

Solvable Model for Inheriting the Regularization through Knowledge Distillation

Luca Saglietti

SPOC laboratory, EPFL, Switzerland.

LUCA.SAGLIETTI@EPFL.CH

Lenka Zdeborová

SPOC laboratory, EPFL, Switzerland.

LENKA.ZDEBOROVA@EPFL.CH

Editors: Joan Bruna, Jan S Hesthaven, Lenka Zdeborova

Abstract

In recent years the empirical success of transfer learning with neural networks has stimulated an increasing interest in obtaining a theoretical understanding of its core properties. Knowledge Distillation where a smaller neural network is trained using the outputs of a larger neural network is a particularly interesting case of transfer learning. In the present work, we introduce a statistical physics framework that allows an analytic characterization of the properties of knowledge distillation (KD) in shallow neural networks. Focusing the analysis on a solvable model that exhibits a non-trivial generalization gap, we investigate the effectiveness of KD. We are able to show that, through KD, the regularization properties of the larger teacher model can be inherited by the smaller student and that the yielded generalization performance is closely linked to and limited by the optimality of the teacher. Finally, we analyze the double descent phenomenology that can arise in the considered KD setting.

Keywords: Knowledge Distillation, Transfer Learning, Logistic Regression, Gaussian Mixture, Replica Method

1. Introduction

Deep learning practice in the past decade has repeatedly confirmed a remarkable observation: Stochastic Gradient Descent (SGD) based training of neural networks becomes easier and more effective as the number of tunable parameters increases. While a higher model complexity could in principle entail high risks of over-fitting, large scale Deep Neural Networks (DNNs) display surprising generalization capabilities, allegedly allowed by an “implicit regularization” mechanism (Neyshabur et al. (2015); Zhang et al. (2017)) that still escapes clear theoretical understanding. On the flip side, the steady scaling up of DNN architectures also carries a massive increase in associated inference and memory costs.

Many attempts at achieving better quality-computation trade-offs have been proposed in the last decade (Han et al. (2015, 2016); Jacob et al. (2018); Frankle and Carbin (2019)), often based on the observation that good generalization scores can be retained if one first trains a more complex DNN and then derives a lighter model from it. In this context, Knowledge Distillation (KD) (Hinton et al. (2015)) has established itself as one of the most popular transfer learning and network compression strategies (Anil et al. (2018); Chen et al. (2018, 2017); Yim et al. (2017); Yu et al. (2017); Kim and Rush (2016)). The general idea of KD is to try and transfer the generalization properties from a larger capacity model (teacher) to a weaker model (student) at training: instead of

learning directly from the vector encodings of the ground truth labels, the KD student learns from the outputs (“dark knowledge”) produced by the teacher model on the same training dataset. Not only the KD optimization step, with real-valued outputs, seems to be generally more well-behaved than usual training, numerical experiments also show that with very little fine-tuning at the level of the employed regularization and optimization heuristics one can reach competitive generalization scores (Tang et al. (2020)).

Despite its effectiveness, KD is still not very well understood from a theoretical standpoint. Few recent works (Celik et al. (2017); Phuong and Lampert (2019); Tang et al. (2020); Rahbar et al. (2020); Yuan et al. (2019); Furlanello et al. (2018)) have attempted to analyse KD in controlled settings, breaking down its net positive impact into separate contributions and proposing a connection between the effectiveness of KD and the role of label smoothing and sample reweighting strategies and that of priors on data geometry (Yuan et al. (2019); Furlanello et al. (2018)). In this work we approach the problem from a statistical physics perspective, aiming to characterize the typical generalization performance achieved by a KD student in the asymptotic limit of large input dimension and dataset size. The main questions we want to investigate are how KD can transfer the regularization properties between mismatched models and when the KD student can display an improvement upon the best score achievable with usual training procedures.

In order to allow a mathematical definition of knowledge distillation, consider a typical classification problem where, given a large dataset of input-output associations $\mathcal{D} = \{\mathbf{x}^\mu, y^\mu\}_{\mu=1}^M$ the task is to learn a parametrized rule $f(\mathbf{x}^\mu, \{\mathbf{w}\})$ (f representing a neural network and $\{\mathbf{w}\}$ its parameters or weights) that allows correct classification of test data points not seen in the training set. As customary, learning can be framed as an empirical risk minimization problem, introducing a regularized loss-function:

$$\mathcal{L}(\{\mathbf{w}\}, \mathcal{D}; \lambda) = \sum_{\mu=1}^M \ell(y^\mu, \sigma(f(\mathbf{x}^\mu, \{\mathbf{w}\}))) + \frac{\lambda}{2} \|\mathbf{w}\|_2^2 \quad (1)$$

where, in the binary classification case, $\sigma(\cdot)$ is the sigmoid activation function $\sigma(x) = (1 + \exp(-x))^{-1}$ (*softmax* in the multi-class case), and where a typical choice would be a cross-entropy loss $\ell(p, q) = \mathcal{H}(p, q)$:

$$\mathcal{H}(p, q) = -p \log q - (1 - p) \log(1 - q) \quad (2)$$

and an L_2 -norm regularization.

In knowledge distillation (KD), one assumes to be granted access to the outputs $\tilde{f}(\mathbf{x}^\mu, \{\tilde{\mathbf{w}}\})$ produced over the training set by a (more complex) teacher model $\{\tilde{\mathbf{w}}\}$, and one aims at training a (weaker) student model through the modified loss:

$$\mathcal{L}_{KD}(\{\mathbf{w}\}, \{\tilde{\mathbf{w}}\}, \mathcal{D}; \lambda, \chi) = \sum_{\mu=1}^N \ell_{KD}(y^\mu, \tilde{f}(\mathbf{x}^\mu, \{\tilde{\mathbf{w}}\}), f(\mathbf{x}^\mu, \{\mathbf{w}\}), \chi) + \frac{\lambda}{2} \|\mathbf{w}\|_2^2 \quad (3)$$

where:

$$\ell_{KD}(y, p, q, \chi) = (1 - \chi) \mathcal{H}(y, \sigma(q)) + \chi \mathcal{H}(\sigma(p), \sigma(q)). \quad (4)$$

The student is thus mixing the usual data-fitting approach with a goal of approximating the behavior of the teacher, the external parameter χ serving to balance between fitting the ground truth labels and the teacher outputs. A key idea behind distillation is that for the student the optimization

process becomes more transparent, as it can rely on the more explicit knowledge derived from the teacher outputs: the softer outputs can prevent student overconfidence on noisy data points and highlight correlations among different labels. Note that, in multi-class problems one typically considers also a distillation temperature T , reweighing teacher and student outputs: we will ignore this additional external parameter in the following, since we focus on a binary classification setting (a short analysis of its effect can be found in Appendix B).

