/1742-5468/2012/08/p08009. URL http://dx.doi.org/10.1088/1742-5468/2012/08/P08009.

Léo Miolane and Andrea Montanari. The distribution of the lasso: Uniform control over sparse balls and adaptive parameter tuning. *The Annals of Statistics*, 49(4), 2021.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. doi: 10.1017/CBO9780511804441.

Neal Parikh and Stephen Boyd. Proximal algorithms. *Found. Trends Optim.*, 1(3):127–239, jan 2014. ISSN 2167-3888. doi: 10.1561/2400000003. URL https://doi.org/10.1561/2400000003.

Stephen J. Wright, Robert D. Nowak, and MÁrio A. T. Figueiredo. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57(7):2479–2493, 2009. doi: 10.1109/TSP.2009.2016892.

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

Sundeep Rangan, Philip Schniter, Erwin Riegler, Alyson K. Fletcher, and Volkan Cevher. Fixed points of generalized approximate message passing with arbitrary matrices. *IEEE Transactions on Information Theory*, 62(12):7464–7474, Dec 2016. ISSN 1557-9654. doi: 10.1109/TIT.2016.2619365.

Lucas Andry Clarté, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. On double-descent in uncertainty quantification in over-parametrized models. volume 206, pages 7089–7125. PMLR Proceedings of Machine Learning Research, 2023b. URL http://infoscience.epfl.ch/record/302224.

Luis M Briceño-Arias, Giovanni Chierchia, Emilie Chouzenoux, and Jean-Christophe Pesquet. A random block-coordinate douglas–rachford splitting method with low computational complexity for binary logistic regression. *Computational Optimization and Applications*, 72:707–726, 2019.

Steven George Krantz and Harold R Parks. *The implicit function theorem: history, theory, and applications*. Springer Science & Business Media, 2002.

Donald B. Owen. Tables for Computing Bivariate Normal Probabilities. *The Annals of Mathematical Statistics*, 27(4):1075 – 1090, 1956. doi: 10.1214/aoms/1177728074. URL https://doi.org/10.1214/aoms/1177728074.

Edward W. Ng and Murray Geller. A table of integrals of the error functions. *Journal of Research of the National Bureau of Standards, Section B: Mathematical Sciences*, page 1, 1969. URL https://api.semanticscholar.org/CorpusID:121798145.

Nikolai E Korotkov and Alexander N Korotkov. *Integrals related to the error function*. Chapman and Hall/CRC, 2020.

D. B. Owen. A table of normal integrals. *Communications in Statistics - Simulation and Computation*, 9(4):389–419, 1980. doi: 10.1080/03610918008812164. URL https://doi.org/10.1080/03610918008812164.

M. Mezard, G. Parisi, and M.A. Virasoro. *Spin Glass Theory And Beyond: An Introduction To The Replica Method And Its Applications*. World Scientific Lecture Notes In Physics. World Scientific Publishing Company, 1987. ISBN 9789813103917. URL https://books.google.ch/books?id=DwY8DQAAQBAJ.

Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

Checklist

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]

(b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]

(c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:

(a) Statements of the full set of assumptions of all theoretical results. [Yes]

(b) Complete proofs of all theoretical results. [Yes]

(c) Clear explanations of any assumptions. [Yes]

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]

(b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]

(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]

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

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]

(b) The license information of the assets, if applicable. [Yes]

(c) New assets either in the supplemental material or as a URL, if applicable. [Yes]

(d) Information about consent from data providers/curators. [Yes]

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [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. [Not Applicable]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Kasimir Tanner, Matteo Vilucchio, Bruno Loureiro, Florent Krzakala

Supplementary Materials

In Appendix A we present a rigorous proof of the main theoretical result introduced in the main body of the paper. This section aims to provide a derivation of the main result based on the body of literature on the use of AMP and CGMT. In Appendix B we derive the theoretical formulas for the the error metrics in terms of the overlap solutions of the fixed-point equations presented in Section 3. Again based on the previous literature we explain the derivation of these formulas and clarify explicitly how the new error metric of the class-preserving error is derived. In Appendix C we expand the fixed-point equations from Section 3 in the high sample complexity regime, offering insights into how the solutions behave as the complexity increases. The expansion provides important implications for practical applications as it studies the limiting performances that can be reached even with infinite number of data. Additionally we provide further results related to the adversarial generalisation error, including its relation to Owen's T function. In Appendix D we show how to rewrite the adversarial problem in terms of a data-dependent regularisation, providing an alternative formulation that aids in better understanding the adversarial setup and its effect on the bias it gives to the solution. In Appendix F we consider another type of attack and provide a similar high-dimensional description of it. In Appendix E we explore the relationship between the attack geometry and the defense geometry, providing geometric interpretations and insights into how different adversarial attack strategies can influence and be mitigated by corresponding defense mechanisms. In Appendix G, we demonstrate the robustness metrics and various defense strategies on the Cifar-10 Alex (2009) and FashionMNIST Xiao et al. (2017) datasets. Detailed experimental results highlight the performance and effectiveness of the proposed methods. Appendix H provides a comprehensive explanation of the figure settings used in the main text is provided here, ensuring reproducibility. In Appendix I we derive in detail the same result presented in Appendix A with the use of the statistical physics's replica method. This offers a different perspective on the problem, serving as a complementary approach to the formal proof.

A HIGH-DIMENSIONAL ASYMPTOTICS

In this section, we present a comprehensive proof of the fixed-point equations outlined in Theorem 3.1. Our approach leverages recent advancements in high-dimensional statistics and convex optimization to establish the asymptotic behavior of adversarial training in the high-dimensional limit.

The proof unfolds in three main stages

Problem Reformulation We begin by recasting the original optimization problem into an equivalent form with additional constraint variables P and $\hat{P}$. This reformulation allows us to apply techniques such as the mapping to a GAMP algorithm.

Algorithmic Interpretation We demonstrate that the reformulated problem can be solved by a Generalized Approximate Message Passing (GAMP) algorithm. This connection not only provides an algorithmic perspective but also establishes the optimality of the GAMP solution at convergence.

Asymptotic Analysis Building upon the well-established literature about high-dimensional asymptotics (Donoho et al., 2010; Krzakala et al., 2012; Miolane and Montanari, 2021; Loureiro et al., 2022a), we derive the low-dimensional asymptotics of our system. This step is crucial in bridging the gap between finite-dimensional system describing our system and the d-dimensional inital problem.

For readers well-versed in statistical physics, we provide an alternative derivation using the replica method in Appendix I.

A.1 Notations and Definitions

In this paper, we extensively employ the concepts of Moreau envelopes and proximal operators, pivotal elements in convex analysis frequently encountered in recent works on high-dimensional asymptotic of convex problems Boyd and Vandenberghe (2004); Parikh and Boyd (2014). For an in-depth analysis of their properties,

we refer the reader to the cited literature. Here, we briefly outline their definition and the main properties for context.

The Moreau envelope and the proximal operator associated to a scalar function $f : \mathbb{R} \to \mathbb{R}$ are defined as

$$\mathcal{M}_{Vf(\cdot)}(\omega) = \min_x \left[\frac{(x - \omega)^2}{2V} + f(x) \right], \quad \mathcal{P}_{Vf(\cdot)}(\omega) = \arg\min_x \left[\frac{(x - \omega)^2}{2V} + f(x) \right], \tag{23}$$

where the $(\cdot)$ indicates the variable considered if the function is of more variables. Generally one can consider the Moreau envelope or the proximal with respect to just one of the inputs of a function depending on more variables.

In the rest of the paper we will use the following properties of the Moreau and Proximal that can be found in Parikh and Boyd (2014). We will be using the *envelope theorem* which states

$$\partial_\omega \mathcal{M}_{Vg(y,\cdot)}(\omega) = V^{-1} \left(\omega - \mathcal{P}_{Vg(y,\cdot)}(\omega) \right) . \tag{24}$$

Additionally we will use the two following results

$$\mathcal{M}_{Vf(\cdot+u)}(\omega) = \mathcal{M}_{Vf(\cdot)}(\omega + u), \quad \mathcal{P}_{Vf(\cdot+u)}(\omega) = u + \mathcal{P}_{Vf(\cdot)}(\omega + u). \tag{25}$$

A.2 Approximate Message Passing

In the landscape of high-dimensional statistical inference, Approximate Message Passing (AMP) algorithms have emerged as a cornerstone for efficiently solving problems like compressed sensing Donoho et al. (2010); Krzakala et al. (2012). At their hearts AMP algorithms are iterative schemes, inspired by the ISTA algorithm Wright et al. (2009), that in addition leverage the statistical properties of high-dimensional random matrices to remove correlation at each step. A key feature of AMP algorithms is their connection with state evolution (SE), a powerful analytical tool that tracks the evolution of the AMP algorithm's performance across iterations. State evolution provides a set of deterministic equations that accurately predict the algorithm's behaviour in the limit of large system sizes, thus offering deep insights into the convergence properties and asymptotic accuracy of AMP algorithms. This method has been extensively used to understand problems like the learning of Gaussian mixtures Loureiro et al. (2021) or learning curves of ensembling methods Loureiro et al. (2022b). This kind of connection and algorithm has been also know by statistical physicist Zdeborová and Krzakala (2016).

In the case of estimation of an i.i.d. random vector observed through a linear transform followed by a component-wise, probabilistic (possibly nonlinear) measurement channel, an optimal algorithm, called generalised approximate message passing (GAMP) and its respective SE has been introduced by Rangan (2011). The main result that associates the evolution of GAMP to the SE has been proven later by Javanmard and Montanari (2013). Thus the idea is that the fixed point equations could be seen as the state evolution equations for a specific Generalised AMP algorithm that minimises the equivalent minimisation problem in Eq. (34). This way of proving the result also gives the advantage of defining an algorithm that, upon convergence, returns the vector $\hat{\boldsymbol{\theta}}$.

We will now define in general a GAMP sequence and state here the result about the general state evolution. Consider a sequence Gaussian matrices $A(n)_{ij} \sim \mathcal{N}(0, 1)$, with $i \in \{1, \ldots n\}, j \in \{1, \ldots d\}$, with i.i.d. Gaussian entries. For each $n, d \in \mathbb{N}$, consider two sequences of pseudo-Lipschitz functions

$$\{\boldsymbol{h}_t : \mathbb{R}^n \to \mathbb{R}^n\}_{t \in \mathbb{N}} \quad \{\boldsymbol{e}_t : \mathbb{R}^d \to \mathbb{R}^d\}_{t \in \mathbb{N}} \tag{26}$$

and recursively define $\boldsymbol{u}^i \in \mathbb{R}^d$ and $\boldsymbol{v}^i \in \mathbb{R}^n$ as

$$\begin{aligned} \boldsymbol{u}^{t+1} &= A^\top \boldsymbol{h}_t \left(\boldsymbol{v}^t \right) - \boldsymbol{e}_t \left(\boldsymbol{u}^t \right) \langle \boldsymbol{h}_t' \rangle^\top \\ \boldsymbol{v}^t &= A \boldsymbol{e}_t \left(\boldsymbol{u}^t \right) - \boldsymbol{h}_{t-1} \left(\boldsymbol{v}^{t-1} \right) \langle \boldsymbol{e}_t' \rangle^\top \end{aligned} \tag{27}$$

where we define the Onsager terms as

$$\langle \boldsymbol{h}_t' \rangle = \frac{1}{d} \sum_{i=1}^{n} \frac{\partial \boldsymbol{h}_t^i}{\partial \boldsymbol{v}_i} \left(\boldsymbol{v}^t \right) \in \mathbb{R} \quad \langle \boldsymbol{e}_t' \rangle = \frac{1}{d} \sum_{i=1}^{d} \frac{\partial \boldsymbol{e}_t^i}{\partial \boldsymbol{u}_i} \left(\boldsymbol{u}^t \right) \in \mathbb{R} \tag{28}$$

The previous recursive relation is defined for a suitable initial condition that has a well-defined high-dimensional limit.

For any d (and respective n) we have a series of vectors defined iteratively through Eq. (27).

We then define the state evolution from two sets of

$$Q_{t+1,s} = Q_{s,t+1} = \lim_{d \to \infty} \frac{1}{d} \mathbb{E} \left[\boldsymbol{e}_s \left(\hat{Z}^s \right)^\top \boldsymbol{e}_{t+1} \left(\hat{Z}^{t+1} \right) \right] \in \mathbb{R}$$

$$\hat{Q}_{t+1,s+1} = \hat{Q}_{s+1,t+1} = \lim_{d \to \infty} \frac{1}{d} \mathbb{E} \left[\boldsymbol{h}_s \left(Z^s \right)^\top \boldsymbol{h}_t \left(Z^t \right) \right] \in \mathbb{R} \tag{29}$$

where $\left(Z^0, \ldots, Z^{t-1} \right) \sim \mathcal{N} \left(0, \{ Q_{r,s} \}_{0 \leqslant r,s \leqslant t-1} \otimes I_n \right), \left(\hat{Z}^1, \ldots, \hat{Z}^r \right) \sim \mathcal{N} \left(0, \left\{ \hat{Q}_{r,s} \right\}_{1 \leqslant r,s \leqslant t} \otimes I_d \right)$. Where the initial values are coherent with the ones of the iterates in Eq. (27).

Theorem A.1 (Javanmard and Montanari (2013)). *In the setting of the previous paragraph, for any sequence of pseudo-Lipschitz functions $\phi_n : (\mathbb{R}^n \times \mathbb{R}^d)^t \to \mathbb{R}$ we have that*

$$\phi_n \left(\boldsymbol{u}^0, \boldsymbol{v}^0, \boldsymbol{u}^1, \boldsymbol{v}^1, \ldots, \boldsymbol{v}^{t-1}, \boldsymbol{u}^t \right) \overset{\mathrm{P}}{\sim} \mathbb{E} \left[\phi_n \left(\boldsymbol{u}^0, Z^0, \hat{Z}^1, Z^1, \ldots, Z^{t-1}, \hat{Z}^t \right) \right] \tag{30}$$

where $\left(Z^0, \ldots, Z^{t-1} \right) \sim \mathcal{N} \left(0, \{ Q_{r,s} \}_{0 < r,s \leqslant t-1} \otimes I_n \right), \left(\hat{Z}^1, \ldots, \hat{Z}^t \right) \sim \mathcal{N} \left(0, \left\{ \hat{Q}_{r,s} \right\}_{1 \leq r,s \leqslant t} \otimes I_n \right).$

The previous theorem states the fact that for the limit of any pseudo lipshitz function of the iterates, and thus of the limiting value, one has that its behaviour concentrates to the expectation of gaussian variables with specific covariances.

A.3   Assumptions

For our results to hold we will need all the assumptions required in Loureiro et al. (2022a) for the application of their results on the saddle point equations. Our assumptions are thus all of the ones contained in (Loureiro et al., 2022a, Appendix B.1). Additionally we need

(A1) The attack and defence geometry $\boldsymbol{\Sigma_\delta}$ and $\boldsymbol{\Sigma_v}$ should be $\boldsymbol{\Sigma_v}, \boldsymbol{\Sigma_\delta} \succ 0$. The spectral distributions of the matrices $\boldsymbol{\Sigma_\delta}, \boldsymbol{\Sigma_v}$ converge to distributions such that the overlaps

$$P = \frac{1}{d} \mathbb{E}_{\mathcal{D}} \left[\hat{\boldsymbol{\theta}}^\top \boldsymbol{\Sigma_\delta} \hat{\boldsymbol{\theta}} \right], \quad A = \frac{1}{d} \mathbb{E}_{\mathcal{D}} \left[\hat{\boldsymbol{\theta}}^\top \boldsymbol{\Sigma_v} \hat{\boldsymbol{\theta}} \right], \quad F = \frac{1}{d} \mathbb{E}_{\mathcal{D}} \left[\boldsymbol{\theta}_0^\top \boldsymbol{\Sigma_v} \hat{\boldsymbol{\theta}} \right], \tag{31}$$

are well-defined. Additionally, the maximum singular values of them are bounded with high probability when $n, p \to \infty$.

(A2) The values $\zeta_i = \mathrm{diag}(\mathrm{S}\boldsymbol{\Sigma_\delta}\mathrm{S}^\top)_i$, $v_i = \mathrm{diag}(\mathrm{S}\boldsymbol{\Sigma_v}\mathrm{S}^\top)_i$ and $\mathbf{f}_i = (\mathrm{S}\boldsymbol{\Sigma_v}^\top \boldsymbol{\theta}_0 / \sqrt{\rho})_i$, where additionally $\rho = \boldsymbol{\theta}_0^\top \boldsymbol{\Sigma_x} \boldsymbol{\theta}_0 / d$, $a = \boldsymbol{\theta}_0^\top \boldsymbol{\Sigma_v} \boldsymbol{\theta}_0 / d$ and $\bar{\boldsymbol{\theta}} = \mathrm{S}\boldsymbol{\Sigma_x}^\top \boldsymbol{\theta}_0 / \sqrt{\rho}$ should have jointly a well defined limit when $n, p \to \infty$. More formally

$$\frac{1}{d} \sum_{i=1}^{d} \delta(\omega - \omega_i) \delta(\bar{\theta} - \bar{\theta}_i) \delta(\zeta - \zeta_i) \delta(v - v_i) \delta(f - f_i) \to \mu \tag{32}$$

should converge for $n, p \to \infty$.

(A3) The choice of the matrices $\boldsymbol{\Sigma_\delta}$ and $\boldsymbol{\Sigma_v}$ should be independent of the teacher vector $\boldsymbol{\theta}_0$.

A.4   Reformulation of the problem

Our analysis begins with the adversarial training problem introduced in Eq. (4) of the main text. For clarity, we restate it here

$$\sum_{i=1}^{n} \max_{\|\boldsymbol{\delta}_i\|_{\boldsymbol{\Sigma}_{\boldsymbol{\delta}}^{-1}} \leq \varepsilon_t} g\left(y_i \frac{\boldsymbol{\theta}^\top (\boldsymbol{x}_i + \boldsymbol{\delta}_i)}{\sqrt{d}}\right) + r(\boldsymbol{\theta}), \tag{33}$$

where we remind that the regularisation function $r : \mathbb{R}^d \to \mathbb{R}$ is convex and the loss function $g : \mathbb{R} \to \mathbb{R}$ is non-increasing, meaning that $x_1 \leq x_2$ implies $g(x_1) \geq g(x_2)$.

The non-increasing property of g allows us to simplify the inner maximization, leading to

$$\sum_{i=1}^{n} g\left(y_i \frac{\boldsymbol{\theta}^\top \boldsymbol{x}_i}{\sqrt{d}} - \varepsilon_t \frac{\sqrt{\boldsymbol{\theta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\delta}} \boldsymbol{\theta}}}{\sqrt{d}}\right) + r(\boldsymbol{\theta}). \tag{34}$$

We can introduce two constrains to rewrite the minimisation as

$$\sum_{i=1}^{n} g\left(y_i, \frac{\boldsymbol{\theta}^\top \boldsymbol{x}_i}{\sqrt{d}} - y_i \varepsilon_t \sqrt{P}\right) + r(\boldsymbol{\theta}) \quad \text{such that} \quad dP = \boldsymbol{\theta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\delta}} \boldsymbol{\theta}, \tag{35}$$

The Lagrangian form of this problem reads

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{s}, P, \hat{P}) = g\left(\boldsymbol{z} - \varepsilon_t \sqrt{P} \boldsymbol{y}\right) + r(\boldsymbol{\theta}) + \boldsymbol{s}\left(\frac{1}{\sqrt{d}} \boldsymbol{X} \boldsymbol{\theta} - \boldsymbol{z}\right) + \hat{P}\left(\boldsymbol{\theta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\delta}} \boldsymbol{\theta} - dP\right) \tag{36}$$

where we have simplified the notations by introducing

$$g\left(\boldsymbol{z} - \varepsilon_t \sqrt{P} \boldsymbol{y}\right) = \sum_{i=1}^{n} g\left(y_i, z_i - \varepsilon_t \sqrt{P}\right), \tag{37}$$

and defined the feature matrix $\boldsymbol{X} \in \mathbb{R}^{n \times d}$.