1.1. Main contributions

In this manuscript, we develop and apply an analytic framework to study knowledge distillation in models where the learning performance is solvable with the replica method. We then consider specifically a Gaussian mixture model where the student is constrained to be sparse, and we perform a series of controlled studies that allowed an investigation of the inheritance properties of knowledge distillation. All analytical results are crosschecked with numerical experiments. Our main results can be summarized in the following qualitative observations:

- Without any fine-tuning at the level of the student loss function, using KD allows a transfer of the (possibly fine-tuned) regularization properties of the teacher, even if the two models are mismatched and even if the regularization strategy in the teacher training is not known explicitly.
- When the regularization mechanism employed for regularizing the teacher can also be applied directly to the student, fine-tuning the direct regularization and fine-tuning the parameters of the KD loss leads to comparable generalization performance. No improvement is observed in this setting.
- If one can access a trained network with superior generalization performance and employ it as a teacher in a KD process, also the KD student will inherit superior generalization properties.
- In the limit of zero direct regularization on the student, the KD loss gives rise to a hybrid double-descent phenomenology, displaying both logistic regression and linear regression types of cusps.

Of course, the results we derived in the simple Gaussian mixture model may not generalize directly to more complex network architectures or different types of model mismatches. We argue, however, that the observed qualitative behavior is in line with the empirical observations about KD practice in deep learning (Hinton et al. (2015); Rahbar et al. (2020)), and support the propositions that transferring knowledge from larger (implicitly regularized) neural network models is almost automatically beneficial for the test performance of weaker students. Moreover, the development of a general theoretical framework for this type of study could stimulate a similar analysis in more realistic settings.

In the next section, we introduce our analytical framework, yielding an asymptotic description of training through knowledge distillation. In section 3, we present a solvable model where the test performance associated with logistic regression is largely sub-optimal and we define in which sense the considered student network is a smaller model than the teacher. In section 4, we apply the analytical framework to the model and we derive a set of deterministic fixed-point equations that allow an estimate of the KD student generalization performance. In section 5, we showcase

the main results of this work, comparing our predictions with numerical simulations. In particular, we characterize the inheritance properties and the limits of KD and show when KD can potentially lead to improved generalization with respect to typical logistic regression. In section 6, we focus on the double-descent phenomenology that appears in our simple transfer learning setup. Finally, in section 7 we discuss our results and propose some future research perspectives.

2. Statistical physics framework to analyze knowledge distillation

The main technical contribution of this work consists in the introduction of an analytic framework based on the replica formalism (Mézard et al. (1987); Mezard and Montanari (2009)), that allows the characterization of learning through knowledge distillation in tractable models.

The proposed analytic setup stems from the simple observation that KD can naturally be framed as a 2-level problem: in the first step one trains a teacher model with the true labels and the dataset \mathcal{D} ; in the second step a student network is trained with the same inputs and the outputs produced by the teacher. From the perspective of the replica method, the fact that the two systems are sharing the same inputs (same quenched disorder) is effectively coupling them. However, because of the fixed order of the two training procedures, the teacher model is not affected by the presence of the student and therefore its statistical properties can be determined self-consistently. On the other hand, the statistical measure of the student is directly dependent on the specific realization of the teacher.

In particular, we can characterize the typical KD student by considering the disordered partition function:

$$Z(\{\tilde{\mathbf{w}}\}, \mathcal{D}) = \lim_{\beta \rightarrow \infty} \int d\mathbf{w} e^{-\beta \mathcal{L}_{KD}(\{\mathbf{w}\}, \{\tilde{\mathbf{w}}\}, \mathcal{D})} \quad (5)$$

in the limit $\beta \rightarrow \infty$, where the measure focuses on the minimizers of the KD loss functions, and then evaluate the free-entropy Φ of the model by performing an external average over the realization of the dataset and an internal average over the measure of the trained teacher:

$$\Phi = \frac{1}{N} \left\langle \left\langle \log Z(\{\tilde{\mathbf{w}}\}, \mathcal{D}) \right\rangle_{\{\tilde{\mathbf{w}}\}} \right\rangle_{\mathcal{D}}. \quad (6)$$

Quantities of interest such as the value of test error of the student are then readily derived from this free entropy in ways standard to statistical physics (Mezard and Montanari (2009)). Obtaining a close formula for the free entropy Φ is hence the key difficulty that can be overcome using the replica trick (Mézard et al. (1987)). The replica formalism required for evaluating the double average in Eq. (6) is equivalent to a Franz-Parisi potential computation (Franz and Parisi (1997)), where one first samples a configuration from an independent equilibrium measure and then evaluates the free-energy of a coupled system sharing the same realization of the disorder.