A.5   Fixed Point Equations for the variables

We can decompose this problem into separate optimizations over $\boldsymbol{\theta}$, $\boldsymbol{z}$, and P To obtain the equations for $P, \hat{P}$ we go back to the complete optimisation problem

$$\begin{aligned}
\sup_{\boldsymbol{s}, \hat{P}} \inf_{\boldsymbol{\theta}, \boldsymbol{z}, P} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{s}, P, \hat{P}) &= \sup_{\boldsymbol{s}, \hat{P}} \left[\inf_{\boldsymbol{\theta}} \left[r(\boldsymbol{\theta}) + \hat{P} \boldsymbol{\theta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\delta}} \boldsymbol{\theta} + \frac{1}{\sqrt{d}} \boldsymbol{s}^\top \boldsymbol{X} \boldsymbol{\theta} \right] + \inf_{\boldsymbol{z}, P} \left[g\left(\boldsymbol{z} - \varepsilon_t \sqrt{P} \boldsymbol{y}\right) - \boldsymbol{s}^\top \boldsymbol{z} - d\hat{P} P \right] \right] \\
&= \sup_{\boldsymbol{s}, \hat{P}} \left[\inf_{\boldsymbol{\theta}} \left[f(\boldsymbol{\theta}) + \hat{P} \boldsymbol{\theta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\delta}} \boldsymbol{\theta} + \frac{1}{\sqrt{d}} \boldsymbol{s}^\top \boldsymbol{X} \boldsymbol{\theta} \right] + \inf_{\boldsymbol{z}} \left[g(\boldsymbol{z}) - \boldsymbol{s}^\top \boldsymbol{z} \right] + \inf_{P} \left[\varepsilon_t \sqrt{P} \boldsymbol{s}^\top \boldsymbol{y} - d\hat{P} P \right] \right]
\end{aligned} \tag{38}$$

We also consider the gradients with respect to the new variables

$$\frac{\partial}{\partial P} = \frac{\varepsilon_t}{2\sqrt{P}} \boldsymbol{s}^\top \boldsymbol{y} - d\hat{P}, \qquad\qquad \frac{\partial}{\partial \hat{P}} = \boldsymbol{\theta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\delta}} \boldsymbol{\theta} - dP, \tag{39}$$

A.6   Generalised AMP mapping of our problem

In Algorithm 1, we present the advGAMP that will be the object of the study of this subsection. The notation $\odot$ represents the component-wise product and $\oslash$ the component-wise division. After an initialisation, we have to update the variables alternating by output channel variables and input channel variables until some convergence condition of the variables is reached.

Algorithm 1: Adversarial Generalised Approximate Message Passing (advGAMP)

Input: Matrix $\boldsymbol{X} \in \mathbb{R}^{n \times d}$, functions $r(\boldsymbol{\theta}) : \mathbb{R}^d \to \mathbb{R}, g(\boldsymbol{z}) : \mathbb{R}^n \to \mathbb{R}$.
Output: An estimate $\boldsymbol{\theta} \in \mathbb{R}^d$

1 $t \leftarrow 0$
2 Initialise $\boldsymbol{\theta}^t \in \mathbb{R}^d, \boldsymbol{\tau}_w^t \in \mathbb{R}_+^d$
3 $\boldsymbol{s}^{t-1} \leftarrow \boldsymbol{0} \in \mathbb{R}^n$
4 $\boldsymbol{F} \leftarrow \boldsymbol{X} \odot \boldsymbol{X}$
5 **repeat**
      // Output node update
6 $\boldsymbol{\tau}_\omega^t \leftarrow \boldsymbol{F}\boldsymbol{\tau}_w^t$
7 $\boldsymbol{\omega}^t \leftarrow \boldsymbol{X}\boldsymbol{\theta}^t - \boldsymbol{s}^{t-1} \odot \boldsymbol{\tau}_\omega^t$
8 $P^t \leftarrow \frac{1}{d}\boldsymbol{\theta}^t \boldsymbol{\Sigma_\delta} \boldsymbol{\theta}^t$
9 $\boldsymbol{z}^t \leftarrow \mathcal{P}_{\boldsymbol{\tau}_p^t}[g(\cdot; P^t)](\boldsymbol{\omega}^t)$
10 $\boldsymbol{\tau}_z^t \leftarrow \boldsymbol{\tau}_p^t \odot \mathcal{P}'_{\boldsymbol{\tau}_p^t}[g(\cdot; P^t)](\boldsymbol{\omega}^t)$
11 $\boldsymbol{s}^t \leftarrow (\boldsymbol{z}^t - \boldsymbol{\omega}^t) \oslash \boldsymbol{\tau}_p^t$
12 $\boldsymbol{\tau}_s^t \leftarrow (\boldsymbol{1} - \boldsymbol{\tau}_z^t \oslash \boldsymbol{\tau}_p^t) \oslash \boldsymbol{\tau}_p^t$
      // Input node update
13 $\hat{P}^t \leftarrow \frac{\varepsilon_t}{2\sqrt{P^t}} \frac{1}{d} \boldsymbol{s}^\top \boldsymbol{y}$
14 $\boldsymbol{\tau}_r^t \leftarrow \boldsymbol{1} \oslash (\boldsymbol{F}^\top \boldsymbol{\tau}_s^t)$
15 $\boldsymbol{r}^t \leftarrow \boldsymbol{\theta}^t + \boldsymbol{\tau}_r^t \odot \boldsymbol{x}^\top \boldsymbol{s}^t$
16 $\boldsymbol{\theta}^{t+1} \leftarrow \mathcal{P}_{\boldsymbol{\tau}_r^t}[r(\cdot; \hat{P}^t)](\boldsymbol{r}^t)$
17 $\boldsymbol{\tau}_x^{t+1} \leftarrow \boldsymbol{\tau}_r^t \odot \mathcal{P}'_{\boldsymbol{\tau}_r^t}[r(\cdot, \hat{P}^t)](\boldsymbol{r}^t)$
18 **until** *Termination condition*

This algorithm can be seen as a specialisation of a GAMP algorithm where the denoising function change at each iteration because dependant on the constants $P, \hat{P}$ that are updated in Lines 8 and 13.

Enforcing these constraints at each step guarantees that the $\hat{P}$ (respectively Line 13) variable will be such that they minimise the Lagrangian with respect to the primal variables. Additionally for the way the P variables are updated in Algorithm 1 (respectively Line 8) it is guaranteed that the definition is satisfied. With these considerations in mind one can prove that upon convergence the $\boldsymbol{\theta}$ found solves the problem in Eq. (34). The steps are based on interpreting the algorithm as a ADMM in a similar fashion as in Rangan et al. (2016, Theorem 1).

We are left with proving that the low dimensional equations are the same presented in the main text in the next subsection.

A.7 Mapping of Saddle point Equations

The SE equations corresponding to the GAMP algorithm defined in Rangan (2011); Rangan et al. (2016) correspond exactly the the equations in Eq. (41) with the dependence on the parameters $P, \hat{P}$. To prove the state evolution form for the new variables one can apply Theorem 1 in Javanmard and Montanari (2013) to the update for $P, \hat{P}$ in Eq. (39) and obtain the results in Eqs. (42) and (43).

To prove the form of the fixed point equations we could also map them to already proven set of fixed point equations proved in Loureiro et al. (2022a).

If we consider the values of the overlaps to be values we have that one can prove the equations for m, q, V and $\hat{m}, \hat{q}, \hat{V}$ as being a case of a specific loss for Loureiro et al. (2022a, Theorem 1). For each fixed value of $P, \hat{P}$ which will be specified afterwards. The mapping from the notation of this current paper to the other

one involves a different loss function and regularisation

$$g(z) \leftrightarrow g\left(z - \varepsilon_t \sqrt{P}\right), \qquad r(\boldsymbol{\theta}) \leftrightarrow r(\boldsymbol{\theta}) + \hat{P}\boldsymbol{\theta}^\top \boldsymbol{\Sigma_\delta}\boldsymbol{\theta}, \tag{40}$$

where the first ones are the notations used in Loureiro et al. (2022a) and the second ones the one used in this paper. For fixed values of $P, \hat{P}$ this can be seen by comparing the Lagrangian formulation in Eq. (36) to the one in Eq. (B.61) of Loureiro et al. (2022a). If we consider that the regularisation is ℓ_2 we have that we can apply the equations with the effective regularisation that is $\boldsymbol{\theta}^\top ((\lambda/2)\mathbb{1} + \hat{P}\boldsymbol{\Sigma_\delta})\boldsymbol{\theta}$, under the previously stated assumptions

$$
\begin{aligned}
\hat{m} &= \alpha\mathbb{E}_\xi\left[\int_\mathbb{R} \mathrm{d}y\, \partial_\omega \mathcal{Z}_0 f_g(\sqrt{q}\xi, P)\right] & m &= \mathbb{E}_\mu\left[\frac{\hat{m}\bar{\theta}^2}{\lambda + \hat{V}\omega + \hat{P}\delta}\right] \\
\hat{q} &= \alpha\mathbb{E}_\xi\left[\int_\mathbb{R} \mathrm{d}y\, \mathcal{Z}_0 f_g^2(\sqrt{q}\xi, P)\right] & q &= \mathbb{E}_\mu\left[\frac{\hat{m}^2\bar{\theta}^2\omega + \hat{q}\omega^2}{(\lambda + \hat{V}\omega + \hat{P}\delta)^2}\right] \\
\hat{V} &= -\alpha\mathbb{E}_\xi\left[\int_\mathbb{R} \mathrm{d}y\, \mathcal{Z}_0 \partial_\omega f_g(\sqrt{q}\xi, P)\right] & V &= \mathbb{E}_\mu\left[\frac{\omega}{\lambda + \hat{V}\omega + \hat{P}\delta}\right]
\end{aligned}
\tag{41}
$$

where we have the same definitions for $\mathcal{Z}_0$ and f_g as in the Theorem 3.1.

At optimality we want the gradients in Eq. (39) to be equal to zero. Thus these equations should be considered as equalities to zero also in the limit. To find the limiting form of these equations we would like to apply Theorem 5 of Loureiro et al. (2022a). We can start from the condition for the gradients of the dual variables. The function satisfies the assumptions of the theorem and thus can be applied. To obtain the specific form

$$P = \mathbb{E}\left[\zeta\frac{\hat{m}^2\bar{\theta}^2 + \hat{q}\omega}{(\lambda + \hat{V}\omega + \hat{P}\delta)^2}\right] \tag{42}$$

For the other two equations we can remember that at optimality the value of $\boldsymbol{s}$ is connected to the proximal operator and the function f_g. We have that these functions are pseudo-Lipschitz and thus satisfy the assumptions of the previously applied theorem. Thus the optimality conditions read in the limit are

$$\hat{P} = \alpha\frac{\varepsilon_t}{2\sqrt{P}}\mathbb{E}_\xi\left[\int \mathrm{d}y\, \mathcal{Z}_0 y f_g(\sqrt{q}\xi, P, \varepsilon_t)\right], \tag{43}$$

thus proving the set of equations.

A.8 Form for the overlaps A and F

Similarly to before we want to find the high dimensional form for

$$A = \frac{1}{d}\hat{\boldsymbol{\theta}}^\top \boldsymbol{\Sigma_v}\hat{\boldsymbol{\theta}}, \quad F = \frac{1}{d}\hat{\boldsymbol{\theta}}^\top \boldsymbol{\Sigma_v}\boldsymbol{\theta}_0 \tag{44}$$

Again we can leverage Loureiro et al. (2022a, Theorem 5) or Javanmard and Montanari (2013, Theorem 1) to obtain in both cases

$$A = \mathbb{E}_\mu\left[v\frac{\hat{m}^2\bar{\theta}^2\omega + \hat{q}\omega^2}{(\lambda + \hat{V}\omega + \hat{P}\delta)^2}\right], \quad F = \mathbb{E}_\mu\left[\frac{\hat{m}f\bar{\theta}}{\lambda + \hat{V}\omega + \hat{P}\delta}\right] \tag{45}$$

as explained before.

A.9 Interpretation of the Result

The parameters m, q, P, N, A and F appearing in the previous sections have a simple interpretation, they are the values around which the teacher-student and student-student overlaps concentrate in high dimension

$$
\begin{aligned}
m &= \frac{1}{d}\mathbb{E}_\mathcal{D}\left[\boldsymbol{\theta}_0^\top \boldsymbol{\Sigma_x}\hat{\boldsymbol{\theta}}\right], & q &= \frac{1}{d}\mathbb{E}_\mathcal{D}\left[\hat{\boldsymbol{\theta}}^\top \boldsymbol{\Sigma_x}\hat{\boldsymbol{\theta}}\right], & P &= \frac{1}{d}\mathbb{E}_\mathcal{D}\left[\hat{\boldsymbol{\theta}}^\top \boldsymbol{\Sigma_\delta}\hat{\boldsymbol{\theta}}\right], \\
A &= \frac{1}{d}\mathbb{E}_\mathcal{D}\left[\hat{\boldsymbol{\theta}}^\top \boldsymbol{\Sigma_v}\hat{\boldsymbol{\theta}}\right], & F &= \frac{1}{d}\mathbb{E}_\mathcal{D}\left[\boldsymbol{\theta}_0^\top \boldsymbol{\Sigma_v}\hat{\boldsymbol{\theta}}\right].
\end{aligned}
\tag{46}
$$

The overlap m describes the angle between student estimate $\hat{\boldsymbol{\theta}}$ and the ground truth teacher vector $\boldsymbol{\theta}_0$, the overlap q represents the data-weighted norm of the weights, P quantifies the norm of the weights scaled by the defence direction, A quantifies how much the student lies in the attack geometry and F quantifies the overlap between teacher and student in the attack geometry. Note that contrary to Eqs. (11) and (13), these expressions cannot be used to efficiently obtain sufficient statistics as they depend on average over dataset realisation of the trained weights $\hat{\boldsymbol{\theta}}$.

It is important to note that all the summary statistics involved in the statement of the theorem are finite-dimensional as the dimension increases, and therefore the result is a fully asymptotic characterisation, in the sense that it does not involve any high-dimensional object. With this theorem, we can avoid solving Eq. (5) (a high dimensional problem) and instead solve Eqs. (11) and (13) (eight dimensional problem): all quantities of interest can be expressed of scalar parameters/sufficient statistics that concentrate in the high-dimensional limit.

B   ERROR METRICS

As is common in machine learning, we want to see how the model trained performs on different metrics. This section is devoted to defining the metrics of our interest and expressing them as a function of the overlaps solutions of the fixed-point equations in Theorem 3.1. In general, we distinguish between the strength of the training attack and indicate it as ε_t and the strength of the actual attacker considered in generalisation, which we call ε_g. In each subsection, we provide the formula to compute the value from the overlap solution of Theorem 3.1 or directly from a data set generated with the distribution explained in Section 2.

Some of these metrics will be computed using the method of "local fields" where one can suppose the jointly Gaussian behaviour for $(\hat{\boldsymbol{\theta}}^\top \boldsymbol{x}/\sqrt{d}, \boldsymbol{\theta}_0^\top \boldsymbol{x}/\sqrt{d})$ with mean zero and covariance $\boldsymbol{\sigma} = \left(\begin{smallmatrix} \rho & m \\ m & q \end{smallmatrix} \right)$. We will refer to the following probability distribution

$$\mathrm{d}\mu\left(\nu,\lambda\right) = \frac{1}{2\pi\sqrt{\det \boldsymbol{\sigma}}} \exp\left(-\frac{1}{2} \begin{pmatrix} \nu \\ \lambda \end{pmatrix}^\top \boldsymbol{\sigma}^{-1} \begin{pmatrix} \nu \\ \lambda \end{pmatrix} \right) \tag{47}$$

This method is explained at length in Clarté et al. (2023a) and applies also in the case considered here. For other computations we refer to the places where the computation can be found in detail.

B.1   Generalisation

In machine learning, particularly in the context of adversarial training, we are concerned with how well our model, referred to as the student model, can make correct predictions on data that has not been altered or perturbed. We quantify this ability using a metric called the generalisation error. The generalisation error, denoted by E_{gen}, is the expected value over all possible data points of whether the model's prediction $\hat{y}$ is incorrect. When the model's predictions are based on the estimated parameter vector $\hat{\boldsymbol{\theta}}$, the generalisation error is mathematically expressed as

$$E_{\mathrm{gen}} = \mathbb{E}_{y,\boldsymbol{x}} \left[\mathbb{1}[y \neq \hat{y}(\hat{\boldsymbol{\theta}}; \boldsymbol{x})] \right] \tag{48}$$

This can also be represented in terms of overlaps using

$$E_{\mathrm{gen}} = \int \mathrm{d}y \, \mathrm{d}\mu\left(\nu,\lambda\right) P(y \mid \lambda) \mathbb{1}[y \neq f(\nu)] = \frac{1}{\pi} \arccos\left(\frac{m}{\sqrt{(\rho + \tau^2)q}} \right) \tag{49}$$

where this integral can be simplified further in the case of simple models, *e.g.* the case of a noiseless channel. The final form in the case of in the noiseless case has been presented in Aubin et al. (2020) Appendix II. For the noisy case the derivation can be found in Clarté et al. (2023a,b). The general form can be found in (Gerace et al., 2021, Appendix D).

However, in the case of adversarial learning, we're not only interested in the generalisation error under normal conditions but also under adversarial attacks. The adversarial generalisation error measures the model's

robustness against such attacks by evaluating the probability of misclassification when the input data is perturbed within a certain norm bound determined by ε_g and the covariance matrix $\boldsymbol{\Sigma_v}$.

Formally, the adversarial generalisation error is given by

$$E_{\text{adv}} = \mathbb{E}_{y,\boldsymbol{x}} \left[\max_{\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma_v}^{-2}} \leq \varepsilon_g} \mathbb{1}[y \neq \hat{y}(\hat{\boldsymbol{\theta}}(\alpha); \boldsymbol{x} + \boldsymbol{\delta})] \right] \tag{50}$$

where the inner maximisation, for a fixed choice of $\boldsymbol{x}$ and $\hat{\boldsymbol{\theta}}$ can be solved explicitly.

Again by using the idea of local fields one can compute the value of the adversarial error as a function of the overlaps. The definitions of the overlap parameters A, F is fixed after training and thus it doesn't get included in the average. We have that the form of the error is

$$\begin{aligned} E_{\text{adv}} &= \int_0^\infty \text{erfc}\left(\frac{\frac{m}{\sqrt{q}}\xi}{\sqrt{2\left(\rho + \tau^2 - m^2/q\right)}} \right) \frac{e^{-\frac{\xi^2}{2}}}{\sqrt{2\pi}} \, \mathrm{d}\xi \\ &+ \int_0^{\varepsilon_g \frac{\sqrt{A}}{\sqrt{q}}} \text{erfc}\left(\frac{-\frac{m}{\sqrt{q}}\xi}{\sqrt{2\left(\rho + \tau^2 - m^2/q\right)}} \right) \frac{e^{-\frac{\xi^2}{2}}}{\sqrt{2\pi}} \, \mathrm{d}\xi \\ &= E_{\text{gen}} + \int_0^{\varepsilon_g \frac{\sqrt{A}}{\sqrt{q}}} \text{erfc}\left(\frac{-\frac{m}{\sqrt{q}}\xi}{\sqrt{2\left(\rho + \tau^2 - m^2/q\right)}} \right) \frac{e^{-\frac{\xi^2}{2}}}{\sqrt{2\pi}} \, \mathrm{d}\xi \end{aligned} \tag{51}$$

One could modify the equation even more to relate it to the Owen's T function as we show in Appendix C.5.

B.2   Training

The training error reflects the model's performance on the dataset it was trained on. The goal during the training phase is to minimise this error and we can expect to reach an optimal zero training error for the estimator $\hat{\boldsymbol{\theta}}$ on noise less dataset. The training error is expressed as

$$E_{\text{train}} = \sum_{i=1}^n \mathbb{1}\left[\hat{y}(\hat{\boldsymbol{\theta}}; \boldsymbol{x}_i) \neq y_i \right] \tag{52}$$

where the sum is over the same dataset used to find $\hat{\boldsymbol{\theta}}$.

Following Gerace et al. (2021, Appendix D.2), one can find the following form of the training error as a function of overlaps as

$$E_{\text{train}} = \frac{1}{2}\mathbb{E}_\xi \left[\int \mathrm{d}y \, \mathcal{Z}_0(y, \sqrt{\eta}\xi, \rho - \rho\eta) \mathbb{1}\left[\text{sign}\left(\mathcal{P}_{Vg(\cdot; y, P)}(\sqrt{q}\xi) \right) \neq y \right] \right] \tag{53}$$

where $\eta = \frac{m}{\sqrt{\rho q}}$.

Arguably more interesting than the training error is the training loss since it is a part of the actual risk to minimise. This metric also brings more information than the training error since usually the losses used for classification are also sensitive to the norm of the solution and not only the direction of $\hat{\boldsymbol{\theta}}$. The definition as a function of the predicted ERM weights is

$$\ell_{\text{train}} = \sum_{i=1}^n g\left(y_i \frac{\hat{\boldsymbol{\theta}}^\top \boldsymbol{x}_i}{\sqrt{d}} \right) \tag{54}$$

and as a function of overlaps

$$\ell_{\text{train}} = \mathbb{E}_{y,\xi} \left[\mathcal{Z}_0(y, \sqrt{\eta}\xi, \rho - \rho\eta) \, g\left(y, \mathcal{P}_{Vg(\cdot; y, P)}(\sqrt{q}\xi), \varepsilon_t \sqrt{P} \right) \right] \tag{55}$$

B.3 Expression for Teacher Usefulness and Robustness

The equations for the usefulness and the robustness can be found using the same approach as before. We remember assumption **(A1)** and **(A4)** of Loureiro et al. (2022a) in Appendix B for the assumptions on the teacher. We explicitly have that

$$
\mathcal{U}_{\boldsymbol{\theta}_0} = \frac{1}{\sqrt{d}} \mathbb{E}_{\boldsymbol{x},y}[y\boldsymbol{\theta}_0^\top \boldsymbol{x}] = \sqrt{\frac{2}{\pi}} \frac{\rho}{\sqrt{\rho + \tau^2}}\,,
$$

$$
\mathcal{R}_{\boldsymbol{\theta}_0} = \frac{1}{\sqrt{d}} \mathbb{E}_{\boldsymbol{x},y}\left[\inf_{\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}_{\boldsymbol{v}}^{-2}} \leq \varepsilon_g} y\boldsymbol{\theta}_0^\top (\boldsymbol{x} + \boldsymbol{\delta}) \right] \tag{56}
$$

$$
= \frac{1}{\sqrt{d}} \mathbb{E}_{\boldsymbol{x},y}[y\boldsymbol{\theta}_0^\top \boldsymbol{x}] - \frac{\varepsilon_g}{\sqrt{d}} \mathbb{E}\left[\sqrt{\boldsymbol{\theta}^\top \boldsymbol{\Sigma}_{\boldsymbol{v}} \boldsymbol{\theta}} \right] = \mathcal{U}_{\boldsymbol{\theta}_0} - \varepsilon_g \sqrt{a}\,.
$$

and in the end we have integrated using the local fields method in the case of averaged teacher.