The computation starts from a chain of identities based on two separate replica tricks:

$$\Phi = \frac{1}{N} \left\langle \left\langle \log \lim_{\beta \rightarrow \infty} \int d\mathbf{w} e^{-\beta \mathcal{L}_{KD}(\{\mathbf{w}\}, \{\tilde{\mathbf{w}}\}, \mathcal{D})} \right\rangle_{\{\tilde{\mathbf{w}}\}} \right\rangle_{\mathcal{D}} = \quad (7)$$

$$\frac{1}{N} \left\langle \left\langle \lim_{n \rightarrow 0} \frac{\partial}{\partial n} \lim_{\beta \rightarrow \infty} \int \prod_{a=1}^n d\mathbf{w}^a e^{-\beta \mathcal{L}_{KD}(\{\mathbf{w}^a\}, \{\tilde{\mathbf{w}}\}, \mathcal{D})} \right\rangle_{\{\tilde{\mathbf{w}}\}} \right\rangle_{\mathcal{D}} = \quad (8)$$

$$\frac{1}{N} \left\langle \lim_{\tilde{n}, n \rightarrow 0} \frac{\partial}{\partial n} \lim_{\tilde{\beta}, \beta \rightarrow \infty} \int \prod_{c=1}^{\tilde{n}} d\tilde{\mathbf{w}}^c e^{-\tilde{\beta} \mathcal{L}(\{\tilde{\mathbf{w}}^c\}, \mathcal{D})} \int \prod_{a=1}^n d\mathbf{w}^a e^{-\beta \mathcal{L}_{KD}(\{\mathbf{w}^a\}, \{\tilde{\mathbf{w}}^1\}, \mathcal{D})} \right\rangle_{\mathcal{D}}. \quad (9)$$

In order to evaluate the disorder average, in the second line the logarithm is removed by replicating the student configuration $\{\mathbf{w}^a\}_{a=1}^n$ (using the identity $\log x = \lim_{n \rightarrow 0} \partial_n x^n$). In the third line, instead, the average over the teacher is removed by introducing $\tilde{n} - 1$ non-interacting and a single interacting replica of the teacher $\{\tilde{\mathbf{w}}^c\}_{c=1}^{\tilde{n}}$, so that in the $\tilde{n} \rightarrow 0$ limit one can recover the expectation over its measure.

Because of concentration properties in high-dimensions, the coupled free-entropy asymptotically converges to a deterministic function of a narrow set of order parameters that capture the geometrical distribution of teacher and student configurations. Enforcing a saddle-point condition for the free-energy allows the derivation of a system of fixed-point equations that can yield an asymptotic prediction for these order parameters, to be compared with the results of numerical simulations.

The proposed formalism is general and may be applied to analyze knowledge distillation in any learning model which is amenable of a description through the replica method. Note that the entailed computation is quite standard in statistical physics and is believed to be exact, although non-rigorous in general. Moreover, an important remark is that there currently are strong technical limitations which restrict the set of models tractable with the replica method to a class of shallow network architectures (Barbier et al. (2019); Aubin et al. (2018)), but these limitations might be lifted with future progress in the field.

3. Gaussian Mixture Model

We now provide a brief introduction to a model recently analyzed in Mignacco et al. (2020) with the replica method, which will be employed as a prototypical study case in the rest of the paper. The same models can be studied with other tools, some of them rigorous, but here we focus on the replica solution of the model because that is the one that can readily be extended to analyze the knowledge distillation along the lines of section 2. We consider a high-dimensional binary classification problem where data is generated according to a Gaussian mixture and the learning model is a linear classifier trained through L_2 -regularized logistic regression.

In particular, let N denote the input dimension and M denote the size of the training set \mathcal{D} . We assume the data points in \mathcal{D} to be Gaussian distributed around two centroids, located on the same axis and positioned respectively at $\pm \frac{\mathbf{v}}{\sqrt{N}}$, $\mathbf{v} \in \mathbb{R}^N$. Moreover, we assume the two clusters to contain respectively a fraction ρ and $(1 - \rho)$ of the points. Thus, data $\{\mathbf{x}^\mu, y^\mu\}_{\mu=1}^M$ is generated according to the process:

$$\mathbf{x}^\mu = (2y^\mu - 1) \frac{\mathbf{v}}{\sqrt{N}} + \sqrt{\Delta} \mathbf{z} \quad (10)$$

where each component of the signal \mathbf{v} and noise \mathbf{z} is i.i.d. Gaussian, $v_i, z_i \sim \mathcal{N}(0, 1)$. The binary labels $y_\mu \in \{0, 1\}$ that determine the cluster membership of the points follow the skewed distribution $y^\mu \sim \rho \delta(y^\mu - 1) + (1 - \rho) \delta(y^\mu)$. We also specialize to the case of a single layer network, with:

$$f(\mathbf{x}^\mu, \{\mathbf{w}\}) = \frac{\mathbf{x}^\mu \cdot \mathbf{w}}{\sqrt{N}} + b \quad (11)$$

where the weights $\mathbf{w} \in \mathbb{R}^N$ and the bias $b \in \mathbb{R}$ represent the tunable parameters of the model. Note that, as soon as the training set is no longer linearly separable, the optimal learning strategy is to try and align the weights in the direction of the signal \mathbf{v} , so that the probability of a correct labeling of the data points is maximized.

Non-trivial behaviour was described in this model in the scaling limit where both $N, M \rightarrow \infty$, while their ratio $\alpha = M/N$ remains of $\mathcal{O}(1)$ (Mignacco et al. (2020)). Asymptotically, the model is fully solvable and one can characterize the typical learning performance as a function of the model parameters and of the regularization intensity λ . For the reader’s convenience we report some results in Appendix A, but we reference Mignacco et al. (2020) for the full details on the properties of this model.

3.1. Sub-optimal performance of logistic regression

One of the main motivations for considering this simple model in the present knowledge distillation study comes from the sub-optimal generalization behavior of regularized logistic regression in the unbalanced cluster case, at $\rho < 0.5$. As reported in Mignacco et al. (2020), given the overlap between the weight configuration and the signal, $m = \frac{w \cdot v}{N}$ and the norm $q = \frac{w \cdot w}{N}$, one obtains the asymptotic generalization error of the trained configuration from the following analytic expression:

$$\epsilon_g = \rho H\left(\frac{m+b}{\sqrt{\Delta q}}\right) + (1-\rho)H\left(\frac{m-b}{\sqrt{\Delta q}}\right), \quad (12)$$

where $H(x) = \int_x^\infty \frac{dt}{\sqrt{2\pi}} \exp(-t^2/2)$ is the Gaussian tail function. This score can then be compared with the Bayes optimal generalization error (computed by matching the inference model with the generative one), which represents a lower bound for the performance of any learning algorithm.