C LARGE SAMPLE COMPLEXITY ASYMPTOTICS

In this Appendix, we provide expansion of the fixed point equations in Theorem 3.1. In practical terms, the high sample complexity regime is challenging to access due to the computational resources it demands to simulate the ERM. By providing a theoretical expansion, we offer a lens through which the system's behaviour under high sample complexity can be understood and predicted at less computational cost.

C.1 Self Consistent equations in the large α limit

We consider the following behaviour as a function of α for the order parameters

$$
\begin{aligned}
q &\underset{\alpha\to\infty}{=} q_0\,, & \hat{q} &\underset{\alpha\to\infty}{=} \hat{q}_0\alpha\,, \\
m &\underset{\alpha\to\infty}{=} m_0\,, & \hat{m} &\underset{\alpha\to\infty}{=} \hat{m}_0\alpha\,, \\
V &\underset{\alpha\to\infty}{=} \frac{V_0}{\alpha}\,, & \hat{V} &\underset{\alpha\to\infty}{=} \hat{V}_0\alpha\,, \\
P &\underset{\alpha\to\infty}{=} P_0\,, & \hat{P} &\underset{\alpha\to\infty}{=} \hat{P}_0\alpha\,, \\
A &\underset{\alpha\to\infty}{=} A_0\,, & \hat{A} &\underset{\alpha\to\infty}{=} 0\,, \\
F &\underset{\alpha\to\infty}{=} F_0\,, & \hat{F} &\underset{\alpha\to\infty}{=} 0\,,
\end{aligned} \tag{57}
$$

where the sub scripted quantities are independent of α. We remark that this scaling is the same found in Vilucchio et al. (2024b).

These scaling assumptions break down whenever $\varepsilon_t = \tau = 0$, in this case, the first row of overlaps except V scale linearly in α. This case is already studied in Aubin et al. (2020). This specific scaling ansatz is justified not only because of consistency of the saddle point equations but also because it is the one found empirically in the study of the phenomena.

The study of the proximity operator for the logistic loss function $f(v) : v \mapsto \log(1 + \exp(-v))$ has been studied in (Briceño-Arias et al., 2019, Section 3.3). Explicitly from (Briceño-Arias et al., 2019, Propositon 2) one has that

$$
\mathcal{P}_{Vf(\cdot)}(\omega) = \omega + \mathbf{W}_{\exp(-\omega)}(V\exp(-\omega)) \tag{58}
$$

where specifically $\mathbf{W}$ is the generalised Lambert function that satisfies

$$
(\forall \bar{v} \in \mathbb{R})(\forall v \in \mathbb{R})(\forall r \in]0, +\infty[) \quad \bar{v}(\exp(\bar{v}) + r) = v \quad \Leftrightarrow \quad \bar{v} = \mathbf{W}_r(v). \tag{59}
$$

Translating this result for our loss function $g(y, \cdot; \varepsilon_t, P)$ gives us the following proximal operator

$$
\mathcal{P}_{Vg(\cdot;y,P)}(\sqrt{q}\xi) = \sqrt{q}\xi + y\mathbf{W}_{\exp\left(-y\sqrt{q}\xi + \varepsilon_t\sqrt{P}\right)}\left(V\exp\left(-\sqrt{q}\xi + \varepsilon_t\sqrt{P}\right) \right) \tag{60}
$$

To simplify the proximal operator we start by its formulation then expand

$$\mathcal{P}_{Vg(\cdot;y,P)}(\sqrt{q}\xi) \underset{\alpha \to \infty}{=} \sqrt{q_0}\xi + \frac{c(\xi)}{\alpha}, \qquad c(\xi) = yV_0\frac{\exp(-y\sqrt{q_0}\xi + \varepsilon_t\sqrt{P_0})}{1 + \exp(-y\sqrt{q_0}\xi + \varepsilon_t\sqrt{P_0})}. \tag{61}$$

With this, we can simplify the channel and the prior equations as follows where we remember that $\eta_0 = m_0^2/q_0$ and $\hat{\eta} = \hat{m}^2/\hat{q}$

$$\begin{cases} \hat{m}_0 &= \frac{1}{V_0}\mathbb{E}_{y,\xi}\left[\partial_\omega \mathcal{Z}_0(y, \sqrt{\eta_0}\xi, \rho - \eta_0)c(\xi)\right] \\ \hat{q}_0 &= \frac{1}{V_0^2}\mathbb{E}_{y,\xi}\left[\mathcal{Z}_0(y, \sqrt{\eta_0}\xi, \rho - \eta_0)c(\xi)^2\right] \\ \hat{V}_0 &= \mathbb{E}_{y,\xi}\left[\mathcal{Z}_0(y, \sqrt{\eta_0}\xi, \rho - \eta_0)\partial^2 g(y, \sqrt{q_0}\xi, P_0, \varepsilon_t))\right] \\ \hat{P}_0 &= \frac{\varepsilon_t}{\sqrt{N_0 V_0}}\mathbb{E}_{y,\xi}\left[\mathcal{Z}_0(y, \sqrt{\eta_0}\xi, \rho - \eta_0)yc(\xi)\right] \end{cases} \tag{62}$$

$$\begin{cases} m_0 &= \frac{1}{d}\operatorname{tr}\left[\hat{m}_0\mathbf{\Sigma}_{\boldsymbol{x}}^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\mathbf{\Sigma}_{\boldsymbol{x}}\mathbf{\Lambda}_0^{-1}\right] \\ q_0 &= \frac{1}{d}\operatorname{tr}\left[\hat{m}_0^2\mathbf{\Sigma}_{\boldsymbol{x}}^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\mathbf{\Sigma}_{\boldsymbol{x}}\mathbf{\Sigma}_{\boldsymbol{x}}\mathbf{\Lambda}_0^{-2}\right] \\ V_0 &= \frac{1}{d}\operatorname{tr}\left[\mathbf{\Sigma}_{\boldsymbol{x}}\mathbf{\Lambda}_0^{-1}\right] \\ P_0 &= \frac{1}{d}\operatorname{tr}\left[\hat{m}_0^2\mathbf{\Sigma}_{\boldsymbol{x}}^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\mathbf{\Sigma}_{\boldsymbol{x}}\mathbf{\Sigma}_{\boldsymbol{\delta}}\mathbf{\Lambda}_0^{-2}\right] \end{cases} \tag{63}$$

with $\mathbf{\Lambda}_0 = \lambda_1\mathbb{1} + \hat{V}_0\mathbf{\Sigma}_{\boldsymbol{x}} + \hat{P}_0\mathbf{\Sigma}_{\boldsymbol{\delta}}$ where $\lambda = \lambda_1\alpha$. Additional explanation on this behaviour can be found in Vilucchio et al. (2024b, Appendix D).

The additional values are

$$A_0 = \frac{1}{d}\operatorname{tr}\left[\hat{m}_0^2\mathbf{\Sigma}_{\boldsymbol{x}}^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\mathbf{\Sigma}_{\boldsymbol{x}}\mathbf{\Sigma}_{\boldsymbol{v}}\mathbf{\Lambda}_0^{-2}\right], \qquad F_0 = \frac{1}{d}\operatorname{tr}\left[\hat{m}_0^2\mathbf{\Sigma}_{\boldsymbol{x}}^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\mathbf{\Sigma}_{\boldsymbol{v}}\mathbf{\Lambda}_0^{-1}\right]. \tag{64}$$

This formulation can be evaluated using the set of equations in Theorem 3.1 upon convergence.

C.2 Large α limit of E_{adv}

Using above simplifications, we can simplify the following quantities in the large sample complexity limit.

We can write the following

$$\lim_{\alpha \to \infty} \frac{m}{\sqrt{(\rho + \tau^2)q}} = \frac{\operatorname{tr}\left[\mathbf{\Sigma}_{\boldsymbol{x}}^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\mathbf{\Sigma}_{\boldsymbol{x}}\mathbf{\Lambda}_0^{-1}\right]}{\sqrt{(\boldsymbol{\theta}_0^\top\mathbf{\Sigma}_{\boldsymbol{x}}\boldsymbol{\theta}_0 + \tau^2)\operatorname{tr}\left[\mathbf{\Sigma}_{\boldsymbol{x}}^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\mathbf{\Sigma}_{\boldsymbol{x}}\mathbf{\Sigma}_{\boldsymbol{x}}\mathbf{\Lambda}_0^{-2}\right]}} \tag{65}$$

and the overlap ratio determining how strong an attack is, is given by

$$\lim_{\alpha \to \infty} \frac{\sqrt{A}}{\sqrt{q}} = \frac{\sqrt{\operatorname{tr}\left[\mathbf{\Sigma}_{\boldsymbol{x}}^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\mathbf{\Sigma}_{\boldsymbol{x}}\mathbf{\Sigma}_{\boldsymbol{v}}\mathbf{\Lambda}_0^{-2}\right]}}{\sqrt{\operatorname{tr}\left[\mathbf{\Sigma}_{\boldsymbol{x}}^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\mathbf{\Sigma}_{\boldsymbol{x}}\mathbf{\Sigma}_{\boldsymbol{x}}\mathbf{\Lambda}_0^{-2}\right]}} \tag{66}$$

Notably, the previous equations no longer depend on $\hat{q}_0$ and $\hat{m}_0$, with this the adversarial generalisation error is characterised by the following equation, which is a function of the limiting values we just defined. Nonetheless the previous equation still depends on the values of $\hat{V}_0, \hat{P}_0$ through $\mathbf{\Lambda}_0$. The values are constants evaluated at the fixed point of Eqs. (62) and (63).

We also want to show that the plateau is reached with a derivative that is going to zero. To do so we first look at the derivative of the angle variables $m/\sqrt{\rho q}$ and $\sqrt{A}/\sqrt{q}$.

Specifically we have that

$$
\begin{aligned}
\frac{m}{\sqrt{\rho q}} &= \frac{\operatorname{tr}\left[\hat{m}_0 \boldsymbol{\Sigma}_{\boldsymbol{x}}^\top \boldsymbol{\theta}_0 \boldsymbol{\theta}_0^\top \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{\Lambda}_0^{-1}\right]}{\sqrt{\boldsymbol{\theta}_0^\top \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{\theta}_0 \operatorname{tr}\left[\left(\hat{m}_0^2 \boldsymbol{\Sigma}_{\boldsymbol{x}}^\top \boldsymbol{\theta}_0 \boldsymbol{\theta}_0^\top \boldsymbol{\Sigma}_{\boldsymbol{x}} + \frac{\hat{q}_0}{\alpha} \boldsymbol{\Sigma}_{\boldsymbol{x}}\right) \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{\Lambda}_0^{-2}\right]}} \\[2ex]
\frac{\sqrt{A}}{\sqrt{q}} &= \frac{\sqrt{\operatorname{tr}\left[\left(\hat{m}_0^2 \boldsymbol{\Sigma}_{\boldsymbol{x}}^\top \boldsymbol{\theta}_0 \boldsymbol{\theta}_0^\top \boldsymbol{\Sigma}_{\boldsymbol{x}} + \frac{\hat{q}_0}{\alpha} \boldsymbol{\Sigma}_{\boldsymbol{x}}\right) \boldsymbol{\Sigma}_{\boldsymbol{v}} \boldsymbol{\Lambda}_0^{-2}\right]}}{\sqrt{\operatorname{tr}\left[\left(\hat{m}_0^2 \boldsymbol{\Sigma}_{\boldsymbol{x}}^\top \boldsymbol{\theta}_0 \boldsymbol{\theta}_0^\top \boldsymbol{\Sigma}_{\boldsymbol{x}} + \frac{\hat{q}_0}{\alpha} \boldsymbol{\Sigma}_{\boldsymbol{x}}\right) \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{\Lambda}_0^{-2}\right]}}
\end{aligned}
\tag{67}
$$

To compute the derivative, we use the additivity of the trace. Note that we compute the derivative with respect to a scalar that is not dependent on the trace itself. Thus, every derivative in both terms will lead to a factor of the form $\frac{c}{\alpha^2} \operatorname{tr}(M)$, which goes to zero as $\alpha \to \infty$.

Thus the adversarial generalisation error approaches a constant in the large sample complexity limit

$$
\partial_\alpha E_{\mathrm{adv}} \underset{\alpha \to \infty}{=} 0, \quad E_{\mathrm{adv}} \underset{\alpha \to \infty}{=} cst.
\tag{68}
$$

C.3 Specific Case of Single Block Model

In this subsection we consider the case where it only exists one kind of feature and we show that this will lead to results that are independent of adversarial training in the $\alpha \to \infty$ limit.

Specifically we consider the fact of having a Block Feature Data Model as explained in Section 2.2 with a single block. We note that the data model of Javanmard et al. (2020) fall under this category.

In this specific case the self consistent equations in Eq. (63) simplify as

$$
\begin{aligned}
m_0 &= \frac{\hat{m}_0 \psi^2 t}{\lambda_1 + \hat{V}_0 \psi}, & q_0 &= \frac{\hat{m}_0^2 \psi^3 t}{(\lambda_1 + \hat{V}_0 \psi)^2}, & V_0 &= \frac{1}{\lambda_1 + \hat{V}_0 \psi}, \\[2ex]
P_0 &= \frac{\hat{m}_0^2 \psi^2 \Delta t}{(\lambda_1 + \hat{V}_0 \psi)^2}, & A_0 &= \frac{\hat{m}_0^2 \psi^2 t \Upsilon}{(\lambda_1 + \hat{V}_0 \psi)^2}, & F_0 &= \frac{\hat{m}_0 \psi t \Upsilon}{\lambda_1 + \hat{V}_0 \psi}.
\end{aligned}
\tag{69}
$$

From the analysis done in the previous section we have that

$$
\lim_{\alpha \to \infty} \frac{m}{\sqrt{(\rho + \tau^2) q}} = \frac{\psi^2 t}{\sqrt{(\psi t + \tau^2) \psi^3 t}}, \quad \lim_{\alpha \to \infty} \frac{\sqrt{A}}{\sqrt{q}} = \frac{\psi^2 t \Upsilon}{\sqrt{\psi^2 t \psi^3 t}}.
\tag{70}
$$

We notice that these equations are independent from the value of ε_t.

By writing the system of saddle point equations to next leading order, one could solve for the next leading order of the overlap parameters (m_1, q_1, V_1, P_1). In general these do depend on the value of ε_t but by the integer scaling of all the analytic functions we have that the overlap parameters at that order differ of an order $1/\alpha$. And the same order of difference is there between the terms of E_{gen} and E_{bnd}.

We have that the second leading order can be bounded by a constant times $1/\alpha$ thus the claim follows.

C.4 Change of Defence Direction

We can rewrite the adversarial error as

$$
E_{\mathrm{adv}} = \frac{1}{\pi} \arccos\left(\frac{\vartheta}{\sqrt{(1 + \frac{\tau^2}{\rho})}}\right) + \int_0^{\varepsilon_g u} \operatorname{erfc}\left(\frac{-\vartheta \nu}{\sqrt{2\left((1 + \frac{\tau^2}{\rho}) - \vartheta^2\right)}}\right) \frac{e^{-\frac{1}{2}\nu^2}}{\sqrt{2\pi}} \, d\nu
\tag{71}
$$

where $\vartheta = m/\sqrt{\rho q}$ and $u = \sqrt{A}/\sqrt{q}$ as there is no dependance on the result of the saddle point equations.

And the formal power series is given by

$$E_{\text{gen}} = \frac{1}{\pi} \arccos\left(\frac{\vartheta(0)}{\sqrt{(1 + \frac{\tau^2}{\rho})}}\right) - \frac{\vartheta'(0)}{\pi\sqrt{1 - \vartheta(0)^2}} \varrho + \mathcal{O}(\varrho^2) \tag{72}$$

$$E_{\text{bnd}} = \int_0^{\varepsilon_g u(0)} \text{erfc}\left(\frac{-\vartheta(0)\nu}{\sqrt{2\left((1 + \frac{\tau^2}{\rho}) - \vartheta(0)^2\right)}}\right) \frac{e^{-\frac{1}{2}\nu^2}}{\sqrt{2\pi}} \, d\nu$$

$$+ \varepsilon_g \frac{e^{-\frac{1}{2}u(0)^2\varepsilon_g^2}}{\sqrt{2\pi}} u'(0) \, \text{erfc}\left(-\frac{\vartheta(0)u(0)\varepsilon_g}{\sqrt{2 - 2\vartheta(0)^2}}\right) \varrho - \frac{e^{-\frac{u(0)^2\varepsilon_g^2}{2(1-\vartheta(0)^2)}} - 1}{\pi\sqrt{1 - \vartheta(0)^2}} \theta'(0)\varrho + \mathcal{O}(\varrho^2) \tag{73}$$

We now need to consider the sign of the derivatives of $u'(0)$ and $\vartheta'(0)$. The main point is that one can expand to get the derivatives as a function of the solution.

We consider the case of two features as in Proposition 4.2. We also suppose that we have an attack that is $\boldsymbol{\Sigma_v} = \mathbb{1}$. We also suppose that $\psi_1 > \psi_2$ without loss of generality.

$$\frac{m^2}{\rho q} = \frac{\left(\phi_1\psi_1^2 b + \phi_2\psi_2^2 a\right)^2}{(\phi_1\psi_1 + \phi_2\psi_2)(\phi_1\psi_1^3 b^2 + \phi_2\psi_2^3 a^2)} \qquad \frac{A}{\sqrt{q}} = \frac{\phi_1\psi_1^2 b^2 + \phi_2\psi_2^2 a^2}{\sqrt{\phi_1\psi_1^3 b^2 + \phi_2\psi_2^3 a^2}} \tag{74}$$

where $a = \lambda + \hat{V}\psi_1 + \hat{P}(1 + \delta_1\varrho) + \hat{N}$ and $b = \lambda + \hat{V}\psi_2 + \hat{P}(1 + \delta_2\varrho) + \hat{N}$.

By the use of the implicit function theorem (Krantz and Parks, 2002) one has that the any overlap parameter u solution in the limit $\varrho \to 0$ can be written as $u = u_0 + u_1\varrho + \mathcal{O}(\varrho^2)$, where u_0 is the same overlap that is solution for $\varrho = 0$. Thus we can expand and thus at first order the following

$$\frac{m}{\sqrt{\rho q}} = \frac{m_0}{\sqrt{\rho q_0}} + \frac{\mathcal{N}_\vartheta}{\mathcal{D}_\vartheta}\varrho + \mathcal{O}(\varrho^2), \qquad \frac{A}{\sqrt{q}} = \frac{A_0}{\sqrt{q_0}} + \frac{\mathcal{N}_u}{\mathcal{D}_u}\varrho + \mathcal{O}(\varrho^2), \tag{75}$$

where we have that both $\mathcal{D}_\vartheta > 0$ and $\mathcal{D}_u > 0$ independently of the values $\psi_1, \psi_2 > 0$ and δ_1, δ_2. Additionaly we have

$$\mathcal{N}_\vartheta = \hat{P}_0 t_1 t_2 \psi_1^2 \psi_2^2 (\Delta_2\psi_1 - \Delta_1\psi_2)(-\Delta_1\hat{P}_0\hat{V}_1\psi_2 + \Delta_2\hat{P}_0\hat{V}_1\psi_1 + \Delta_1\hat{P}_1\hat{V}_0\psi_2 + \Delta_2(-\hat{P}_1)\hat{V}_0\psi_1)$$

$$- \hat{P}_0 t_1 t_2 \psi_1^2 \psi_2^2 (\Delta_2\psi_1 - \Delta_1\psi_2)\left(\Delta_1\hat{P}_0^2 + \hat{P}_0\hat{V}_0\psi_1\right)\delta_2 \tag{76}$$

$$+ \hat{P}_0 t_1 t_2 \psi_1^2 \psi_2^2 (\Delta_2\psi_1 - \Delta_1\psi_2)\left(\Delta_2\hat{P}_0^2 + \hat{P}_0\hat{V}_0\psi_2\right)\delta_1$$

and

$$\mathcal{N}_u = t_1 t_2 \psi_1^2 \psi_2^2 (\psi_2 - \psi_1)(\Delta_1\hat{P}_0 + \hat{V}_0\psi_1)(\Delta_2\hat{P}_0 + \hat{V}_0\psi_2)(\Delta_1\hat{P}_0\hat{V}_1\psi_2 - \Delta_2\hat{P}_0\hat{V}_1\psi_1 - \Delta_1\hat{P}_1\hat{V}_0\psi_2 + \Delta_2\hat{P}_1\hat{V}_0\psi_1)$$

$$- t_1 t_2 \psi_1^2 \psi_2^2 (\psi_1 - \psi_2)(\Delta_1\hat{P}_0 + \hat{V}_0\psi_1)\left(\Delta_1\hat{P}_0^2 + \hat{P}_0\hat{V}_0\psi_1\right)(\Delta_2\hat{P}_0 + \hat{V}_0\psi_2)\delta_2$$

$$+ t_1 t_2 \psi_1^2 \psi_2^2 (\psi_1 - \psi_2)(\Delta_1\hat{P}_0 + \hat{V}_0\psi_1)(\Delta_2\hat{P}_0 + \hat{V}_0\psi_2)\left(\Delta_2\hat{P}_0^2 + \hat{P}_0\hat{V}_0\psi_2\right)\delta_1$$

$$\tag{77}$$

We see that for the term $\mathcal{N}_u$ we have that the sign of the coefficients in front of δ_1 and δ_2 have the same sign as $\pm(\psi_1 - \psi_2)$ respectively.

For the therm $\mathcal{N}_\vartheta$ the analysis is a little bit more complicated.

In the case where $\Delta_1 = \Delta_2 = 1$ we see from this that $\psi_1 \gtrless \psi_2$ and we chose $\delta_1 \lessgtr \delta_2$ we have that both of the numerators are positive.

$$\Delta_2\psi_1 - \Delta_1\psi_2 \geq 0 \tag{78}$$

if this condition is satisfied then the claim holds, otherwise not. Take into consideration that as long as $\Delta_2 \geq \Delta_1$ this is also the case.

C.5 Adversarial Error and Owen's T function

We demonstrate that the adversarial error from Eq. (51) can be written in terms of Owen's T function, we begin by explicitly stating the adversarial error. We believe that given the usual definition of the Owen's T function this is a interesting geometrical approach.

We start from the definition of the adversarial error

$$
\begin{aligned}
E_{\text{adv}}(\varepsilon_g) = &\int_0^\infty \text{erfc}\left(\frac{\frac{m}{\sqrt{q}}\xi}{\sqrt{2\left(\rho + \tau^2 - m^2/q\right)}}\right) \frac{e^{-\frac{\xi^2}{2}}}{\sqrt{2\pi}}\, d\xi \\
&+ \int_0^{\varepsilon_g \frac{\sqrt{A}}{\sqrt{q}}} \text{erfc}\left(\frac{-\frac{m}{\sqrt{q}}\xi}{\sqrt{2\left(\rho + \tau^2 - m^2/q\right)}}\right) \frac{e^{-\frac{\xi^2}{2}}}{\sqrt{2\pi}}\, d\xi
\end{aligned}
\tag{79}
$$

where we did a change of variable with an opposite sign and we changed the order of integration.

Now we would like to use some identities to deal with the non-adversarial part of the generalisation error. We use the formula

$$
\int_0^\infty \text{erfc}(ax)e^{b^2 x^2}\, dx = \frac{1}{2\sqrt{\pi}b}\ln\left[\frac{a+b}{a-b}\right], \quad b \text{ may be complex}, \quad |\arg a| < \frac{\pi}{4}
\tag{80}
$$

Finally, we obtain an expression for the adversarial generalisation error as a function of the standard generalisation error and an integral for the boundary error.

$$
E_{\text{adv}} = \frac{1}{\pi}\arccos\left(\frac{m}{\sqrt{(\rho+\tau^2)q}}\right) + \int_0^{\varepsilon_g \frac{\sqrt{A}}{\sqrt{q}}} \text{erfc}\left(\frac{-\frac{m}{\sqrt{q}}\xi}{\sqrt{2\left(\rho + \tau^2 - m^2/q\right)}}\right) \frac{e^{-\frac{\xi^2}{2}}}{\sqrt{2\pi}}\, d\xi
\tag{81}
$$

We would like to simplify this part more and obtain a relation to the Owen T function definition Owen (1956). From Ng and Geller (1969); Korotkov and Korotkov (2020) we use the following identity

$$
\int_0^b e^{-x^2}\text{erfc}(ax)\, dx = \frac{1}{2\sqrt{\pi}a}\left[-4\pi\sqrt{a^2}T\left(\sqrt{2}\sqrt{a^2}b, \frac{1}{\sqrt{a^2}}\right) + \pi a\,\text{erf}(b)\,\text{erfc}(ab) + 2a\cot^{-1}(a)\right]
\tag{82}
$$

where we are using Owen's T function which is defined as

$$
T(h, a) = \frac{1}{2\pi}\int_0^a \frac{e^{-\frac{1}{2}h^2\left(1+x^2\right)}}{1+x^2}\, dx \quad (-\infty < h, a < +\infty).
\tag{83}
$$

By performing a change of variables $\xi' = \xi/\sqrt{2}$ we can simplify further the second term of Eq. (81) as

$$
\begin{aligned}
E_{\text{bnd}} = &2T\left(\frac{m}{\sqrt{(\rho+\tau^2)q - m^2}}\varepsilon_g\frac{\sqrt{A}}{\sqrt{q}}, \frac{\sqrt{(\rho+\tau^2)q - m^2}}{m}\right) \\
&+ \frac{1}{2}\text{erf}\left(\varepsilon_g\frac{\sqrt{A}}{\sqrt{2q}}\right)\text{erfc}\left(\frac{-m}{\sqrt{2((\rho+\tau^2)q - m^2)}}\varepsilon_g\frac{\sqrt{A}}{\sqrt{q}}\right) \\
&+ \frac{1}{\pi}\cot^{-1}\left(\frac{-m}{\sqrt{(\rho+\tau^2)q - m^2}}\right)
\end{aligned}
\tag{84}
$$

Given that the adversarial test error in the case where it is not attacked is equal to the standard test error, it should be possible to simplify the expression for the owen's T function. Indeed, the following property Owen (1980) holds

$$
T(0, a) = \frac{1}{2\pi}\arctan(a) = \frac{1}{2\pi}\arccos\left(\frac{a}{\sqrt{1+a^2}}\right)
\tag{85}
$$

and this leads us to the unperturbed version of the generalisation error.
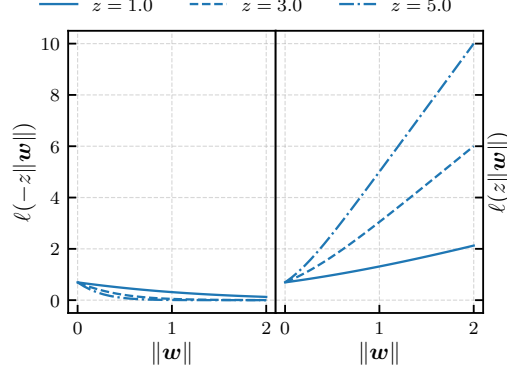
Figure 4: Behaviour of the loss function for positive and negative margins as a function of the norm of the student estimate. A negative margin favours a zero norm solution.

C.6 Simplified Expressions for Generalisation and Boundary Errors

We can rewrite the key quantity describing generalisation error as follows

$$\frac{m_0}{\sqrt{(\rho + \tau^2)q_0}} = \frac{m_0}{\sqrt{\rho q_0}} \frac{1}{\sqrt{1 + \frac{\tau^2}{\rho}}} \ . \tag{86}$$

Whenever $\vartheta = \frac{m_0}{\sqrt{\rho q_0}} \to 1$ when $\alpha \to \infty$, we can leverage, that the angle between teacher and student is zero, and the actual vectors differ only up to a constant, that is $\boldsymbol{\theta} = c\boldsymbol{\theta}_0$.

$$E_{\text{gen}} \underset{\alpha \to \infty}{=} \frac{1}{\pi} \arccos\left(\vartheta \sqrt{\frac{\pi}{2}} \frac{\mathcal{U}_{\boldsymbol{\theta}_0}}{\sqrt{\rho}}\right) = \frac{1}{\pi} \arccos\left(\frac{\sqrt{\rho}}{\sqrt{\rho + \tau^2}}\right) \tag{87}$$

Furthermore, the quantity $\frac{\sqrt{A}}{\sqrt{q}}$ converges to $\frac{a}{\sqrt{\rho}}$ by definition. Writing the boundary error as a function of $\mathcal{U}_{\boldsymbol{\theta}_0} - \mathcal{R}_{\boldsymbol{\theta}_0}$, leads thus

$$E_{\text{bnd}} = \int_0^{\varepsilon_g \frac{1}{\sqrt{q}}(\mathcal{U}_{\boldsymbol{\theta}_0} - \mathcal{R}_{\boldsymbol{\theta}_0})} f\left(\xi; \vartheta, \frac{\mathcal{U}_{\boldsymbol{\theta}_0}}{\sqrt{\rho}}\right) d\xi \underset{\alpha \to \infty}{=} \int_0^{\varepsilon_g \frac{a}{\sqrt{\rho n}}} f\left(\xi; \frac{1}{\sqrt{1 + \frac{\tau^2}{\rho}}}\right) d\xi \ . \tag{88}$$

We can simplify this expression further under the assumption of a uniform attack in a BFM with a singular block. In this case, the boundary term can be simplified and we obtain

$$E_{\text{bnd}} \underset{\alpha \to \infty}{=} \int_0^{\varepsilon_g \frac{1}{\sqrt{\psi_\ell}}} f\left(\xi; \frac{\sqrt{\rho}}{\sqrt{\rho + \tau^2}}\right) d\xi \ . \tag{89}$$

In this simple setting, the variance of the data influences inverse proportionally the integration bound.

D ADVERSARIAL PROBLEM AS AN APPROXIMATE DATA-DEPENDANT REGULARISATION

In other settings, adversarial training has been studied as a form of data dependent regularisation Roth et al. (2020); Ribeiro et al. (2024). In this Appendix, we provide an approximately equivalent loss function that can also be understood as a data dependent regularisation. Here by data dependant regularisation we mean an explicit term in the minimisation function that is separate from the loss function term, even if it can still depend on the dataset $\mathcal{D}$.

Recall the minimisation problem in Eq. (5) is

$$\sum_{\mu=1}^{n} g\left(y^{\mu}\frac{\boldsymbol{\theta}^{\top}\boldsymbol{x}^{\mu}}{\sqrt{d}} - \varepsilon_t\frac{\sqrt{\boldsymbol{\theta}^{\top}\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\boldsymbol{\theta}}}{\sqrt{d}}\right) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2 \tag{90}$$

where we specified the ℓ_2 regularisation. Also in this Appendix we will use g to specifically indicate the logistic loss, the same analysis could be performed for other losses similarly.

We want to split the sum into a first part, where the shifted margin is positive and a second part, where the shifted margin is negative

$$\sum_{\mu=1}^{n_1} g\left(y^{\mu}\frac{\boldsymbol{\theta}^{\top}\boldsymbol{x}^{\mu}}{\sqrt{d}} - \varepsilon_t\frac{\sqrt{\boldsymbol{\theta}^{\top}\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\boldsymbol{\theta}}}{\sqrt{d}}\right) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2 + \sum_{\mu=n_1+1}^{n} g\left(y^{\mu}\frac{\boldsymbol{\theta}^{\top}\boldsymbol{x}^{\mu}}{\sqrt{d}} - \varepsilon_t\frac{\sqrt{\boldsymbol{\theta}^{\top}\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\boldsymbol{\theta}}}{\sqrt{d}}\right). \tag{91}$$

Here, n_1 is the number of points that have shifted positive margin. We call a margin positive, when $y^{\mu}\frac{\boldsymbol{\theta}^{\top}\boldsymbol{x}^{\mu}}{\sqrt{d}} - \varepsilon_t\frac{\sqrt{\boldsymbol{\theta}^{\top}\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\boldsymbol{\theta}}}{\sqrt{d}} > 0$. We expand the term with the negative margin around zero for the logistic loss to obtain

$$\sum_{\mu=1}^{n_1} g(\dots) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2 - \frac{1}{2}\sum_{\mu=n_1+1}^{n}\left[y^{\mu}\frac{\boldsymbol{\theta}^{\top}\boldsymbol{x}^{\mu}}{\sqrt{d}} - \varepsilon_t\frac{\sqrt{\boldsymbol{\theta}^{\top}\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\boldsymbol{\theta}}}{\sqrt{d}}\right] + \frac{1}{8}\sum_{\mu=n_1+1}^{n}\left[y^{\mu}\frac{\boldsymbol{\theta}^{\top}\boldsymbol{x}^{\mu}}{\sqrt{d}} - \varepsilon_t\frac{\sqrt{\boldsymbol{\theta}^{\top}\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\boldsymbol{\theta}}}{\sqrt{d}}\right]^2 \tag{92}$$

The expansion around zero is justified, as a negative margin favours a zero norm solution, which we show in Fig. 4.

A first approximation is to not keep the mixed terms and resum all of them to obtain

$$\sum_{\mu=1}^{n_1} g(\dots) + \sum_{\mu=n_1+1}^{n} g\left(y^{\mu}\frac{\boldsymbol{\theta}^{\top}\boldsymbol{x}^{\mu}}{\sqrt{d}}\right) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2 + (n - n_1)g\left(\frac{\varepsilon_t}{\sqrt{d}}\sqrt{\boldsymbol{\theta}^{\top}\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\boldsymbol{\theta}}\right) \tag{93}$$

Note that one can also consider the case where ε_t is small and thus keep just the first terms in ε_t to get something like the following

$$\begin{aligned}
\sum_{\mu=1}^{n_1} g(\dots) &+ \sum_{\mu=n_1+1}^{n} g\left(y^{\mu}\frac{\boldsymbol{\theta}^{\top}\boldsymbol{x}^{\mu}}{\sqrt{d}}\right) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2 + (n - n_1)\frac{\varepsilon_t}{2}\frac{\sqrt{\boldsymbol{\theta}^{\top}\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\boldsymbol{\theta}}}{\sqrt{d}} \\
&- \frac{\varepsilon_t}{4}\frac{\sqrt{\boldsymbol{\theta}^{\top}\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\boldsymbol{\theta}}}{\sqrt{d}}\sum_{\mu=n_1+1}^{n} y^{\mu}\frac{\boldsymbol{\theta}^{\top}\boldsymbol{x}^{\mu}}{\sqrt{d}} \\
&+ \frac{\varepsilon_t^2}{8}\frac{\boldsymbol{\theta}^{\top}\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\boldsymbol{\theta}}{d}(n - n_1).
\end{aligned} \tag{94}$$

In both cases in the limit $\varepsilon_t \to 0$ there are two types of regularisation, the first is an $\sqrt{\ell_2}$ regularisation and the other is a negative shifted margin dependent ℓ_2 regularisation.

In our setting we suppose that the coefficient for the second term is positive as we are considering Gaussian data and in high dimension most of the data lies very close to the boundary.

The loss can then be written as

$$\sum_{i=1}^{n} g\left(y_i\frac{\boldsymbol{\theta}^{\top}\boldsymbol{x}_i}{\sqrt{d}}\right) + \tilde{\lambda}_1\sqrt{\boldsymbol{\theta}^{\top}\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\boldsymbol{\theta}} + \tilde{\lambda}_2\boldsymbol{\theta}^{\top}\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\boldsymbol{\theta}. \tag{95}$$

where the value for $\tilde{\lambda}_1$ and $\tilde{\lambda}_2$ can be read from the previous expansion.

E DEFENDABLE ATTACKS AND INEVITABLE TRADE-OFFS

In this Appendix we want to study the relation between the attack geometry $\boldsymbol{\Sigma_v}$ and the defence geometry $\boldsymbol{\Sigma_\delta}$. We assume to know the attack geometry. If we can successfully defend against a geometry, without incurring a trade-off, we call the geometry *defendable*, otherwise *not defendable*. A trade-off does not occur, if the adversarial error is dominated by the generalisation error and if the generalisation error behaves roughly as a constant as a function of the defence strength ε_t.

To make a fair comparison between the geometries we are going to consider a normalised version of the matrices $\boldsymbol{\Sigma_v}$ and $\boldsymbol{\Sigma_\delta}$ and consider the total cost of the attack/defence to be tuned by $\varepsilon_g/\varepsilon_t$. We are asking how to spend a fixed budget among the defence directions, to protect most effectively against a given attack. Since both matrices are symmetric we can write them as $\boldsymbol{\Sigma} = \sum_{i=1}^d \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^\top$ where $\sum_{i=1}^d \lambda_i = 1$ and the eigenvectors are normalised $\|\boldsymbol{v}_i\|_2 = 1$.

To keep things simple, we can suppose that the attack matrix is $\boldsymbol{\Sigma_v} = \boldsymbol{v}\boldsymbol{v}^\top$. We will refer to $\boldsymbol{v}$ as the unit vector that is composed of a sum between $\boldsymbol{\theta}_0$ and another vector $\boldsymbol{u}$ perpendicular to $\boldsymbol{\theta}_0$.

We can thus decompose the attack matrix as

$$\boldsymbol{\Sigma_v} = (\boldsymbol{\theta}_0 + \boldsymbol{u})(\boldsymbol{\theta}_0 + \boldsymbol{u})^\top = \boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top + \boldsymbol{\theta}_0\boldsymbol{u}^\top + \boldsymbol{u}\boldsymbol{\theta}_0^\top + \boldsymbol{u}\boldsymbol{u}^\top \tag{96}$$

To reduce the attack strength, we want to minimise the the attack overlap A which is $\boldsymbol{\theta}^\top\boldsymbol{\Sigma_v}\boldsymbol{\theta}$.

To minimise the term $\hat{\boldsymbol{\theta}}^\top \boldsymbol{u}\boldsymbol{u}^\top \hat{\boldsymbol{\theta}}$, we would like to make the ERM procedure chose a vector $\boldsymbol{\theta}$ that is perpendicular to $\boldsymbol{u}$. This can be done without affecting generalisation performances as $\boldsymbol{u} \perp \boldsymbol{\theta}_0$. A way to remove this, is to regularise in the direction $\boldsymbol{u}$. This can be done with a directional ℓ_2 regularisation or with adversarial training and choosing $\boldsymbol{\Sigma_\delta} = \boldsymbol{u}\boldsymbol{u}^\top$. With this choice of $\boldsymbol{\Sigma_\delta}$, the values $\hat{\boldsymbol{\theta}}^\top \boldsymbol{\theta}_0\boldsymbol{u}^\top\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}^\top \boldsymbol{u}\boldsymbol{\theta}_0^\top\hat{\boldsymbol{\theta}}$ can be decreased in the same way.

For the last term $\hat{\boldsymbol{\theta}}^\top \boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\hat{\boldsymbol{\theta}}$, this recipe does not work. We can interpret this term as being proportional to the square of the overlap m for isotropic data. Thus, it is obvious that protecting such a direction, can only be achieved by reducing the norm of $\hat{\boldsymbol{\theta}}$, which reduces the overlap m, which in turn cannot be changed without hurting the angle between teacher and student and thus generalisation error. This term is thus at the heart of the trade-off between generalisation and boundary term.

Thus shown that all the components of the eigenvectors of $\boldsymbol{\Sigma_v}$ that are perpendicular to the teacher can be "regularised away".

F ANALYSIS OF THE FAST GRADIENT METHOD ADVERSARIAL TRAINING

Sometimes in practical adversarial one does not solve exactly the each-sample maximisation as it is to much computationally intensive for general machine learning models. Thus people turn to Fast Gradient Method (FGM) that corresponds to taking a single gradient step in the direction of max gradient with the biggest possible norm. Unlike explicitly maximising the loss with respect to the data, which requires iterative and computationally intensive optimisation, FGM provides a more efficient, single-step approach. This efficiency makes FGM practical for training robust models on large datasets.

In the case of linear models we have that the objective is

$$\sum_{i=1}^n g\left(y_i \frac{\boldsymbol{\theta}^\top \boldsymbol{x}_i}{\sqrt{d}} - \varepsilon_t \frac{\boldsymbol{\theta}^\top \boldsymbol{\Sigma_\delta}\boldsymbol{\theta}}{\sqrt{d}\|\boldsymbol{\theta}\|_2}\right) \tag{97}$$

Our theoretical toolkit is versatile enough to be adapted to this problem as well. We provide in Appendix I a statistical physics derivation of the two results in parallel. We also believe that also this results holds formally as a theorem.

F.1 Class-preserving adversarial attacks

The ability of explicitly defining a direction allows us to reason about attacks that do not cross the teacher margin that we derive in this paragraph.

We propose a more refined metric called the class-preserving generalisation error. This metric only considers attacks that mislead the student model while not affecting the teacher model's classification or its confidence in that classification. The teacher's confidence is quantified by a margin, denoted as γ, which represents the minimum allowable distance from the decision boundary to consider a classification confident.

This is an artificial metric, as it would be inaccessible to a student. We are interested in measuring the error with respect to the noisy labels by only considering attacks which are fair with respect to the ground truth.