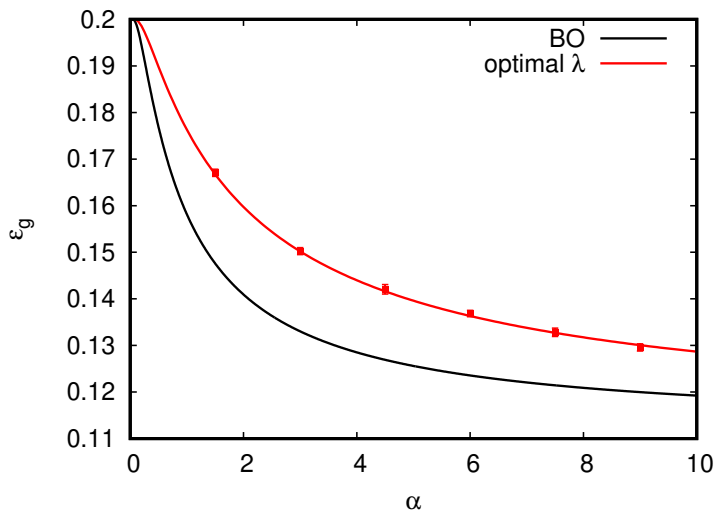


Figure 1: Generalization performance of L_2 -regularized logistic regression compared to the Bayes optimal lower bound, in a Gaussian Mixture with $\rho = 0.2$ and $\Delta = 1$. Red line: regularized logistic regression with optimal intensity λ . Black line: Bayes optimal performance. The data points with error bars represent the results of numerical experiments at $N = 4000$ (10 samples per point).

Remarkably, in this specific model there always exist at least a point estimator that achieves such Bayes-optimal performance, constructed according to a Hebbian principle (Hebb (2005)):

$$\mathbf{w}_{BO} = \frac{1}{\alpha\sqrt{N}} \sum_{\mu=1}^{\alpha N} (2y_{\mu} - 1)\mathbf{x}_{\mu}, \quad b_{BO} = \frac{\Delta\|\mathbf{w}_{BO}\|_2^2}{2} \log \frac{\rho}{1-\rho}. \quad (13)$$

Thus, the question is whether one can set an optimal value of the regularization such that logistic regression can perform similarly. The somewhat surprising answer is that, at $\rho < 0.5$ and any value of α and Δ , one observes a sizable gap between the best generalization performance obtained through logistic regression and the optimal one. This phenomenon can be clearly seen in Fig. 1. In section 5 we will investigate whether KD can help the student close this performance gap. Note that this sub-optimal behavior does not appear with balanced clusters $\rho = 0.5$, where the optimal regularization level is obtained in the limit $\lambda \rightarrow \infty$ (Mignacco et al. (2020)).

3.2. Teacher-student mismatch

In the present work, we want to analyze a setting where teacher and student model classes are mismatched, so that the weaker student is not able to exactly replicate the behavior of the teacher. Introducing some type of mismatch is not only closer to KD practice, but also crucial for inducing a richer model phenomenology. As a matter of fact, it was shown in Phuong and Lampert (2019) that as soon as the density of patterns becomes larger than $\alpha > 1$ a linear KD student can trivially recover the teacher weight configuration.

We thus consider a scenario where the student model can only train a fraction $0 < \eta < 1$ of its weights while the rest is set to 0 *ab initio*. In this way it will be impossible for the student to exactly infer (and achieve the same performance as) the teacher, and we can focus on the transfer of knowledge between the two models. Of course, this type of mismatch is much simpler than the architectural mismatches typically entailed in KD learning procedures, but we will see it is sufficient to display non-trivial phenomenology.

Note that, because of the simple nature of the considered generative model, setting a fraction $\eta < 1$ is equivalent to rescaling the effective signal-to-noise ratio in the student learning problem. In fact, the two inference tasks with $\{\eta, \alpha, \Delta\}$ and with $\{1, \alpha/\eta, \Delta/\eta\}$ are information-theoretically equivalent in the asymptotic limit. Moreover, one can easily define a Bayes optimal lower bound also in the sparse sub-space spanned by the student, achieved by a point-estimator with the same bias b_{BO} as in Eq. (13) and trimmed weights:

$$(w_{BO})_i = \begin{cases} \frac{1}{\alpha\sqrt{N}} \sum (2y^{\mu} - 1) x_i^{\mu}, & \text{for } i \leq \eta N \\ 0, & \text{for } i > \eta N \end{cases}. \quad (14)$$

As expected, one can easily prove that the associated performance coincides with the typical Bayesian generalization obtained after rescaling α and Δ by a factor $1/\eta$.

4. Knowledge Distillation in the Gaussian Mixture Model

We apply the replica framework sketched in Sec. 2 to derive a set of deterministic equations characterizing typical knowledge distillation processes in the above introduced logistic regression setting, where one first trains a teacher linear classifier and then employs the KD loss Eq. (4) to train a

sparsified linear student. The convex nature of the two nested optimization problems justifies the employment of the so-called Replica Symmetric ansatz (Mignacco et al. (2020)), which simplifies the analysis considerably. As the replica computation is still quite involved, we defer a detailed description to the Appendix, and report here the obtained final expressions.

We remind that the main parameters of the setting we analyze are the noise variance Δ and the label fraction ρ , the number of samples per dimension $\alpha = M/N$ and the student-sparsity level η . Specializing Eq. (9) to our study case we get:

$$\begin{aligned} \Phi = & \frac{1}{N} \lim_{n, \tilde{n} \rightarrow 0} \partial_n \left\langle \lim_{\tilde{\beta} \rightarrow \infty} \lim_{\beta \rightarrow \infty} \int \prod_{c=1}^{\tilde{n}} d\tilde{w}^c e^{-\frac{\beta \tilde{\lambda}}{2} \|\tilde{w}^c\|_2^2} \int \prod_{c=1}^{\tilde{n}} d\tilde{b}^c \prod_{\mu, c} e^{-\frac{\tilde{\beta}}{2} \ell \left(y^\mu, \sigma \left(\sum_{i=1}^N \frac{\tilde{w}_i^c x_i^\mu}{\sqrt{N}} + \tilde{b}^c \right) \right)} \right. \\ & \times \left. \int \prod_{a=1}^n d\mathbf{w}^a e^{-\frac{\beta \lambda}{2} \|\mathbf{w}^a\|_2^2} \int \prod_{a=1}^n db^a \prod_{\mu, a} e^{-\frac{\beta}{2} \ell' \left(y^\mu, \sigma \left(\sum_{i=1}^N \frac{\tilde{w}_i^1 x_i^\mu}{\sqrt{N}} + \tilde{b}^1 \right), \sigma \left(\sum_{i=1}^N \frac{w_i^a x_i^\mu}{\sqrt{N}} + b^a \right), \chi \right)} \right\rangle_{\{\mathbf{x}^\mu, y^\mu\}}, \end{aligned} \quad (15)$$

where b, \tilde{b} are the biases, and $\lambda, \tilde{\lambda}$ the strength of the regularization of the student and teacher.

In order to perform disorder average and remove the dependency on the specific realization of the Gaussian mixture dataset, one can first isolate teacher and student preactivations by introducing the associated Dirac's δ -functions. Then, it becomes possible to factorize over the samples and the input components and take the expectation over x_i^μ . Finally, one can discard the $o(N^{-1})$ terms and obtain the effective interaction between the various replicas, mediated by a set of overlap order parameters.

Thus, in the replica symmetric assumption, the only relevant quantities that fully characterize the studied model are:

- The overlap between teacher \tilde{w} and student w weight configurations with the signal v denoted $\tilde{m} = \frac{\tilde{w} \cdot v}{N}$, $m = \frac{w \cdot v}{N}$.
- The norms $\tilde{q} = \frac{\tilde{w} \cdot \tilde{w}}{N}$ and $q = \frac{w \cdot w}{N}$.
- The teacher-student overlap $S = \frac{w \cdot \tilde{w}}{N}$.
- The vanishing variances $\delta \tilde{q}$, δq and δS (see Appendix A for a detailed definition), opportunely rescaled in the $\tilde{\beta}, \beta \rightarrow \infty$ limit.

and their associated conjugate variables, denoted with a hat symbol in the following. Then, after some calculations, one can express the free-entropy as an extremum operation:

$$\Phi = \text{extr}_{b, m, q, \delta q, S, \delta S, \hat{m}, \hat{q}, \delta \hat{q}, \hat{S}, \delta \hat{S}} - \left(\hat{m} m + \frac{1}{2} (\hat{q} \delta q - \delta \hat{q} q) + (\hat{S} \delta S + \delta \hat{S} S) \right) + \eta g_s + \alpha g_e, \quad (16)$$

where we introduced the entropic and energetic contributions:

$$g_s = \frac{1}{2} \frac{\left(\hat{m} + \hat{m} \frac{\delta \hat{S}}{\lambda + \delta \hat{q}} \right)^2 + \hat{q} + 2 \frac{\delta \hat{S} \hat{S}}{\lambda + \delta \hat{q}} + \frac{\hat{q} \delta \hat{S}^2}{(\lambda + \delta \hat{q})^2}}{\lambda + \delta \hat{q}} \quad (17)$$

$$g_e = \left\langle \int \mathcal{D}z \int \mathcal{D}\tilde{z} M_E^* \right\rangle_y \quad (18)$$

with $\mathcal{D}z$ and $\mathcal{D}\tilde{z}$ denoting independent normalized Gaussian measures and the average $\langle \cdot \rangle_y$ taken over the distribution of the cluster labels. The argument of Eq. (18) is obtained from a one-dimensional optimization problem:

$$M_E^* = \max_u \left\{ -\frac{1}{2}u^2 - \frac{1}{2}\ell' \left(y, \sigma \left(\tilde{h}(\tilde{u}^*) \right), \sigma \left(h(u, \tilde{u}^*) \right), \chi \right) \right\} \quad (19)$$

where \tilde{h} and h represent the teacher's and student's output pre-activations, given respectively by:

$$\tilde{h}(\tilde{u}) = \sqrt{\Delta\delta\tilde{q}}\tilde{u} + \sqrt{\Delta\tilde{q}}\tilde{z} + (2y-1)\tilde{m} + \tilde{b} \quad (20)$$

$$h(u, \tilde{u}) = \sqrt{\Delta\delta q}u + \sqrt{\Delta\left(q - \frac{S^2}{\tilde{q}}\right)}z + \sqrt{\Delta}\frac{\delta S}{\sqrt{\delta\tilde{q}}}\tilde{u} + \sqrt{\Delta}\frac{S}{\sqrt{\tilde{q}}}\tilde{z} + (2y-1)m + b \quad (21)$$

$$\tilde{u}^* = \operatorname{argmax}_{\tilde{u}} \left\{ -\frac{1}{2}\tilde{u}^2 - \frac{1}{2}\ell \left[y, \sigma \left(\tilde{h}(\tilde{u}) \right) \right] \right\}. \quad (22)$$

Note that the 2-level structure of Knowledge Distillation clearly appears in the concatenated optimization entailed in Eqs. (19) and (22).

Since the teacher measure does not depend on the student, the value of the associated order parameters can be determined independently by optimizing a simpler free-entropy. The corresponding fixed-point equations are reported in Appendix A (and are equivalent to those presented in Mignacco et al. (2020)).

Once the saddle-point values for the order parameters of the models are evaluated for a given set of parameters $\{\alpha, \Delta, \rho, \eta, \tilde{\lambda}, \lambda, \chi\}$, the corresponding generalization can be obtained via Eq. (12).

In the present work we do not seek a rigorous proof of the replica predictions (e.g., following a similar Gordon Minimax approach as in Mignacco et al. (2020)). We will, however, provide numerical confirmation of the consistency of the analysis in the next section.

5. Main Results

In this section we will consider a series of learning settings encompassed in the analytic framework in order to investigate the effectiveness and properties of knowledge distillation.

The external parameters of the studied model are the dataset size to input dimension ratio α , the cluster spread Δ , the relative size of the two clusters ρ , and the sparsity level of the student η . In the following we will focus on a representative case, with normal Gaussian noise $\Delta = 1$, unbalanced clusters $\rho = 0.2$ and a half-sparse student $\eta = 0.5$, and explore various ranges of α (some experiments in the balanced case are reported in Appendix D). Note that different choices for the external parameters do not affect any qualitative result presented in the following, but the present setting was conveniently chosen to induce a sizable performance gap between the dense teacher model and the sparse student.