We focus on attacks that attempt to deceive the student model within a specified norm bound but also ensure that the teacher model does not misclassify the perturbed data. The class-preserving generalisation error, incorporating these considerations, is defined as:

$$
\begin{aligned}
E_{\mathrm{CP}}(\varepsilon_g, \gamma) = & \mathbb{E}_{y,\boldsymbol{x}} \left[\max_{\substack{\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}_{\boldsymbol{v}}^{-2}} \leq \varepsilon_g \\ y\boldsymbol{\theta}_0^\top (\boldsymbol{x}+\boldsymbol{\delta}) > \sqrt{d}\gamma}} \mathbb{1}(y \neq \hat{y}(\hat{\boldsymbol{\theta}}, \boldsymbol{x}+\boldsymbol{\delta})) \mathbb{1}(y\boldsymbol{\theta}_0^\top \boldsymbol{x} > \sqrt{d}\gamma) \right] \\
& + \mathbb{E}_{y,\boldsymbol{x}} \left[\mathbb{1}(y \neq \hat{y}(\hat{\boldsymbol{\theta}}, \boldsymbol{x})) \mathbb{1}(y\boldsymbol{\theta}_0^\top \boldsymbol{x} \leq \sqrt{d}\gamma) \right]
\end{aligned}
\tag{98}
$$

The constraint within the maximisation ensure that the perturbation does not cause the teacher model's confidence classification is not diminished beyond the margin γ and that the total length of the perturbation doesn't surpass the value ε_g. The two constraints can be rewritten as

$$
\boldsymbol{\delta}^\top \boldsymbol{\Sigma}_{\boldsymbol{v}}^{-2} \boldsymbol{\delta} \leq \varepsilon_g^2 \,,
\tag{99}
$$

$$
y\boldsymbol{\theta}_0^\top \boldsymbol{\delta} > \gamma - y\boldsymbol{\theta}_0^\top \boldsymbol{x} \,,
\tag{100}
$$

this means that the projection of the perturbation on the teacher should always be greater in absolute value than the projection of the data, *i.e.* the data point classified from the teacher should not change.

To simplify further one can consider the vanishing margin case as in the main text, formally

$$
E_{\mathrm{CP}}(\varepsilon_g) = \lim_{\gamma \to 0^+} E_{\mathrm{CP}}(\varepsilon_g, \gamma) \,.
\tag{101}
$$

We want to fool the student, so we consider a perturbation in the direction of $\hat{\boldsymbol{\theta}}$. Based on different conditions we have different attacks and thus different norm of this perturbation. We will indicate the norm as

$$
\boldsymbol{\delta}_{(\cdot)} = \alpha_{(\cdot)} \frac{\boldsymbol{\Sigma}_{\boldsymbol{v}} \hat{\boldsymbol{\theta}}}{\left\| \hat{\boldsymbol{\theta}} \right\|_2}
\tag{102}
$$

where the subscript can indicate any of the cases and we will define them later.

The first case is if the point $\boldsymbol{x}$ already has a margin with the teacher that it is smaller than γ. The condition for this to happen is that

$$
y\boldsymbol{\theta}_0^\top \boldsymbol{x} < \gamma \quad \implies \quad \alpha_{\mathrm{LM}} = 0 \,, \quad \boldsymbol{\delta}_{\mathrm{LM}} = \boldsymbol{0} \,,
\tag{103}
$$

in this case we do not perturb the input data point.

By just considering Eq. (99) we have that the maximum norm the perturbation can have is

$$
\alpha_{\mathrm{MAX}} = -y\varepsilon_g \,, \quad \boldsymbol{\delta}_{\mathrm{MAX}} = -y\varepsilon_g \frac{\boldsymbol{\Sigma}_{\boldsymbol{v}} \hat{\boldsymbol{\theta}}}{\left\| \hat{\boldsymbol{\theta}} \right\|_2}
\tag{104}
$$

this is the strongest attack, but we can only use it if the perturbed image does not cross the margin γ of the teacher $\boldsymbol{\theta}_0$. In equations, it means that Eq. (100) is satisfied, which means that

$$y\frac{\boldsymbol{\theta}_0^\top (\boldsymbol{x} + \boldsymbol{\delta}_{\mathrm{MAX}})}{\sqrt{d}} \geq \gamma \iff y\frac{\boldsymbol{\theta}_0^\top \boldsymbol{x}}{\sqrt{d}} - \varepsilon_g \frac{\boldsymbol{\theta}_0^\top \boldsymbol{\Sigma_v} \hat{\boldsymbol{\theta}}}{\sqrt{d}\left\|\hat{\boldsymbol{\theta}}\right\|_2} \geq \gamma \tag{105}$$

thus if this previous condition is satisfied we can perturb with $\boldsymbol{\delta}_{\mathrm{MAX}}$.

If instead Eqs. (100) and (105) are not satisfied we can proceed the attack but with a smaller norm, that is found by imposing that the final margin is equal to γ. By doing that we can solve for the final norm and obtain

$$\alpha_\gamma = \frac{\sqrt{d}\left\|\hat{\boldsymbol{\theta}}\right\|_2}{\boldsymbol{\theta}_0^\top \boldsymbol{\Sigma_v} \hat{\boldsymbol{\theta}}}\left(y\gamma - \frac{\boldsymbol{\theta}_0^\top \boldsymbol{x}}{\sqrt{d}}\right), \quad \boldsymbol{\delta}_\gamma = \frac{\sqrt{d}\boldsymbol{\Sigma_v}\hat{\boldsymbol{\theta}}}{\boldsymbol{\theta}_0^\top \boldsymbol{\Sigma_v}\hat{\boldsymbol{\theta}}}\left(y\gamma - \frac{\boldsymbol{\theta}_0^\top \boldsymbol{x}}{\sqrt{d}}\right) \tag{106}$$

where remember that y is just a sign.

One can thus rewrite explicitly the maximisation in Eq. (98) as follows

$$
\begin{aligned}
E_{\mathrm{CP}}(\varepsilon_g, \gamma) = {}& \mathbb{E}_{y,\boldsymbol{x}}\left[\mathbb{1}\left(y \neq \hat{y}(\hat{\boldsymbol{\theta}}, \boldsymbol{x} + \boldsymbol{\delta}_\gamma)\right)\mathbb{1}\left(y\frac{\boldsymbol{\theta}_0^\top \boldsymbol{x}}{\sqrt{d}} - \varepsilon_g \frac{\boldsymbol{\theta}_0^\top \boldsymbol{\Sigma_v}\hat{\boldsymbol{\theta}}}{\sqrt{d}\left\|\hat{\boldsymbol{\theta}}\right\|_2} < \gamma\right)\mathbb{1}(y\boldsymbol{\theta}_0^\top \boldsymbol{x} > \sqrt{d}\gamma)\right] \\
& + \mathbb{E}_{y,\boldsymbol{x}}\left[\mathbb{1}\left(y \neq \hat{y}(\hat{\boldsymbol{\theta}}, \boldsymbol{x} + \boldsymbol{\delta}_{\mathrm{MAX}})\right)\mathbb{1}\left(y\frac{\boldsymbol{\theta}_0^\top \boldsymbol{x}}{\sqrt{d}} - \varepsilon_g \frac{\boldsymbol{\theta}_0^\top \boldsymbol{\Sigma_v}\hat{\boldsymbol{\theta}}}{\sqrt{d}\left\|\hat{\boldsymbol{\theta}}\right\|_2} \geq \gamma\right)\mathbb{1}(y\boldsymbol{\theta}_0^\top \boldsymbol{x} > \sqrt{d}\gamma)\right] \\
& + \mathbb{E}_{y,\boldsymbol{x}}\left[\mathbb{1}(y \neq \hat{y}(\hat{\boldsymbol{\theta}}, \boldsymbol{x}))\mathbb{1}(y\boldsymbol{\theta}_0^\top \boldsymbol{x} \leq \sqrt{d}\gamma)\right]
\end{aligned}
\tag{107}
$$

where each one of the lines correspond to one of the previously explained cases. To compute the expectations in Eq. (107), because of the presence of y one should consider the actual channel that generates the data. Nevertheless the form in Eq. (107) is well suited for estimation from a given test dataset by replacing the expectations with empirical averages over that training dataset.

F.1.1 Noiseless Channel

In the case that the channel is a noiseless sign channel we have

$$
\begin{aligned}
E_{\mathrm{CP}}(\varepsilon_g, \gamma) = {}& \int \mathrm{d}\mu(\nu, \lambda)\,\mathbb{1}\left[\mathrm{sign}(\nu) \neq \mathrm{sign}\left(\lambda + \mathrm{sign}(\nu)\frac{A}{F}(\gamma - |\nu|)\right)\right]\mathbb{1}\left[|\nu| - \varepsilon_g \frac{F}{\sqrt{N}} < \gamma\right]\mathbb{1}[|\nu| > \gamma] \\
& + \int \mathrm{d}\mu(\nu, \lambda)\,\mathbb{1}\left[\mathrm{sign}(\nu) \neq \mathrm{sign}\left(\lambda - \mathrm{sign}(\nu)\varepsilon_g\frac{A}{\sqrt{N}}\right)\right]\mathbb{1}\left[|\nu| - \varepsilon_g\frac{F}{\sqrt{N}} \geq \gamma\right]\mathbb{1}[|\nu| > \gamma] \\
& + \int \mathrm{d}\mu(\nu, \lambda)\,\mathbb{1}[\mathrm{sign}(\nu) \neq \mathrm{sign}(\lambda)]\mathbb{1}[|\nu| \leq \gamma]
\end{aligned}
\tag{108}
$$

where we have that the local fields ν, λ are jointly Gaussian with zero mean and covariance $\left(\begin{smallmatrix} \rho & m \\ m & q \end{smallmatrix}\right)$, this is the probability distribution $\mathrm{d}\mu(\nu, \lambda)$.

To simplify the double integral in Eq. (108) it is a matter of integrating over λ and then computing the integration over ν numerically, as it is in a single variable. In the end, we obtain

$$
\begin{aligned}
E_{\mathrm{CP}}(\varepsilon_g, \gamma) = {}& \int_0^\gamma \mathrm{erfc}\left(\frac{m\nu}{\sqrt{2\rho(q\rho - m^2)}}\right)\mathcal{D}_\rho[\nu] + \int_{\gamma^\star}^\infty \mathrm{erfc}\left(\frac{m\nu - \varepsilon_g\sqrt{N}\rho}{\sqrt{2\rho(q\rho - m^2)}}\right)\mathcal{D}_\rho[\nu] \\
& + \frac{1}{2}\left[\int_\gamma^{\gamma^\star} \mathrm{erfc}\left(\frac{A\rho\gamma + \nu(mF - A\rho)}{F\sqrt{2\rho(q\rho - m^2)}}\right)\mathcal{D}_\rho[\nu] + \int_{-\gamma^\star}^{-\gamma} \mathrm{erfc}\left(\frac{A\rho\gamma - \nu(mF - A\rho)}{F\sqrt{2\rho(q\rho - m^2)}}\right)\mathcal{D}_\rho[\nu]\right]
\end{aligned}
\tag{109}
$$

where the notation $\mathcal{D}_\rho[\nu]$ indicates the p.d.f. of a random variable $\nu \sim \mathcal{N}(0, \rho^2)$ and $\gamma^\star = \max(\gamma, \gamma + \varepsilon_g F/\sqrt{N})$.
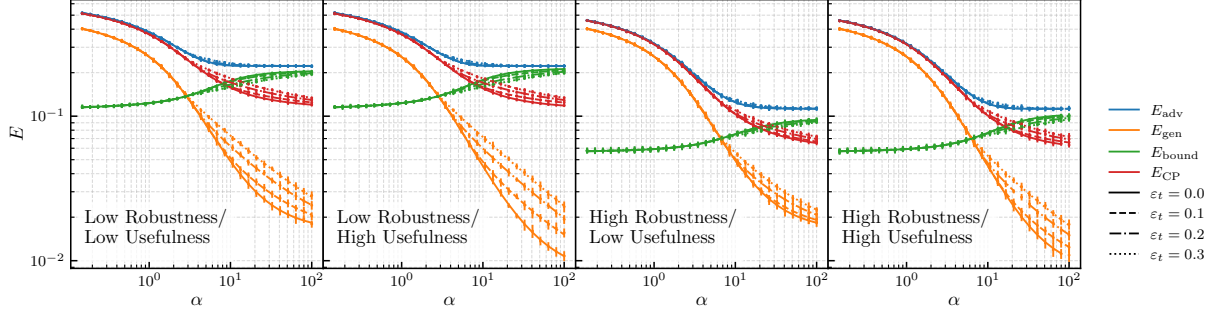
Figure 5: FGM training for single block SWFM with the same setting as in Fig. 1. We see that the qualitative behaviour is very similar among the two figures.
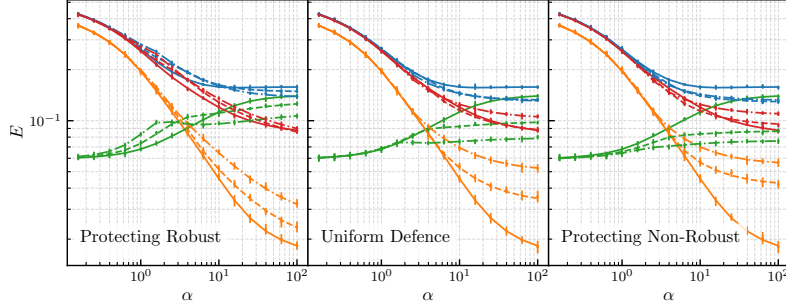
Figure 6: FGM training with the same setting as in Fig. 2 (Left). We see that the qualitative behaviour is very similar among the two figures.

F.1.2 Probit Channel

In the case where the channel is a probit channel we have that the class preserving error could be computed similarly at the addition of one integral. The integration is done over the measure $\mathrm{d}\mu(\nu, \lambda)$ as before but one should add the probability distribution of the channel. Thus we have that the integration becomes

$$
\begin{aligned}
&E_{\mathrm{CP}}(\varepsilon_g, \gamma) \\
&= \int \mathrm{d}y \, \mathrm{d}\mu(\nu, \lambda) \, P(y \mid \nu) \mathbb{1}\left[y \neq \mathrm{sign}\left(\lambda + \mathrm{sign}(\nu + z)\frac{A}{F}(\gamma - |\nu|)\right)\right] \mathbb{1}\left[y\nu - \varepsilon_g \frac{F}{\sqrt{N}} < \gamma\right] \mathbb{1}[y\nu > \gamma] \\
&+ \int \mathrm{d}y \, \mathrm{d}\mu(\nu, \lambda) \, P(y \mid \nu) \mathbb{1}\left[y \neq \mathrm{sign}\left(\lambda - \mathrm{sign}(\nu)\varepsilon_g \frac{A}{\sqrt{N}}\right)\right] \mathbb{1}\left[y\nu - \varepsilon_g \frac{F}{\sqrt{N}} \geq \gamma\right] \mathbb{1}[y\nu > \gamma] \\
&+ \int \mathrm{d}y \, \mathrm{d}\mu(\nu, \lambda) \, P(y \mid \nu) \mathbb{1}[y \neq \mathrm{sign}(\lambda)] \mathbb{1}[y\nu \leq \gamma]
\end{aligned}
\tag{110}
$$

where we have introduced the channel probability distribution.

F.2 Comparison between full minimisation and FGM

From the look of Figs. 5 and 6, that correspond to the same setting as Figs. 1 and 2, we see that the qualitative behaviours are very similar among the two cases.

- In both cases we have that the adversarial error E_{adv} is dominated by generalisation error in the low sample complexity regime and it is dominated by boundary error in the high sample complexity regime.

- In both cases we see that usefulness determines the plateau for generalisation error and robustness the one for boundary error.

Table 1: We isolate robust features and demonstrate that attacking the non-robust features contributes most to boundary error. For CIFAR10, we keep the first 2500 principle components, and evaluate the target 1 vs 9. For FashionMNIST, we keep the first 500 principle components and evaluate odd versus even targets.

CIFAR 10			
ATTACK DIRECTION	E_{train}	E_{gen}	E_{adv}
UNIFORM	0.0749	0.2225	0.2645
ROBUST	0.0749	0.2225	0.2230
NON-ROBUST	0.0749	0.2225	0.2645

FASHION MNIST			
ATTACK DIRECTION	E_{train}	E_{gen}	E_{adv}
UNIFORM	0.0326	0.0396	0.0635
ROBUST	0.0326	0.0396	0.0427
NON-ROBUST	0.0326	0.0396	0.0595

- Among the different protection strategies protecting the non robust features leads to a smaller adversarial generalisation error in the high sample complexity regime.

A small difference that one can notice is that the difference in performances between the three cases considered in Fig. 6 is more marked than the one considered in Fig. 2.

In both plots we the red curves correspond to the class-preserving error explained in Appendix F.1. We can see that this kind of error metric behaves as the adversarial training error: at high sample complexities it has the same value as the generalistion error while increasing the number of training points decreases its value. One key difference is that the lowest value reached by E_{CP} is lower than the one reached by E_{adv}. This behaviour can be explained as follows. As the sample complexity increases the student vector $\hat{\boldsymbol{\theta}}$ aligns with the teacher vector $\boldsymbol{\theta}_0$ and in the case of E_{adv} a lot of points start crossing the boundary leading to an increase of the boundary error E_{bnd}. In the case of the class preserving error these points will be perturbed to not misclassified as the boundary for the teacher and the students are almost the same and the adversarial perturbation cannot flip the teacher labels. For this reason the class-preserving error is an intermediate error metric between generalisation and adversarial error.

G   EXPERIMENTS ON REAL DATA

In this Appendix we demonstrate the robustness metric and defense-strategies evaluated the Cifar 10 Alex (2009) and FashionMNIST Xiao et al. (2017) datasets.

G.1   Finding Non-Robust Features

The margin-based robustness given in Eq. (7) can be interpreted in terms of the student estimate $\boldsymbol{\theta}$ and it can be written as

$$\mathcal{R}_j^{\text{emp.}} = \sum_{i=1}^{n} y_i \boldsymbol{\theta}_j(\boldsymbol{x}_i)_j - \varepsilon_g \sqrt{\boldsymbol{\theta}^\top \boldsymbol{\Sigma_\delta} \boldsymbol{\theta}} \tag{111}$$

which can be evaluated component-wise.

Our aim is to apply this measure on components obtained by PCA. The motivation for choosing PCA as a method of feature extraction and dimensionality reduction is twofold, on the one hand it creates uncorrelated variables and on the other hand, it creates them in order of explained variance. This produces a data-set similar to the BFM with a power-law distributed spectrum.

We begin by computing the principal components and projecting all test and training data. Next, we optimise the weights with $\varepsilon_t = 0$. Next, we increase ε_t and evaluate the component-wise robustness Eq. (111).

Table 2: We protect the non-robust features and find that identity and inverse proportionality to the explained variance is good compared to protecting proportional to the explained variance.

CIFAR 10			
PROTECTION DIRECTION	E_{train}	E_{gen}	E_{adv}
UNIFORM	0.0802	0.1950	0.2255
PROPORTIONAL	0.0736	0.2210	0.2625
INVERSE	0.0873	0.1965	0.2230

FASHION MNIST			
PROTECTION DIRECTION	E_{train}	E_{gen}	E_{adv}
UNIFORM	0.0352	0.0395	0.0493
PROPORTIONAL	0.0337	0.0391	0.0558
INVERSE	0.0354	0.0392	0.0503

By hand, we choose a cutoff-value (2000 for Cifar10 and for 20000 FashionMNIST) for the component-wise robustness-measure to produce a mask for robust features. All values higher than the cutoff are considered to be robust. We choose the constant such that the boundary error is explained to a large degree through the non-robust features. This is evaluated by constraining the attack to all features, the robust features and the non-robust features. The results of this experiment is shown in Table 1.

G.2 Training the Non-Robust Features Adversarially

A principle finding of ours is the discussion of different defence strategies. Here, we provide some simple experiments on real data.

In the main text, we suggested to defend the non-robust features either proportional to their variance, inverse proportional to their variance or uniformly. We continue the experiment from the previous paragraph and consider the PCA explained variance as the relevant metric for the proportionality. To be explicit, we perform three different adversarial training runs with different directional choices for $\boldsymbol{\Sigma_\delta}$. We define the uniform defence by having $\boldsymbol{\Sigma_\delta}$ equal to the identity, the proportional defence by setting the directions proportional to their explained variance and the inverse defence by setting the directions proportional to the inverse of the explained variance. The results of this experiment are given in Table 2. Indeed, protecting inversely or uniformly appears to work well.

H SETTINGS OF MAIN TEXT FIGURES

In this Appendix we provide the parameters for the figures presented in the main text. For these experimental parts of the paper, we fix the function $g(x) = \log(1 + \exp(-x))$ to be the logistic loss.

H.1 Setting Figure 1

In Fig. 1 we consider we consider a SWFM with a single block. Considering the figures from left to right, the entries for the parameter specifying the block are

Table 3: Parameters for Fig. 1

Plot Label	ψ_ℓ	t_ℓ	Δ_ℓ	Υ_ℓ	ε_g	λ	τ
Low Robustness/Low Usefulness	0.5	2	1	1	0.2	10^{-3}	0.05
Low Robustness/High Usefulness	0.5	8	1	1	0.2	10^{-3}	0.05
High Robustness/Low Usefulness	2	0.5	1	1	0.2	10^{-3}	0.05
High Robustness/High Usefulness	2	2	1	1	0.2	10^{-3}	0.05

The points describe the simulation using ERM routine, where we have used a dimension $d = 1000$ and averaged over 20 runs. We show the mean and the standard deviation on the mean.