We will also adjust the L_2 regularization intensity in the teacher and student losses, $\tilde{\lambda}$ and λ , and the KD mixing parameter, χ (see Eq. (4)), in order to evaluate the variation in the student generalization performance.

5.1. Inheriting the regularization

In the first experiment, we study whether learning from the outputs produced by the teacher instead of the true labels can indirectly regularize the student network and improve its generalization performance.

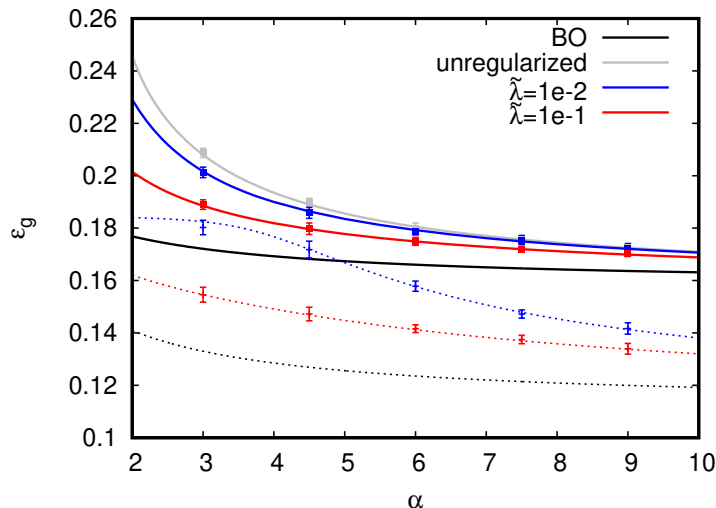


Figure 2: Comparison between the replica prediction for the generalization performance of a $\eta = 0.5$ sparse student (full curves) and the corresponding ridge regularized teacher (dashed curves), in a pure distillation setting ($\rho = 0.2$, $\Delta = 1$, $\chi = 1$), as a function of α . The data points with error bars represent the results of numerical experiments at $N = 4000$ (10 samples per point).

In Fig. 2, we compare the test error of an unregularized student learning from the true labels (grey curve) with the test error obtained by learning from the outputs of an L_2 -regularized teacher (with different intensity levels, blue and red curves). We call this a “pure distillation” setting, with $\chi = 1$ and $\lambda = 0$. In black we show the Bayes optimal lower bounds for teacher and student. The corresponding performance of the teacher is displayed with dashed lines.

The first observation we can make is that, indeed, there is a transfer of the regularization properties from the teacher to the student. In a typical logistic regression setting, the shape of the cross-entropy encourages a large student norm, since the produced outputs can match the ground-truth binary labels only when the magnitude of the student goes to infinity. Instead, the soft outputs of the teacher inform the student against overfitting the training set and growing the norm of the weights disproportionately. The second observation is that better teacher regularization induces better student regularization. In particular, we can see that in the large α regime, a better student performance is attained at the value of $\tilde{\lambda}$ which also optimizes the teacher performance. The student is thus able to inherit the fine-tuning done at the level of the teacher, even though it does not belong to the same model class. Note however that the KD student test error is still far from the displayed Bayes optimal bound.

In this plot we avoided showing the generalization behavior in the low α regime, which will be described in detail in section 6. Moreover, a more thorough analysis of the location of the optimal teacher regularization and the corresponding student performance can be found in Appendix B.

5.2. Limits of KD

Now that we have seen that KD can indirectly regularize the student, we continue by studying whether it can outperform direct regularization methods.

First, we compare the generalization curves obtained in the pure distillation setting described above ($\chi = 1$, learning only from the teacher outputs) with the effect of a simple L_2 penalty directly at the level of the student loss ($\chi = 0$, learning only from the true labels). In the top plot of Fig. 3, we display the test error in the two cases (full lines for $\chi = 1$, dashed lines for $\chi = 0$) at two dataset sizes $\alpha = 1.5$ (red) and $\alpha = 4.5$ (blue). The horizontal dashed lines highlight the best direct regularization performance. The black horizontal lines, instead, show the Bayes optimal bound (top $\alpha = 1.5$, bottom $\alpha = 4.5$). One can see that the two curves differ the most in the low α regime, where the direct regularization is outperforming KD (this will be clarified in Sec. 6). At higher values of α , instead, the effects of regularizing the teacher or directly regularizing the student and the associated generalization performances become almost indistinguishable. This is a positive result: with no fine-tuning at the level of the student loss (i.e., without the usual hyper-parameter optimization), a better teacher directly yields a better student performance. However, it is clear that in this case pure distillation is not leading to any improvements over simple ridge regularization.

We thus consider a second setting, where we add a direct L_2 regularization of intensity λ also on the student weights (in the KD loss), and then vary the mixing parameter χ in order to balance the total amount of regularization. Note that, since the student is sparse, the associated optimal amount of regularization will in general differ from the intensity $\tilde{\lambda}$ that yields the best performance for the dense teacher. In the bottom plot of Fig. 3, we fix $\alpha = 4.5$ and $\tilde{\lambda} = 0.1$ (optimal regularization regime for the teacher) and explore three values of λ for the student: over-regularized case ($\lambda = 0.5$), properly regularized case ($\lambda = 0.2$) and under-regularized case ($\lambda = 0.1$). These three regimes were identified *a posteriori* from extensive simulations at varying regularization levels. Again, the dashed horizontal line and the black horizontal line respectively mark the best performance achieved with perfect fine-tuning of the ridge regularization (at $\chi = 0$) and the Bayes optimal performance. We observe the following: in the over-regularized regime, adding even more regularization through KD is not beneficial and the best value of the mixing parameter is thus $\chi = 0$ (the minimum of the generalization is sub-optimal with respect to the grey line); in the other settings, instead, one finds an optimal value of χ at which the performance associated with optimal direct regularization is matched.