H.2 Setting Figure 2

In Fig. 2 there are three blocks of images, we start from the first block on the left. Here we show an SWFM consisting of two blocks $k = 2$. The values are filled as follows

Table 4: Parameters for Fig. 2 (Left)

Plot Label	(ψ_1, ψ_2)	(t_1, t_2)	(Δ_1, Δ_2)	(Υ_1, Υ_2)	ε_g	λ	τ
Protecting Robust	(5,0.2)	(1,1)	(2,1)	(1,1)	0.2	10^{-3}	0.05
Uniform Defence	(5,0.2)	(1,1)	(1,1)	(1,1)	0.2	10^{-3}	0.05
Protecting Non-Robust	(5,0.2)	(1,1)	(1,2)	(1,1)	0.2	10^{-3}	0.05

All matrices are normalised by their trace. The points describe the simulation using ERM, where we have used a dimension $d = 1000$ and averaged over 20 runs. We show the mean and the standard deviation on the mean.

The figure on the top right, shows a power-law spectrum in ψ_ℓ, we begin the indexing with 1 and choose a coefficient $\beta = 1.5$. Again, all the matrices are normalised by their trace. The values are constructed as follows, we consider the block dimension $d_\ell = 1$, and use the following values

Table 5: Parameters for Fig. 2 (Top, Right) and (Bottom, Left)

ψ_ℓ	t_ℓ	Δ_ℓ	Υ_ℓ	ε_g	λ	τ
$i^{-\beta}$	1	1	1	0.2	10^{-3}	0.05

In the figure on the bottom right, we show different power-law behaviour when $\alpha \to \infty$. Again, we normalise each finite size power-law by their trace, we fix $d = 1000$. The parameters are filled exactly as in the previously described figure, the values can be found in Table 5.

H.3 Setting Figure 3

In Fig. 3 for both the center and rigth plot we considered a BFM with a single block $k = 1$. We fix $\tau = 0.05$ and $\varepsilon_g = 0.2$.

In both figures, we consider $t_\ell = 1$, $\psi_\ell = 1$, $\Delta_\ell = 1$, $\Upsilon_i = 1$ and the values for the ε_t are shown in the legend.

In the center figure, we use our state evolution equations to optimise the regularisation strength in terms of generalisation error. Each simulation point hence is created at its respective optimal regularisation.

In the right figure, we compare the learning curves of standard adversarial training with the learning curves of the approximately equivalent loss, for this figure we fix the regularisation strength $\lambda = 10^{-3}$.

In the left figure we build a SWFM model where we fix two different blocks of features. The percentage for the two features are $(\phi_1, \phi_2) = (0.01, 0.99)$. The experiment is performed at a high sample complexity $\alpha = 10^2$. The other values for the plots are

Table 6: Parameters for Fig. 3

Plot Label	(ψ_1, ψ_2)	(t_1, t_2)	(Δ_1, Δ_2)	(Υ_1, Υ_2)	ε_g	λ	τ
Top	(1,1)	$(10^3,1)$	$\mathrm{diag}(\mathbf{\Sigma}_{\boldsymbol{\delta}}^b)$	$\mathrm{diag}(\mathbf{\Sigma}_{\boldsymbol{\upsilon}}^b)$	0.06	10^{-3}	0.05
Bottom	(1,1)	$(10^3,1)$	$(10^3,1)$	$(10^3,1)$	0.06	10^{-3}	0.05

In the bottom plot, we compose a vector orthogonal to the diagonal to the teacher covariance $\boldsymbol{v} \perp \mathrm{diag}\,\mathbf{\Sigma}_{\boldsymbol{\theta}}$. Then, we choose $\mathbf{\Sigma}_{\boldsymbol{\delta}}^b = \mathbf{\Sigma}_{\boldsymbol{\upsilon}}^b = a\boldsymbol{v}\boldsymbol{v}^\top + N$. We add a variance 1 Gaussian noise matrix N to make the matrix non-singular and we increase the importance of the first component through a factor a. All matrices are normalised by their trace.

I   STATISTICAL PHYSICS DERIVATION OF THE MAIN RESULT

In this Section we give a full derivation of the results in Theorem 3.1 by means of the replica approach, a standard method developed in the realm of statistical physics of disordered systems. Our computational approach can be found in Loureiro et al. (2022a); Gerace et al. (2021); Aubin et al. (2020). For a foundational understanding of this effective yet heuristic method, we suggest the following books Mezard et al. (1987); Mezard and Montanari (2009).

Between the two computations of the FG method and the true minimisation we chose to present the first one as it is the one more involved and one could obtain the other one by a little simplification when one has to derive the final saddle point equations.

I.1   Gibbs minimisation

The starting point is to define the following Gibbs measure over weights $\boldsymbol{\theta} \in \mathbb{R}^d$. We want the most probable states to be the ones that minimises the ERM problem in the first place. Then to select only these states we will consider the zero temperature limit by taking the parameter $\beta \to \infty$. The measure that we are interested in is

$$\mu_\beta(\mathrm{d}\boldsymbol{\theta}) = \frac{1}{\mathcal{Z}_\beta} e^{-\beta\left[\sum_{\mu=1}^n g\left(y^\mu, \boldsymbol{\theta}^\top \boldsymbol{x}^\mu, \boldsymbol{\theta}, \mathbf{\Sigma}_{\boldsymbol{\delta}}, \varepsilon_t\right) + \frac{\lambda}{2} \boldsymbol{\theta}^T \mathbf{\Sigma}_{\boldsymbol{w}} \boldsymbol{\theta}\right]} \mathrm{d}\boldsymbol{\theta} = \frac{1}{\mathcal{Z}_\beta} \underbrace{\prod_{\mu=1}^n e^{-\beta g\left(y^\mu, \boldsymbol{\theta}^\top \boldsymbol{x}^\mu, \boldsymbol{\theta}, \mathbf{\Sigma}_{\boldsymbol{\delta}}, \varepsilon_t\right)}}_{P_g} \underbrace{e^{-\frac{\beta\lambda}{2} \boldsymbol{\theta}^T \mathbf{\Sigma}_{\boldsymbol{w}} \boldsymbol{\theta}} \, \mathrm{d}w_i}_{P_w} \quad (112)$$

Where P_g is the probability distribution associated with the channel and P_w is the prior probability distribution.

Here, $\mathcal{Z}_\beta$, is the partition function that normalises the Gibbs measure and it is given by

$$\mathcal{Z}_\beta = \int_{\mathbb{R}^d} \mathrm{d}\boldsymbol{\theta} e^{-\frac{\beta\lambda}{2} \boldsymbol{\theta}^T \mathbf{\Sigma}_{\boldsymbol{w}} \boldsymbol{\theta}} \prod_{\mu=1}^n e^{-\beta g\left(y^\mu, \boldsymbol{\theta}^\top \boldsymbol{x}^\mu, \boldsymbol{\theta}, \mathbf{\Sigma}_{\boldsymbol{\delta}}, \varepsilon_t\right)} \quad (113)$$

You do need attention, but the free energy density is sufficient. In the zero temperature limit, $\beta \to \infty$ the Gibbs measure in Eq. (112) concentrates around the solutions of the ERM problem. With the replica method, we can compute the free energy density, it is given by:

$$\beta f_\beta = -\lim_{d\to\infty} \frac{1}{d} \mathbb{E}_{\mathcal{D}} \log \mathcal{Z}_\beta \quad (114)$$

To evaluate the quenched average of the free energy is to use the replica trick

$$\lim_{d\to\infty} \frac{1}{d} \mathbb{E}_{\mathcal{D}} \log \mathcal{Z}_\beta = \lim_{r\to 0} \lim_{d\to\infty} \frac{1}{d} \frac{\partial_r \mathbb{E}_{\mathcal{D}} \mathcal{Z}^r}{1} \quad (115)$$

Note that we introduced three limits up to here. The first is the zero temperature limit ensuring that we find the ground state of our Gibbs measure which corresponds to the minimum of our ERM problem. The second is the thermodynamic limit of very large dimension whilst keeping the sampling ratio fixed. And the third limit stems from the replica trick allowing us to compute the logarithm of the partition function, it corresponds to setting the number of replicated systems to zero.

This computation follows for the first part the one in Loureiro et al. (2022a). So we start with the initial definition of replicated partition function the difference we have in our case is that we have a dependence on ε_t on the output probability.

$$
\begin{aligned}
\mathbb{E}_{\mathcal{D}} \mathcal{Z}_\beta^r &= \prod_{\mu=1}^n \mathbb{E}_{\boldsymbol{x}^\mu} \prod_{a=1}^r \int_{\mathbb{R}^d} P_w\left(\mathrm{d}\boldsymbol{\theta}^a\right) P\left(y^\mu \mid \frac{\boldsymbol{x}^\mu \cdot \boldsymbol{\theta}^a}{\sqrt{d}}\right) \\
&= \prod_{\mu=1}^n \int_{\mathbb{R}} \mathrm{d}y^\mu \int_{\mathbb{R}^p} P_{\boldsymbol{\theta}_0}\left(\mathrm{d}\boldsymbol{\theta}_0\right) \int_{\mathbb{R}^{d \times r}} \left(\prod_{a=1}^r P_w\left(\mathrm{d}\boldsymbol{\theta}^a\right)\right) \mathbb{E}_{\boldsymbol{x}^\mu}\left[P_0\left(y^\mu \mid \frac{\boldsymbol{x}^\mu \cdot \boldsymbol{\theta}_0}{\sqrt{d}}\right) \prod_{a=1}^r P_g\left(y^\mu \mid \frac{\boldsymbol{x}^\mu \cdot \boldsymbol{\theta}^a}{\sqrt{d}}, \boldsymbol{\Sigma}_{\boldsymbol{\delta}}, \boldsymbol{\theta}^a, \varepsilon_t\right)\right]
\end{aligned}
$$

(116)

explicitly we have that the term in P_g is in the case of the true adversarial attack

$$
P_g\left(y^\mu \mid \frac{\boldsymbol{x}^\mu \cdot \boldsymbol{\theta}^a}{\sqrt{d}}, \boldsymbol{\Sigma}_{\boldsymbol{\delta}}, \boldsymbol{\theta}^a, \varepsilon_t\right) = \frac{\sqrt{\beta}}{\sqrt{2\pi}} \exp\left(-\beta g\left(y \frac{\boldsymbol{x}^\mu \cdot \boldsymbol{\theta}^a}{\sqrt{d}} - \frac{\varepsilon_t}{\sqrt{d}}\sqrt{\boldsymbol{\theta}^a \boldsymbol{\Sigma}_{\boldsymbol{\delta}} \boldsymbol{\theta}^a}\right)\right)
$$

(117)

or in the case of a FGM attack

$$
P_g\left(y^\mu \mid \frac{\boldsymbol{x}^\mu \cdot \boldsymbol{\theta}^a}{\sqrt{d}}, \boldsymbol{\Sigma}_{\boldsymbol{\delta}}, \boldsymbol{\theta}^a, \varepsilon_t\right) = \frac{\sqrt{\beta}}{\sqrt{2\pi}} \exp\left(-\beta g\left(y \frac{\boldsymbol{x}^\mu \cdot \boldsymbol{\theta}^a}{\sqrt{d}} - \frac{\varepsilon_t}{\sqrt{d}}\frac{\boldsymbol{\theta}^a \boldsymbol{\Sigma}_{\boldsymbol{\delta}} \boldsymbol{\theta}^a}{\|\boldsymbol{\theta}^a\|_2}\right)\right)
$$

(118)

note that P_0 can be a general noisy channel distribution. the expectation part is equal to:

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}^\mu}\left[P_0\left(y^\mu \mid \frac{\boldsymbol{x}^\mu \cdot \boldsymbol{\theta}_0}{\sqrt{d}}\right) \prod_{a=1}^r P_g\left(y^\mu \mid \frac{\boldsymbol{x}^\mu \cdot \boldsymbol{\theta}^a}{\sqrt{d}}, \boldsymbol{\Sigma}_{\boldsymbol{\delta}}, \boldsymbol{\theta}^a, \varepsilon_t\right)\right] \\
&= \int_{\mathbb{R}} \mathrm{d}\nu_\mu P_0\left(y \mid \nu_\mu\right) \int_{\mathbb{R}^r}\left(\prod_{a=1}^r \mathrm{d}\lambda_\mu^a P_g\left(y^\mu \mid \lambda_\mu^a, \boldsymbol{\Sigma}_{\boldsymbol{\delta}}, \boldsymbol{\theta}^a, \varepsilon_t\right)\right) \mathbb{E}_{\boldsymbol{x}^\mu}\left[\delta\left(\nu_\mu - \frac{\boldsymbol{x}^\mu \cdot \boldsymbol{\theta}_0}{\sqrt{d}}\right) \prod_{a=1}^r \delta\left(\lambda_\mu^a - \frac{\boldsymbol{x}^\mu \cdot \boldsymbol{\theta}^a}{\sqrt{d}}\right)\right]
\end{aligned}
$$

(119)

We can still perform the average over the dataset. We have that the new variables will behave again as Gaussians with the following covariances:

$$
\rho \equiv \mathbb{E}\left[\nu_\mu^2\right] = \frac{1}{d}\boldsymbol{\theta}_0^\top \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{\theta}_0, \quad m^a \equiv \mathbb{E}\left[\lambda_\mu^a \nu_\mu\right] = \frac{1}{d}\boldsymbol{\theta}_0^\top \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{\theta}^a, \quad Q^{ab} \equiv \mathbb{E}\left[\lambda_\mu^a \lambda_\mu^b\right] = \frac{1}{d}\boldsymbol{\theta}^{a\top} \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{\theta}^b
$$

(120)

where one can organise them in a single covariance matrix.

Now we want to perform several change of variables. The first one is the one in the matrix of overlaps:

$$
\begin{aligned}
1 &\propto \int_{\mathbb{R}} \mathrm{d}\rho \, \delta\left(d\rho - \boldsymbol{\theta}_0^\top \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{\theta}_0\right) \int_{\mathbb{R}^r} \prod_{a=1}^r \mathrm{d}m^a \, \delta\left(dm^a - \boldsymbol{\theta}_0^\top \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{\theta}^a\right) \int_{\mathbb{R}^{r \times r}} \prod_{1 \le a \le b \le r} \mathrm{d}Q^{ab} \, \delta\left(dQ^{ab} - \boldsymbol{\theta}^{a\top} \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{\theta}^b\right) \\
&= \int_{\mathbb{R}} \frac{\mathrm{d}\rho \, \mathrm{d}\hat\rho}{2\pi} e^{-i\hat\rho\left(d\rho - \boldsymbol{\theta}_0^\top \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{\theta}_0\right)} \int_{\mathbb{R}^r} \prod_{a=1}^r \frac{\mathrm{d}m^a \, \mathrm{d}\hat m^a}{2\pi} e^{-i\sum_{a=1}^r \hat m^a\left(dm^a - \boldsymbol{\theta}_0^\top \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{\theta}^a\right)} \int_{\mathbb{R}^{r \times r}} \\
&\quad \prod_{1 \le a \le b \le r} \frac{\mathrm{d}Q^{ab} \, \mathrm{d}\hat Q^{ab}}{2\pi} e^{-i\hat Q^{ab}\left(dQ^{ab} - \boldsymbol{\theta}^{a\top} \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{\theta}^b\right)}
\end{aligned}
$$

(121)

We also would like to define new overlaps which are

$$
P = \frac{1}{d}\boldsymbol{\theta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\delta}} \boldsymbol{\theta}, \quad A = \frac{1}{d}\boldsymbol{\theta}^\top \boldsymbol{\Sigma}_{\boldsymbol{v}} \boldsymbol{\theta}, \quad F = \frac{1}{d}\boldsymbol{\theta}_0^\top \boldsymbol{\Sigma}_{\boldsymbol{v}} \boldsymbol{\theta}, \quad N = \frac{1}{d}\boldsymbol{\theta}^\top \boldsymbol{\theta},
$$

(122)

which enter into the computation as follows

$$
1 \propto \int \prod_{a=1}^{r} \mathrm{d}P^a \, \delta(dP^a - \boldsymbol{\theta}^a \boldsymbol{\Sigma_\delta} \boldsymbol{\theta}^a) \int \prod_{a=1}^{r} \mathrm{d}A^a \, \delta(dA^a - \boldsymbol{\theta}^a \boldsymbol{\Sigma_v} \boldsymbol{\theta}^a)
$$

$$
\int_{\mathbb{R}^r} \prod_{a=1}^{r} \mathrm{d}F^a \, \delta\left(dF^a - \boldsymbol{\theta}_0^\top \boldsymbol{\Sigma_v} \boldsymbol{\theta}^a\right) \int \prod_{a=1}^{r} \mathrm{d}N^a \, \delta(dN^a - \boldsymbol{\theta}^a \cdot \boldsymbol{\theta}^a)
$$

$$
= \int \prod_{a=1}^{r} \frac{\mathrm{d}P^a \, \mathrm{d}\hat{P}^a}{2\pi} e^{-i\hat{P}^a(dP^a - \boldsymbol{\theta}^a \boldsymbol{\Sigma_\delta} \boldsymbol{\theta}^a)} \int \prod_{a=1}^{r} \frac{\mathrm{d}A^a \, \mathrm{d}\hat{A}^a}{2\pi} e^{-i\hat{A}^a(dA^a - \boldsymbol{\theta}^a \boldsymbol{\Sigma_v} \boldsymbol{\theta}^a)}
$$
(123)

$$
\int_{\mathbb{R}^r} \prod_{a=1}^{r} \frac{\mathrm{d}F^a \, \mathrm{d}\hat{F}^a}{2\pi} e^{-i\sum_{a=1}^{r} \hat{F}^a \left(dF^a - \boldsymbol{\theta}_0^\top \boldsymbol{\Sigma_v} \boldsymbol{\theta}^a\right)} \int \prod_{a=1}^{r} \frac{\mathrm{d}N^a \, \mathrm{d}\hat{N}^a}{2\pi} e^{-i\hat{N}^a(dN^a - \boldsymbol{\theta}^a \cdot \boldsymbol{\theta}^a)}
$$

notice that the overlap N enters the computation only if we are considering the FGM attack and not for the true minimisation attack. In the following of this computation we will try to remain as general as possible considering both cases.

We finally can write our replicated partition function as the integral of a functional as follows

$$
\mathbb{E}_{\mathcal{D}} \mathscr{Z}_\beta^r = \int \frac{\mathrm{d}\rho \, \mathrm{d}\hat{\rho}}{2\pi} \prod_{a=1}^{r} \frac{\mathrm{d}m^a \, \mathrm{d}\hat{m}^a}{2\pi} \frac{\mathrm{d}P^a \, \mathrm{d}\hat{P}^a}{2\pi} \frac{\mathrm{d}A^a \, \mathrm{d}\hat{A}^a}{2\pi} \frac{\mathrm{d}F^a \, \mathrm{d}\hat{F}^a}{2\pi} \frac{\mathrm{d}N^a \, \mathrm{d}\hat{N}^a}{2\pi} \prod_{1 \le a \le b \le r} \frac{\mathrm{d}Q^{ab} \, \mathrm{d}\hat{Q}^{ab}}{2\pi} e^{d\Phi^{(r)}} \quad (124)
$$

where the r times replicated functional $\Phi^{(r)}$ is

$$
\Phi^{(r)} = -\rho\hat{\rho} - \sum_{a=1}^{r} m^a \hat{m}^a - \sum_{1 \le a \le b \le r} Q^{ab} \hat{Q}^{ab} - \sum_{a=1}^{r} N^a \hat{N}^a - \sum_{a=1}^{r} A^a \hat{A}^a - \sum_{a=1}^{r} P^a \hat{P}^a - \sum_{a=1}^{r} F^a \hat{F}^a
$$
(125)
$$
+ \alpha \Psi_y^{(r)} \left(\rho, m^a, Q^{ab}, A^a, N^a, P^A, F^a\right) + \Psi_w^{(r)} \left(\hat{\rho}, \hat{m}^a, \hat{Q}^{ab}, \hat{A}^a, \hat{N}^a, \hat{P}^a, \hat{F}^a\right)
$$

we will refer to the elements in the first line of Eq. (125) as the trace term. Note in Eq. (124) we factored out d such that we can later evaluate the partition function in the thermodynamic limit using Laplace's method. We also have defined the prior part of the free energy Ψ_w to be

$$
\Psi_w^{(r)} = \frac{1}{d} \log \left[\iint_{\mathbb{R}^d} P_{\boldsymbol{\theta}_0} \left(\mathrm{d}\boldsymbol{\theta}_0\right) e^{\hat{\rho}\boldsymbol{\theta}_0^\top \boldsymbol{\Sigma_x} \boldsymbol{\theta}_0} \right.
$$

$$
\left. \int_{\mathbb{R}^{d \times r}} \prod_{a=1}^{r} P_w \left(\mathrm{d}\boldsymbol{\theta}^a\right) e^{\sum_{a=1}^{r} \left(\hat{m}^a \boldsymbol{\theta}_0^\top \boldsymbol{\Sigma_x} \boldsymbol{\theta}^a + \hat{A}^a \boldsymbol{\theta}^{a\top} \boldsymbol{\Sigma_v} \boldsymbol{\theta}^a + \hat{P}^a \boldsymbol{\theta}^{a\top} \boldsymbol{\Sigma_\delta} \boldsymbol{\theta}^a + \hat{F}^a \boldsymbol{\theta}^{a\top} \boldsymbol{\Sigma_v} \boldsymbol{\theta}^a + \hat{N}^a \boldsymbol{\theta}^a \cdot \boldsymbol{\theta}^a\right) + \sum_{1 \le a \le b \le r} \left(\hat{Q}^{ab} \boldsymbol{\theta}^{a\top} \boldsymbol{\Sigma_x} \boldsymbol{\theta}^b\right)} \right]
$$
(126)

and the channel part of the free energy Ψ_y as

$$
\Psi_y^{(r)} = \log \left[\int_{\mathbb{R}} \mathrm{d}y \int_{\mathbb{R}} \mathrm{d}\nu P_0(y \mid \nu) \int \prod_{a=1}^{r} \mathrm{d}\lambda^a P_g \left(y \mid \lambda^a, P^a, N^a, \varepsilon_t\right) \mathcal{N} \left(\nu, \lambda^a; \mathbf{0}, \Sigma^{ab}\right) \right] \quad (127)
$$

where we have used the fact that $(\nu_\mu, \lambda_\mu) \, \mu = 1, \ldots n$ factors over all the data points.