Overall these two experiments show a clear limitation of KD in the studied model: its effect is at best equal to that of a fine-tuned direct regularization scheme, and thus the student performance is still inferior with respect to the Bayes optimal one. In Appendix B we provide further details and take into account also a different setting, where the regularization scheme is based on the introduction of soft-labels: we can report here that also in that case the direct regularization and the inherited KD regularization induce the same generalization performance. It, of course, remains to be seen whether this above limitation of KD extends beyond the studied model.

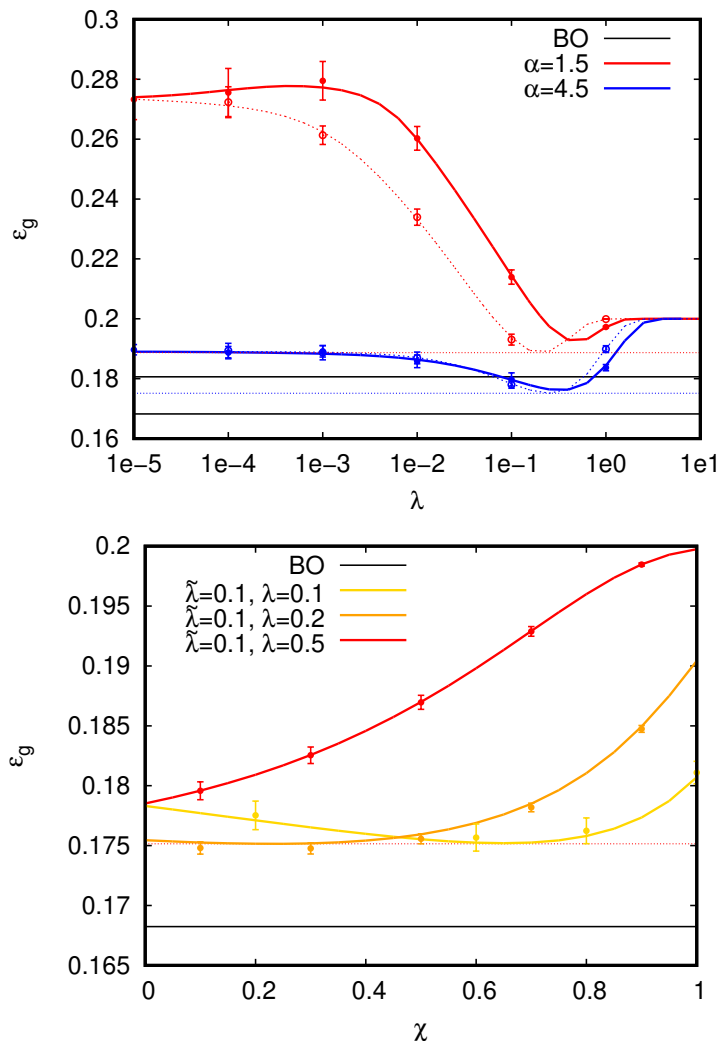


Figure 3: *Top plot:* Comparison between the replica prediction for the test error of a student learning from a regularized teacher (full curves) with $\chi = 1$, and a directly regularized student (dashed curves) with $\chi = 0$, at fixed values of α and at $\rho = 0.2$, $\Delta = 1$, $T = 1$ and $\eta = 0.5$. *Bottom plot:* variation in the test error induced by tuning the mixing parameter χ , at $\alpha = 4.5$, in correspondence of three regularization regimes in the student loss (under-regularized, properly regularized, over-regularized). (Horizontal dashed lines) Test error achieved by setting the optimal L_2 regularization intensity (at $\chi = 0$). (Black lines) Bayes optimal performance. The data points with error bars represent the results of numerical experiments at $N = 4000$ (10 samples per point).

5.3. Learning from a Bayes Optimal teacher

We finally consider a case where the teacher is not trained through an explicit regularization method: since in this setting it is not possible to regularize directly the student in a similar way, a transfer

learning strategy becomes necessary. This construction is meant to mimic more closely the behavior of knowledge distillation in usual deep learning settings, where learning algorithms play an implicit role in regularizing the network and their effectiveness may be dependent on the architecture.

In our framework, we study the performance of a student distilling the knowledge of a Bayes optimal teacher. As mentioned before, in the GM model there exists a point estimator that achieves the Bayes optimal generalization performance Eq. (13), which is characterized by an overlap with the signal v , a norm and a bias respectively equal to:

$$\tilde{m} = \frac{\mathbf{v} \cdot \mathbf{w}_{BO}}{N} = 1, \quad \tilde{q} = \frac{\|\mathbf{w}_{BO}\|^2}{N} = 1 + \Delta/\alpha, \quad \tilde{b} = \frac{\Delta(1 + \Delta/\alpha)}{2} \log\left(\frac{\rho}{1 - \rho}\right). \quad (23)$$

In order to obtain an analytical characterization of this special distillation setting, instead of taking

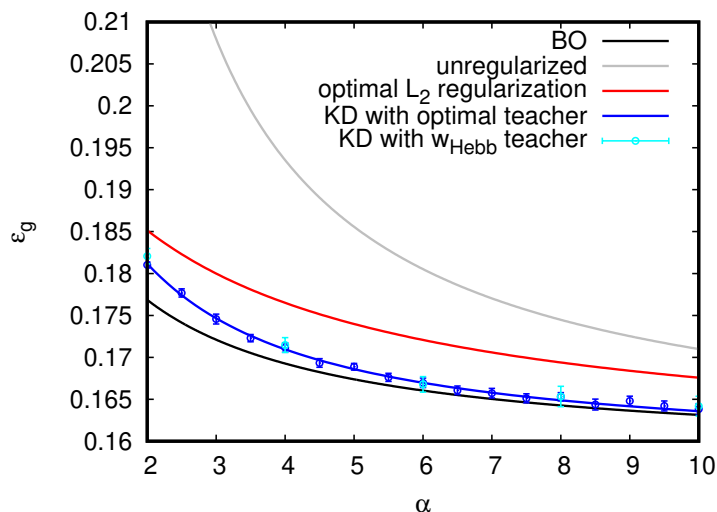


Figure 4: Replica prediction for the generalization performance of a $\eta = 0.5$ sparse student in a GM setting ($\rho = 0.2$, $\Delta = 1$). (Black curve) Performance bound, given by the generalization of the sparsified plug-in estimator. (Grey curve) Unregularized case. (Red curve) Student learning with a direct L_2 regularization of optimal intensity. (Blue curve) Pure distillation student ($\chi = 1$) learning from an optimal teacher (e.g. BO point estimator). The blue data points represent the results of simulations with a teacher produced according to Eq. 24. The cyan data points are instead obtained with \mathbf{w}_{BO} (Eq. 13) as a teacher. All numerical experiments were run at $N = 4000$ (10 samples per point).