In the thermodynamic limit where $d \to \infty$ with n/d fixed, the integral in Eq. (124) concentrates around the values of the overlap parameters that extremize the free entropy $\Phi^{(r)}$ and hence we can get the free energy density as:

$$
\beta f_\beta = -\lim_{r \to 0^+} \frac{1}{r} \operatorname{extr} \Phi^{(r)} = -\lim_{r \to 0^+} \partial_r \operatorname{extr} \Phi^{(r)} \quad (128)
$$

I.2 Replica Symmetric Ansatz

We propose the following Ansatz for the variables that we have to extremise over

$$
\begin{aligned}
m^a &= m & \hat{m}^a &= \hat{m} & \text{for } a = 1, \cdots, r \\
q^{aa} &= Q & \hat{q}^{aa} &= -\tfrac{1}{2}\hat{Q} & \text{for } a = 1, \cdots, r \\
q^{ab} &= q & \hat{q}^{ab} &= \hat{q} & \text{for } 1 \le a < b \le r \\
P^a &= P & \hat{P}^a &= -\tfrac{1}{2}\hat{P} & \text{for } a = 1, \cdots, r \\
N^a &= N & \hat{N}^a &= -\tfrac{1}{2}\hat{N} & \text{for } a = 1, \cdots, r \\
A^a &= A & \hat{A}^a &= -\tfrac{1}{2}\hat{A} & \text{for } a = 1, \cdots, r \\
F^a &= F & \hat{F}^a &= \hat{F} & \text{for } a = 1, \cdots, r
\end{aligned}
\tag{129}
$$

Before we take the replica zero limit, let's check that our ansatz above is well-defined and does not have an order one term in $\Phi^{(r)}$ that diverges. For this, we need to ensure that $\lim_{r \to 0^+} \Phi^{(r)} = 0$. The trace terms depends on r except for $\rho\hat{\rho}$, $\lim_{r \to 0^+} \Psi_y^{(r)} = \lim_{r \to 0^+} \Psi_w^{(r)} = 0$ hold. Note that $\lim_{r \to 0^+} \log C^r = 0$ but $\lim_{r \to 0^+} \partial_r \log C^r = \log C$, and thus all we need to check is the prior part of the free energy in the zero replica limit.

$$
\lim_{r \to 0^+} \Phi^{(r)} = -\rho\hat{\rho}
\tag{130}
$$

For this limit to be zero, we must fix $\hat{\rho} = 0$ and note that ρ is a constant we fixed earlier.

Plugging in the Ansatz, the trace term becomes

$$
-\rho\hat{\rho} - rm\hat{m} - \frac{r(r-1)}{2}q\hat{q} + \frac{r}{2}Q\hat{Q} + \frac{r}{2}A\hat{A} + \frac{r}{2}N\hat{N} + \frac{r}{2}P\hat{P} - rF\hat{F}
\tag{131}
$$

Now we take the limit $r \to 0$ after dividing the trace term T (which is no longer an actual trace as we introduced overlaps beyond the traditional replica matrix ansatz) by r

$$
T = \frac{1}{2}(q\hat{q} + Q\hat{Q}) + \frac{1}{2}P\hat{P} + \frac{1}{2}A\hat{A} + \frac{1}{2}N\hat{N} - m\hat{m} - F\hat{F}
\tag{132}
$$

We now define $V = Q - q$ and $\hat{V} = \hat{Q} + \hat{q}$ and rewrite the trace term as follows (by replacing q and $\hat{Q}$).

$$
\begin{aligned}
T &= \frac{1}{2}(q\hat{q} + (V + q)(\hat{V} - \hat{q})) + \frac{1}{2}P\hat{P} + \frac{1}{2}A\hat{A} + \frac{1}{2}N\hat{N} - m\hat{m} - F\hat{F} \\
&= \frac{1}{2}(V\hat{V} + q\hat{V} - V\hat{q}) + \frac{1}{2}P\hat{P} + \frac{1}{2}A\hat{A} + \frac{1}{2}N\hat{N} - m\hat{m} - F\hat{F}
\end{aligned}
\tag{133}
$$

I.2.1 Prior Replica Zero Limit

Thus we can proceed plug these ansätze inside Eqs. (126) and (127) we obtain the following for the prior term

$$
\begin{aligned}
\Psi_w^{(r)} = \frac{1}{d} \log \Bigg[&\int_{\mathbb{R}^d} P_{\boldsymbol{\theta}_0}\,(\mathrm{d}\boldsymbol{\theta}_0)\, e^{\hat{\rho}\boldsymbol{\theta}_0^\top \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{\theta}_0} \int_{\mathbb{R}^{d \times r}} \prod_{a=1}^r P_w(\mathrm{d}\boldsymbol{\theta}^a) e^{\sum_{a=1}^r \left(\hat{m}\boldsymbol{\theta}_0^\top \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{\theta}^a \hat{F} \boldsymbol{\theta}^{a\top} \boldsymbol{\Sigma}_{\boldsymbol{v}} \boldsymbol{\theta}^a \right) + \sum_{1 \le a < b \le r} \left(\hat{q}\boldsymbol{\theta}^{a\top} \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{\theta}^b \right)} \\
&e^{-\frac{1}{2} \sum_{a=1}^r \left(\hat{Q}\boldsymbol{\theta}^{a\top} \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{\theta}^a + \hat{A}\boldsymbol{\theta}^{a\top} \boldsymbol{\Sigma}_{\boldsymbol{v}} \boldsymbol{\theta}^a + \hat{P}\boldsymbol{\theta}^{a\top} \boldsymbol{\Sigma}_{\boldsymbol{\delta}} \boldsymbol{\theta}^a + \hat{N}\boldsymbol{\theta}^a \cdot \boldsymbol{\theta}^a \right)} \Bigg]
\end{aligned}
\tag{134}
$$

to perform in the following the $r \to 0^+$ limit we can change a bit the integral by factoring out all the terms.

To perform this simplification we will use the multidimensional Hubbard-Stratonovic identity which reads

$$
e^{\frac{1}{2} \sum_{a,b=1}^r \boldsymbol{\theta}^{a\top} [\hat{q}\boldsymbol{\Sigma}_{\boldsymbol{x}}] \boldsymbol{\theta}^b} = \mathbb{E}_{\boldsymbol{\xi}} \left[e^{\boldsymbol{\xi}^\top \sqrt{\hat{q}\boldsymbol{\Sigma}_{\boldsymbol{x}}} \sum_{a=1}^r \boldsymbol{\theta}^a} \right]
\tag{135}
$$

where $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, 1_d)$.

Thus by calling the part inside the log in Eq. (134) with the letter $\mathcal{A}$ we have that (putting in $\hat{\rho} = 0$)

$$
\begin{aligned}
\mathcal{A} =& \mathbb{E}_{\boldsymbol{\theta}_0} \int_{\mathbb{R}^{d \times r}} \prod_{a=1}^{r} P_w\left(\mathrm{d}\boldsymbol{\theta}^a\right) e^{-\sum_{a=1}^{r}\left(\hat{m}\boldsymbol{\theta}_0^\top \boldsymbol{\Sigma}_x \boldsymbol{\theta}^a - \hat{F}\boldsymbol{\theta}_0^\top \boldsymbol{\Sigma}_v \boldsymbol{\theta}^a\right) + \frac{1}{2}\sum_{1 \le a, b \le r} \hat{q}\boldsymbol{\theta}^{a\top} \boldsymbol{\Sigma}_x \boldsymbol{\theta}^b} \\
& \cdot e^{-\frac{1}{2}\sum_{a=1}^{r}\left(\hat{V}\boldsymbol{\theta}^{a\top} \boldsymbol{\Sigma}_x \boldsymbol{\theta}^a + \hat{A}\boldsymbol{\theta}^{a\top} \boldsymbol{\Sigma}_v \boldsymbol{\theta}^a + \hat{P}\boldsymbol{\theta}^{a\top} \boldsymbol{\Sigma}_\delta \boldsymbol{\theta}^a + \hat{N}\boldsymbol{\theta}^a \cdot \boldsymbol{\theta}^a\right)} \\
=& \mathbb{E}_{\boldsymbol{\theta}_0} \int_{\mathbb{R}^{d \times r}} \prod_{a=1}^{r} P_w\left(\mathrm{d}\boldsymbol{\theta}^a\right) e^{-\sum_{a=1}^{r}\left(\frac{\hat{V}}{2}\boldsymbol{\theta}^a \boldsymbol{\Sigma}_x \boldsymbol{\theta}^a + \frac{\hat{P}}{2}\boldsymbol{\theta}^a \boldsymbol{\Sigma}_\delta \boldsymbol{\theta}^a + \frac{\hat{A}}{2}\boldsymbol{\theta}^a \boldsymbol{\Sigma}_v \boldsymbol{\theta}^a + \frac{\hat{N}}{2}\boldsymbol{\theta}^a \boldsymbol{\theta}^a + \hat{m}\boldsymbol{\theta}_0 \boldsymbol{\Sigma}_x \boldsymbol{\theta}^a + \hat{F}\boldsymbol{\theta}_0 \boldsymbol{\Sigma}_v \boldsymbol{\theta}^a\right) + \frac{1}{2}\sum_{1 \le a, b \le r} \hat{q}\boldsymbol{\theta}^{a\top} \boldsymbol{\Sigma}_x \boldsymbol{\theta}^b} \\
=& \mathbb{E}_{\boldsymbol{\theta}_0} \int_{\mathbb{R}^{d \times r}} \prod_{a=1}^{r} P_w\left(\mathrm{d}\boldsymbol{\theta}^a\right) \mathbb{E}_{\boldsymbol{\xi}}\left[e^{-\frac{\hat{V}}{2}\boldsymbol{\theta}^a \boldsymbol{\Sigma}_x \boldsymbol{\theta}^a - \frac{\hat{P}}{2}\boldsymbol{\theta}^a \boldsymbol{\Sigma}_\delta \boldsymbol{\theta}^a - \frac{\hat{A}}{2}\boldsymbol{\theta}^a \boldsymbol{\Sigma}_v \boldsymbol{\theta}^a - \frac{\hat{N}}{2}\boldsymbol{\theta}^a \boldsymbol{\theta}^a - \boldsymbol{\theta}^a\left(\hat{m}\boldsymbol{\Sigma}_x \boldsymbol{\theta}_0 + \hat{F}\boldsymbol{\Sigma}_v \boldsymbol{\theta}_0 - \sqrt{\hat{q}\boldsymbol{\Sigma}_x}\boldsymbol{\xi}\right)}\right] \\
=& \mathbb{E}_{\boldsymbol{\xi}, \boldsymbol{\theta}_0}\left[\left[\int_{\mathbb{R}^d} P_w\left(\mathrm{d}\boldsymbol{\theta}\right) e^{-\frac{\hat{V}}{2}\boldsymbol{\theta}\boldsymbol{\Sigma}_x \boldsymbol{\theta} - \frac{\hat{P}}{2}\boldsymbol{\theta}\boldsymbol{\Sigma}_\delta \boldsymbol{\theta} - \frac{\hat{A}}{2}\boldsymbol{\theta}\boldsymbol{\Sigma}_v \boldsymbol{\theta} - \frac{\hat{N}}{2}\boldsymbol{\theta}\boldsymbol{\theta} - \boldsymbol{\theta}\left(\hat{m}\boldsymbol{\Sigma}_x \boldsymbol{\theta}_0 + \hat{F}\boldsymbol{\Sigma}_v \boldsymbol{\theta}_0 - \sqrt{\hat{q}\boldsymbol{\Sigma}_x}\boldsymbol{\xi}\right)}\right]^r\right]
\end{aligned}
\tag{136}
$$

Then we can take the derivative and limit and obtain

$$
\Psi_w = \lim_{r \to 0^+} \partial_r \Psi_w^{(r)} = \frac{1}{d}\mathbb{E}_{\boldsymbol{\xi}, \boldsymbol{\theta}_0}\left[\log \int_{\mathbb{R}^d} P_w\left(\mathrm{d}\boldsymbol{\theta}\right) e^{-\frac{\hat{V}}{2}\boldsymbol{\theta}\boldsymbol{\Sigma}_x \boldsymbol{\theta} - \frac{\hat{P}}{2}\boldsymbol{\theta}\boldsymbol{\Sigma}_\delta \boldsymbol{\theta} - \frac{\hat{A}}{2}\boldsymbol{\theta}\boldsymbol{\Sigma}_v \boldsymbol{\theta} - \frac{\hat{N}}{2}\boldsymbol{\theta}\boldsymbol{\theta} - \boldsymbol{\theta}\left(\hat{m}\boldsymbol{\Sigma}_x \boldsymbol{\theta}_0 + \hat{F}\boldsymbol{\Sigma}_v \boldsymbol{\theta}_0 - \sqrt{\hat{q}\boldsymbol{\Sigma}_x}\boldsymbol{\xi}\right)}\right]
\tag{137}
$$

where we still need to take the limit $d \to \infty$.

I.2.2 Channel Replica Zero Limit

Now we can focus on the channel term and rewrite it in a more suitable way for taking the $r \to 0^+$ limit. In a very similar fashion as before we would like to simplify

$$
\Psi_y^{(r)} = \log\left[\int_{\mathbb{R}} \mathrm{d}y \int_{\mathbb{R}} \mathrm{d}\nu\, P_0(y \mid \nu) \int \prod_{a=1}^{r} \mathrm{d}\lambda^a\, P_g\left(y \mid \lambda^a, P, N, \varepsilon_t\right) \mathcal{N}\left(\nu, \lambda^a; \mathbf{0}, \Sigma^{ab}\right)\right]
\tag{138}
$$

We will indicate the argument of the log with $\mathcal{B}$. Additionally we have that the martix of covariances is

$$
\Sigma = \begin{pmatrix} \rho & m & m & \dots & m \\ m & Q & q & \dots & q \\ m & q & Q & \dots & q \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ m & q & q & \dots & Q \end{pmatrix}
\tag{139}
$$

and in addition also the inverse matrix has a Replica Symmetric Structure which is given from the following elements

$$
\begin{aligned}
&\left(\Sigma^{-1}\right)_{00} \equiv \tilde{\rho} = \frac{Q + (r-1)q}{\rho(Q + (r-1)q) - rm^2}, && \left(\Sigma^{-1}\right)_{0a} \equiv \tilde{m} = \frac{m}{rm^2 - \rho(Q + (r-1)q)}, \\
&\left(\Sigma^{-1}\right)_{aa} \equiv \tilde{Q} = \frac{\rho(Q + (r-2)q) - (r-1)m^2}{(Q-q)\left(\rho(Q + (r-1)q) - rm^2\right)}, && \left(\Sigma^{-1}\right)_{ab} \equiv \tilde{q} = \frac{m^2 - \rho q}{(Q-q)\left(\rho(Q + (r-1)q) - rm^2\right)}
\end{aligned}
\tag{140}
$$

and thus there is an implicit dependence on r in the covariance. To check that the inverse matrix has a RS structure as well one can think of the formula that is used to evaluate the inverse of a matrix from the cofactors.

Also we look at the determinant of the matrix. There are three different eigenvalue types

$$
\begin{aligned}
&\lambda_1 = Q - q, && \lambda_2 = \frac{1}{2}\left(-Q - q(r-1) - \rho - \tilde{\Delta}\right), && \lambda_3 = \frac{1}{2}\left(-Q - q(r-1) - \rho + \tilde{\Delta}\right), \\
&d_1 = r - 1, && d_2 = 1, && d_3 = 1,
\end{aligned}
\tag{141}
$$

with $\tilde{\Delta} = \sqrt{4m^2r + (Q + q(r-1) - \rho)^2}$ and thus one obtains the determinant. More explicitly we have that

$$
\begin{aligned}
\det\left(2\pi\Sigma\right) &= (2\pi)^{r+1}(Q-q)^{r-1}\frac{1}{4}(-Q - q(r-1) - \rho - \tilde{\Delta})(-Q - q(r-1) - \rho + \tilde{\Delta}) \\
&= (2\pi)^{r+1}(Q-q)^{r-1}(\rho(Q + (r-1)q) - rm^2)
\end{aligned}
\tag{142}
$$

Thus we have that

$$
\begin{aligned}
\mathcal{B} &= \int_{\mathbb{R}} \mathrm{d}y \int_{\mathbb{R}} \mathrm{d}\nu\, P_0(y \mid \nu) e^{-\frac{1}{2}\tilde{\rho}\nu^2} \int \prod_{a=1}^{r} \mathrm{d}\lambda^a\, P_g\left(y \mid \lambda^a, P, N, \varepsilon_t\right) \\
&\quad e^{-\tilde{m}\nu\sum_{a=1}^{r}\lambda^a - \frac{1}{2}\tilde{Q}\sum_{a=1}^{r}(\lambda^a)^2 - \frac{1}{2}\tilde{q}\sum_{1\le a,b\le r, a\ne b}\lambda^a\lambda^b - \frac{1}{2}\log\det(2\pi\Sigma)} \\
&= \mathbb{E}_{\xi} \int_{\mathbb{R}} \mathrm{d}y\, e^{-\frac{1}{2}\log\det(2\pi\Sigma)} \int_{\mathbb{R}} \mathrm{d}\nu\, P_0(y \mid \nu) e^{-\frac{1}{2}\tilde{\rho}\nu^2}\left[\int \mathrm{d}\lambda\, P_g\left(y \mid \lambda, P, N, \varepsilon_t\right) e^{-\frac{\tilde{Q}-\tilde{q}}{2}\lambda^2 + (\sqrt{-\tilde{q}}\xi - \tilde{m}\nu)\lambda}\right]^r
\end{aligned}
\tag{143}
$$

Now we can follow a similar procedure as before and define $V = Q - q$ we have that and the limit is

$$
\begin{aligned}
\Psi_y &= \lim_{r\to 0^+} \partial_r \Psi_y^{(r)} = \mathbb{E}_{\xi}\left[\int_{\mathbb{R}} \mathrm{d}y \int \frac{\mathrm{d}\nu}{\sqrt{2\pi\rho}} P_0(y \mid \nu) e^{-\frac{1}{2\rho}\nu^2} \log\left[\int \frac{\mathrm{d}\lambda}{\sqrt{2\pi}} P_g(y \mid \lambda, P, N, \varepsilon_t) e^{-\frac{1}{2}\frac{\lambda^2}{V} + \left(\frac{\sqrt{q-m^2/\rho}}{V}\xi + \frac{m/\rho}{V}\nu\right)\lambda}\right]\right] \\
&\quad - \frac{1}{2}\log V - \frac{1}{2}\frac{q}{V}
\end{aligned}
\tag{144}
$$

We would like to rewrite the quantities with the help of the following definition

$$
\mathcal{Z}_0(y, \omega, V) = \int \frac{\mathrm{d}x}{\sqrt{2\pi V}} e^{-\frac{1}{2V}(x-\omega)^2} P_0(y \mid x), \quad \mathcal{Z}_y(y, \omega, V, P, n) = \int \frac{\mathrm{d}x}{\sqrt{2\pi V}} e^{-\frac{1}{2V}(x-\omega)^2} P_g(y \mid x, P, N, \varepsilon_t),
\tag{145}
$$

The result becomes thus

$$
\mathbb{E}_{\xi}\left[\int_{\mathbb{R}} \mathrm{d}y\, \mathcal{Z}_0\left(y, \frac{m}{\sqrt{q}}\xi, \rho - \frac{m^2}{q}\right) \log \mathcal{Z}_y(y, \sqrt{q}\xi, V, P, N)\right]
\tag{146}
$$

Now there are two things that we still need to do : find the form for the prior term and take the limit $\beta \to \infty$.

I.3 Prior term for ℓ_2 regularisation

To be as general as possible we would like to include the case of a possible non isotropic regularisation. Thus

$$
P_w(\mathrm{d}\boldsymbol{\theta}) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{\beta\lambda}{2}\boldsymbol{\theta}\boldsymbol{\Sigma_w}\boldsymbol{\theta}\right)\mathrm{d}\boldsymbol{\theta}
\tag{147}
$$

We want to calculate the term inside the log in Eq. (137)

$$
\begin{aligned}
&\int_{\mathbb{R}^d} P_w(\mathrm{d}\boldsymbol{\theta}) e^{-\frac{\hat{V}}{2}\boldsymbol{\theta}\boldsymbol{\Sigma_x}\boldsymbol{\theta} - \frac{\hat{P}}{2}\boldsymbol{\theta}\boldsymbol{\Sigma_\delta}\boldsymbol{\theta} - \frac{\hat{A}}{2}\boldsymbol{\theta}\boldsymbol{\Sigma_v}\boldsymbol{\theta} - \frac{\hat{N}}{2}\boldsymbol{\theta}\boldsymbol{\theta} - \boldsymbol{\theta}\left(\hat{m}\boldsymbol{\Sigma_x}\boldsymbol{\theta}_0 + \hat{F}\boldsymbol{\Sigma_v}\boldsymbol{\theta}_0 - \sqrt{\hat{q}\boldsymbol{\Sigma_x}}\boldsymbol{\xi}\right)} \\
&= \frac{\exp\left(\frac{1}{2}\left(-\hat{m}\boldsymbol{\Sigma_x}^{\top}\boldsymbol{\theta}_0 - \hat{F}(\boldsymbol{\Sigma_v})^{\top}\boldsymbol{\theta}_0 + \sqrt{\hat{q}\boldsymbol{\Sigma_x}}\boldsymbol{\xi}\right)^{\top}\boldsymbol{\Lambda}^{-1}\left(-\hat{m}\boldsymbol{\Sigma_x}^{\top}\boldsymbol{\theta}_0 - \hat{F}(\boldsymbol{\Sigma_v})^{\top}\boldsymbol{\theta}_0 + \sqrt{\hat{q}\boldsymbol{\Sigma_x}}\boldsymbol{\xi}\right)^{\top}\right)}{\sqrt{\det\boldsymbol{\Lambda}}}
\end{aligned}
\tag{148}
$$

where we defined $\boldsymbol{\Lambda} = \beta\lambda\boldsymbol{\Sigma_w} + \hat{V}\boldsymbol{\Sigma_x} + \hat{P}\boldsymbol{\Sigma_\delta} + \hat{A}\boldsymbol{\Sigma_v} + \hat{N}\mathbb{1}$. Now the prior term becomes after taking the log and using the identity $\log\det = \operatorname{tr}\log$

$$
\begin{aligned}
\Psi_w &= \frac{1}{d}\mathbb{E}_{\boldsymbol{\xi},\boldsymbol{\theta}_0}\left[\frac{1}{2}\left(-\hat{m}\boldsymbol{\Sigma_x}^\top\boldsymbol{\theta}_0 - \hat{F}(\boldsymbol{\Sigma_v})^\top\boldsymbol{\theta}_0 + \sqrt{\hat{q}\boldsymbol{\Sigma_x}}\boldsymbol{\xi}\right)^\top\boldsymbol{\Lambda}^{-1}\left(-\hat{m}\boldsymbol{\Sigma_x}^\top\boldsymbol{\theta}_0 - \hat{F}(\boldsymbol{\Sigma_v})^\top\boldsymbol{\theta}_0 + \sqrt{\hat{q}\boldsymbol{\Sigma_x}}\boldsymbol{\xi}\right)^\top\right] - \frac{1}{2d}\operatorname{tr}\log\boldsymbol{\Lambda} \\
&= \frac{1}{2d}\operatorname{tr}\left[\left(\hat{m}^2\boldsymbol{\Sigma_x}^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\boldsymbol{\Sigma_x} + \hat{m}\hat{F}\boldsymbol{\Sigma_x}^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\boldsymbol{\Sigma_v} + \hat{m}\hat{F}(\boldsymbol{\Sigma_v})^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\boldsymbol{\Sigma_x} + \hat{F}^2(\boldsymbol{\Sigma_v})^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\boldsymbol{\Sigma_v} + \hat{q}\boldsymbol{\Sigma_x}\right)\boldsymbol{\Lambda}^{-1}\right] \\
&\quad - \frac{1}{2d}\operatorname{tr}\log\boldsymbol{\Lambda}
\end{aligned}
$$
(149)

The factor $\frac{1}{d}$ comes from the required scaling on d for the free entropy and the expectation from our replica zero limit of the prior term.

I.4 Zero temperature limit

We now need to take the zero temperature limit for this case. The explicit scalings of the parameters are

$$
\begin{array}{ccccccc}
V \to \beta^{-1}V & q \to q & m \to m & A \to A & N \to N & P \to P & F \to F \\
\hat{V} \to \beta\hat{V} & \hat{q} \to \beta^2\hat{q} & \hat{m} \to \beta\hat{m} & \hat{A} \to \beta\hat{A} & \hat{N} \to \beta\hat{N} & \hat{P} \to \beta\hat{P} & \hat{F} \to \beta\hat{F}
\end{array}
$$
(150)

The limit of the prior term is

$$
\begin{aligned}
\Psi_w &= \lim_{\beta\to\infty}\frac{1}{\beta}\Psi_w = \\
&\frac{1}{2d}\operatorname{tr}\left[\left(\hat{m}^2\boldsymbol{\Sigma_x}^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\boldsymbol{\Sigma_x} + \hat{m}\hat{F}\boldsymbol{\Sigma_x}^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\boldsymbol{\Sigma_v} + \hat{m}\hat{F}(\boldsymbol{\Sigma_v})^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\boldsymbol{\Sigma_x} + \hat{F}^2(\boldsymbol{\Sigma_v})^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\boldsymbol{\Sigma_v} + \hat{q}\boldsymbol{\Sigma_x}\right)\boldsymbol{\Lambda}^{-1}\right]
\end{aligned}
$$
(151)

To understand the limit of the channel term, we need to get the following insight for the limit of the channel partition function

$$
\begin{aligned}
\mathcal{Z}_y(y,\omega,V,P,N) &= \int\frac{\mathrm{d}x}{\sqrt{2\pi V}}e^{-\frac{\beta}{2V}(x-\omega)^2}P_g(y\mid x,P,N,\varepsilon_t) \\
&= \sqrt{\beta}\int\frac{\mathrm{d}x}{\sqrt{2\pi V}}e^{-\frac{\beta}{2V}(x-\omega)^2}\frac{1}{\sqrt{2\pi}}e^{\left(-\beta g(yx - \frac{\varepsilon_t}{\sqrt{d}}\frac{P}{\sqrt{N}})\right)} \\
&\underset{\beta\to\infty}{=} e^{-\beta\mathcal{M}_{Vg(y,\cdot)}(\omega)}
\end{aligned}
$$
(152)

where we introduced the Moreau envelope defined in Eq. (23). Notice that the previous computation is done for the case of the FGM attack and it doesn't change much in the case of the true minimisation.

Then the limit of the channel term becomes

$$
\Psi_y = \lim_{\beta\to\infty}\frac{1}{\beta}\Psi_y = -\mathbb{E}_\xi\left[\int\mathrm{d}y\,\mathcal{Z}_0\left(y,\frac{m}{\sqrt{q}}\xi,\rho - \frac{m^2}{q}\right)\mathcal{M}_{Vg(y,\cdot;P,N,\varepsilon_t)}(\sqrt{q}\xi)\right]
$$
(153)

where $\mathcal{M}_{Vg(y,\cdot;A,N,\varepsilon_t)}$ is the Moreau envelope of the modified loss function defined in Eq. (5) with the relevant quantities changed for their overlaps and $\xi \sim \mathcal{N}(0,1)$.

After taking the zero temperature limit, we are left with the following expression for the free energy density

$$
\begin{aligned}
\lim_{\beta\to\infty}f_\beta = \operatorname*{extr}_{\substack{V,q,m,A,N,P,F, \\ \hat{V},\hat{q},\hat{m},\hat{A},\hat{N},\hat{P},\hat{F}}}&\left\{-\frac{1}{2}(q\hat{V} - \hat{q}V) - \frac{1}{2}P\hat{P} - \frac{1}{2}A\hat{A} - \frac{1}{2}N\hat{N} + m\hat{m} + F\hat{F} + \alpha\mathbb{E}_\xi\left[\int\mathrm{d}y\,\mathcal{Z}_0\mathcal{M}_{Vg(y,\cdot)}\right]\right. \\
&\left. - \frac{1}{2d}\operatorname{tr}\left(\hat{m}^2\boldsymbol{\Sigma_x}^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\boldsymbol{\Sigma_x} + \hat{m}\hat{F}\boldsymbol{\Sigma_x}^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\boldsymbol{\Sigma_v} + \hat{m}\hat{F}(\boldsymbol{\Sigma_v})^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\boldsymbol{\Sigma_x} + \hat{F}^2(\boldsymbol{\Sigma_v})^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\boldsymbol{\Sigma_v} + \hat{q}\boldsymbol{\Sigma_x}\right)\boldsymbol{\Lambda}^{-1}\right\}
\end{aligned}
$$
(154)

I.5 Saddle-point equations

The extremisation condition of Eq. (154) can be translated into the overlap needing to satisfy the following

$$
\begin{aligned}
\hat{V} &= 2\alpha\partial_q\Psi_y, & q &= -2\partial_{\hat{V}}\Psi_w \\
\hat{q} &= -2\alpha\partial_V\Psi_y, & V &= 2\partial_{\hat{q}}\Psi_w, \\
\hat{N} &= 2\alpha\partial_N\Psi_y, & N &= -2\partial_{\hat{N}}\Psi_w \\
\hat{P} &= 2\alpha\partial_P\Psi_y, & P &= -2\partial_{\hat{P}}\Psi_w \\
\hat{A} &= 2\alpha\partial_A\Psi_y, & A &= -2\partial_{\hat{A}}\Psi_w \\
\hat{m} &= -\alpha\partial_m\Psi_y, & m &= \partial_{\hat{m}}\Psi_w \\
\hat{F} &= -\alpha\partial_F\Psi_y, & F &= \partial_{\hat{F}}\Psi_w.
\end{aligned}
\tag{155}
$$

As we pre-announced we would like to find the stationary values that dominate the integral and to do so we should derive the exponent with respect to all the order parameters. The saddle points that depend on $m, q, V, \hat{m}, \hat{q}$ and $\hat{V}$ are of a similar form as those found already in Loureiro et al. (2022a). We need thus to derive with respect to $A, N, P, F, \hat{A}, \hat{P}, \hat{F}$ and $\hat{N}$.

I.5.1 The Channel Saddle-Point Equations

Let us begin by looking at the derivatives with respect to P and N. These derivatives amount to computing the derivative of the Moreau-envelope with respect to P and N since we have that

$$
\begin{aligned}
\partial_P\Psi_y &= \mathbb{E}_{y,\xi}\left[\mathcal{Z}_0\left(y, \frac{m}{\sqrt{q}}\xi, \rho - \frac{m^2}{q}\right)\partial_P\mathcal{M}_{Vg(y,\cdot;P,N,\varepsilon_t)}(\sqrt{q}\xi)\right], \\
\partial_N\Psi_y &= \mathbb{E}_{y,\xi}\left[\mathcal{Z}_0\left(y, \frac{m}{\sqrt{q}}\xi, \rho - \frac{m^2}{q}\right)\partial_N\mathcal{M}_{Vg(y,\cdot;P,N,\varepsilon_t)}(\sqrt{q}\xi)\right],
\end{aligned}
\tag{156}
$$

Let's focus on the true minimisation. In this specific case we have that

$$
\mathcal{M}_{Vg(y,\cdot,P,N,\varepsilon_t)}(\omega) = \inf_{x\in\mathbb{R}}\left[\frac{(x-\omega)^2}{2V} + \ell\left(yx - \varepsilon_t\sqrt{P}\right)\right] = \mathcal{M}_{V\ell(y,\cdot)}\left(\omega - y\varepsilon_t\sqrt{P}\right)
\tag{157}
$$

where we remind that this specific form is possible since $y \in \{+1, -1\}$. With this we can relate these function the the derivative of the Moreau envelope with respect to its input as

$$
\partial_P\mathcal{M}_{Vg(y,\cdot,P,N,\varepsilon_t)}(\omega) = -y\frac{\varepsilon_t}{2\sqrt{P}}\mathcal{M}'_{V\ell(y,\cdot)}\left(\omega - y\varepsilon_t\sqrt{P}\right)
\tag{158}
$$

With this, we can write the new equation as

$$
\hat{P} = \alpha\frac{\varepsilon_t}{2\sqrt{P}}\mathbb{E}_\xi\left[\int_\mathbb{R}\mathrm{d}y\, y\mathcal{Z}_0 f_g\right]
\tag{159}
$$

where we have defined

$$
f_g(y, \omega, V, P, N, \varepsilon_t) = -\mathcal{M}'_{Vg(y,\cdot,P,N,\varepsilon_t)}(\omega)
\tag{160}
$$

In the case of the FGM we will also have an equation for N that can be derived similarly.

As the channel term does not depend on the overlaps F and A the hat equations are trivially zero and

$$
\hat{A} = 0, \quad \hat{F} = 0.
\tag{161}
$$

The remaining three equations can be found as in Loureiro et al. (2022a), where the only difference lies in the dependence of f_g on $\hat{P}$ and $\hat{N}$. Note that here and above we denote by z^* the value of the proximal at any given point of integration.

$$
\hat{V} = -\alpha\mathbb{E}_\xi\left[\int_\mathbb{R}\mathrm{d}y\,\mathcal{Z}_0\partial_\omega f_g\right], \quad \hat{q} = \alpha\mathbb{E}_\xi\left[\int_\mathbb{R}\mathrm{d}y\,\mathcal{Z}_0 f_g^2\right], \quad \hat{m} = \alpha\mathbb{E}_\xi\left[\int_\mathbb{R}\mathrm{d}y\,\partial_\omega\mathcal{Z}_0 f_g\right],
\tag{162}
$$

I.5.2 The Prior Saddle-Point Equations

For the prior saddle-point equations our starting point is

$$\Psi_w = \frac{1}{2d}\operatorname{tr}\left[\left(\hat{m}^2\boldsymbol{\Sigma}_{\boldsymbol{x}}^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\boldsymbol{\Sigma}_{\boldsymbol{x}} + \hat{m}\hat{F}\boldsymbol{\Sigma}_{\boldsymbol{x}}^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\boldsymbol{\Sigma}_{\boldsymbol{v}} + \hat{m}\hat{F}(\boldsymbol{\Sigma}_{\boldsymbol{v}})^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\boldsymbol{\Sigma}_{\boldsymbol{x}} + \hat{F}^2(\boldsymbol{\Sigma}_{\boldsymbol{v}})^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\boldsymbol{\Sigma}_{\boldsymbol{v}} + \hat{q}\boldsymbol{\Sigma}_{\boldsymbol{x}}\right)\boldsymbol{\Lambda}^{-1}\right] \tag{163}$$

where for simplicity of notation we define $\boldsymbol{\Lambda} = \beta\lambda\boldsymbol{\Sigma}_{\boldsymbol{w}} + \hat{V}\boldsymbol{\Sigma}_{\boldsymbol{x}} + \hat{P}\boldsymbol{\Sigma}_{\boldsymbol{\delta}} + \hat{A}\boldsymbol{\Sigma}_{\boldsymbol{v}} + \hat{N}\mathbb{1}$ and we will use $\boldsymbol{H} = \hat{m}^2\boldsymbol{\Sigma}_{\boldsymbol{x}}^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\boldsymbol{\Sigma}_{\boldsymbol{x}} + \hat{m}\hat{F}\boldsymbol{\Sigma}_{\boldsymbol{x}}^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\boldsymbol{\Sigma}_{\boldsymbol{v}} + \hat{m}\hat{F}(\boldsymbol{\Sigma}_{\boldsymbol{v}})^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\boldsymbol{\Sigma}_{\boldsymbol{x}} + \hat{F}^2(\boldsymbol{\Sigma}_{\boldsymbol{v}})^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\boldsymbol{\Sigma}_{\boldsymbol{v}} + \hat{q}\boldsymbol{\Sigma}_{\boldsymbol{x}}$.

As the channel equations for $\hat{A}$ and $\hat{F}$ are trivially zero, we want to start with these derivatives as the following expressions will simplify considerably.

$$\begin{aligned}
A = \partial_{\hat{A}}\Psi_w &= \frac{1}{d}\operatorname{tr}\left[\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{v}}\boldsymbol{\Lambda}^{-2}\right] \\
F = \partial_{\hat{F}}\Psi_w &= \frac{1}{d}\operatorname{tr}\left[\hat{m}\left(\boldsymbol{\Sigma}_{\boldsymbol{x}}^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\boldsymbol{\Sigma}_{\boldsymbol{v}}\right)\boldsymbol{\Lambda}^{-1}\right]
\end{aligned} \tag{164}$$

We want to compute a few derivatives of the term Ψ_w to obtain equations for the overlap P.

We begin with the hat-variable

$$P = \partial_{\hat{P}}\Psi_w = \frac{1}{d}\operatorname{tr}\left[\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\boldsymbol{\Lambda}^{-2}\right] \tag{165}$$

Again in the case of the FGM we will have an additional equation that is very similar to the one of P.

As before the derivative w.r.t. $\hat{q}, \hat{a}, \hat{n}$ follow from previous literature as

$$V = \partial_{\hat{q}}\Psi_w = \frac{1}{d}\operatorname{tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\Lambda}^{-1}\right], \quad q = \partial_{\hat{V}}\Psi_w = \frac{1}{d}\operatorname{tr}\left[\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\Lambda}^{-2}\right], \quad m = \partial_{\hat{m}}\Psi_w = \frac{1}{d}\operatorname{tr}\left[\hat{m}\boldsymbol{\Sigma}_{\boldsymbol{x}}^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\Lambda}^{-1}\right]. \tag{166}$$

Note that for the numerical evaluation $\frac{1}{d}\operatorname{tr}$ is just the mean of the eigenspectrum.

I.6 Final set of saddle point equations for ℓ_2 regularisation

We state here our final set of saddle point equations for reference

$$\begin{cases}
\hat{m} = \alpha\mathbb{E}_\xi\left[\int_\mathbb{R} dy\, \partial_\omega\mathcal{Z}_0 f_g(\sqrt{q}\xi, P, N, \varepsilon_t)\right] \\
\hat{q} = \alpha\mathbb{E}_\xi\left[\int_\mathbb{R} dy\, \mathcal{Z}_0 f_g^2(\sqrt{q}\xi, P, N, \varepsilon_t)\right] \\
\hat{V} = -\alpha\mathbb{E}_\xi\left[\int_\mathbb{R} dy\, \mathcal{Z}_0\partial_\omega f_g(\sqrt{q}\xi, P, N, \varepsilon_t)\right] \\
\hat{P} = \varepsilon_t\frac{1}{2\sqrt{P}}\alpha\mathbb{E}_\xi\left[\int_\mathbb{R} dy\, \mathcal{Z}_0 y f_g(\sqrt{q}\xi, P, N, \varepsilon_t)\right]
\end{cases}$$

$$\begin{cases}
m = \frac{1}{d}\operatorname{tr}\left[\hat{m}\boldsymbol{\Sigma}_{\boldsymbol{x}}^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\Lambda}^{-1}\right] \\
q = \frac{1}{d}\operatorname{tr}\left[\left(\hat{m}^2\boldsymbol{\Sigma}_{\boldsymbol{x}}^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\boldsymbol{\Sigma}_{\boldsymbol{x}} + \hat{q}\boldsymbol{\Sigma}_{\boldsymbol{x}}\right)\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\Lambda}^{-2}\right] \\
V = \frac{1}{d}\operatorname{tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\Lambda}^{-1}\right] \\
P = \frac{1}{d}\operatorname{tr}\left[\left(\hat{m}^2\boldsymbol{\Sigma}_{\boldsymbol{x}}^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\boldsymbol{\Sigma}_{\boldsymbol{x}} + \hat{q}\boldsymbol{\Sigma}_{\boldsymbol{x}}\right)\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\boldsymbol{\Lambda}^{-2}\right]
\end{cases} \tag{167}$$

where we remember the definitions of $\boldsymbol{\Lambda} = \lambda\boldsymbol{\Sigma}_{\boldsymbol{w}} + \hat{V}\boldsymbol{\Sigma}_{\boldsymbol{x}} + \hat{P}\boldsymbol{\Sigma}_{\boldsymbol{\delta}}$. Note $\hat{A}, \hat{F}$ are exactly zero because the channel is not dependent on A, F. The values for A and F becomes

$$A = \frac{1}{d}\operatorname{tr}\left[\left(\hat{m}^2\boldsymbol{\Sigma}_{\boldsymbol{x}}^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\boldsymbol{\Sigma}_{\boldsymbol{x}} + \hat{q}\boldsymbol{\Sigma}_{\boldsymbol{x}}\right)\boldsymbol{\Sigma}_{\boldsymbol{v}}\boldsymbol{\Lambda}^{-2}\right], \quad F = \frac{1}{d}\operatorname{tr}\left[\hat{m}\left(\boldsymbol{\Sigma}_{\boldsymbol{x}}^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\boldsymbol{\Sigma}_{\boldsymbol{v}}\right)\boldsymbol{\Lambda}^{-1}\right]. \tag{168}$$

where we remind that f_g is defined in Eq. (23).