the Hebbian estimator of Eq. (13) directly (carrying non-trivial correlations with the noise terms in the training data-points), we consider a teacher with the same bias and weights distributed as:

$$\tilde{\mathbf{w}} = \mathbf{v} + \sqrt{\frac{\Delta}{\alpha}} \mathbf{z}, \quad (24)$$

where each component of \mathbf{z} is i.i.d. Gaussian distributed with unit variance. A typical realization of this teacher will achieve exactly the Bayes optimal generalization performance, so we can use it as

a proxy for studying the distillation setting with the Hebbian estimator as a teacher. The details of the associated analytical calculation are reported in Appendix A.

In Fig. 4 we compare the generalization performance in three different learning settings: (grey line) an unregularized student, (red line) a student with direct L_2 regularization set at the optimal intensity and learning from the ground truth labels, and (blue line) a pure distillation student ($\chi = 1$) learning from a Bayes optimal teacher (according to the above definition). These results are again compared with the Bayes optimal performance bound (black line) relevant for the student.

When the transferred outputs are produced by an optimal teacher, we observe a clear improvement in the distillation test error with respect to a direct L_2 regularization: the gap with the optimal generalization bound is nearly closed (especially at large values of α). This result clearly shows the potential of knowledge distillation: through transfer learning the student can reach performances that are seemingly not achievable with direct regularization schemes, inheriting also the “implicit” regularization of the teacher (similar to what is observed in deep learning experiments (Frankle and Carbin (2019))).

6. Double descent in the KD framework

In this final section we will focus on the low α regime, i.e. number of samples small or comparable to the dimensionality, and the limit of zero direct regularization either in the student or in the teacher losses.

Note that, in the considered model, when the L_2 regularization term is completely switched off, the minimization of non-regularized logistic loss becomes equivalent to maximum likelihood estimation (MLE). In the GMM, with high probability the generated binary data will be linearly separable up until some threshold $\alpha_S(\rho, \Delta, \eta)$, and in this regime the ML estimator is ill-defined (Sur and Candès (2019)). Since we are not interested in addressing this issue in the present work, in the following we will consider a baseline regularization intensity of $\lambda = 0.00001$ as a proxy for the unregularized limit.

Let’s first consider taking an unregularized teacher and a pure distillation student ($\chi = 0$, $\tilde{\lambda}, \lambda \rightarrow 0$). The teacher training problem is perfectly separable below the separability threshold $\tilde{\alpha}_S$: without an explicit regularization its norm will thus diverge, due to the shape of the cross-entropy loss. Therefore, the outputs produced by the teacher will be quasi-binary (since the sigmoid activation will be completely saturated), and the student learning problem will look exactly like the usual logistic regression with binary labels. If we focus on the generalization behavior of the student, we thus expect a peak (and the corresponding double-descent behavior) at the student’s linear separability threshold α_S (Mignacco et al. (2020)). Note that in general $\alpha_S \neq \tilde{\alpha}_S$, since we take $\eta < 1$.

Now, let’s consider instead the case of a regularized teacher. In this case, even below $\tilde{\alpha}_S$, the teacher norm will remain finite and the produced outputs will be continuously distributed in the range $[0, 1]$. In this case, the pure distillation student will try to interpolate a set of non-binary outputs: as long as the number of linear constraints will be lower than the number of trainable weights $\alpha < \eta$, the student will be able to exactly reproduce the teacher outputs. However, just like in a normal regression scenario (Mignacco et al. (2020)), this will give rise to an interpolation peak at $\alpha = \eta$ (and the associated double-descent).

In Fig. 5, we can see the two types of peak (blue and red curves respectively). Note that the deviation of the experimental points from the theoretical predictions, at low α in the first regime, is due

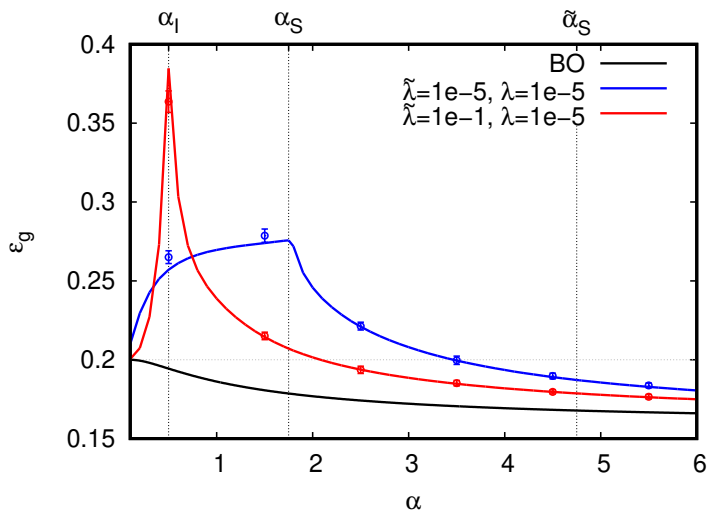


Figure 5: Generalization performance of a $\eta = 0.5$ sparse student in a pure distillation setting ($\rho = 0.2$, $\Delta = 1$, $\chi = 1$). (Black curve) Performance bound, given by the Bayes Optimal generalization. (Blue curve) Distillation curve when both teacher and student are not L_2 regularized. (Red curve) Distillation performance when only the teacher is regularized. α_I , $\tilde{\alpha}_S$ and α_S respectively denote the student interpolation threshold and the teacher and student separability thresholds. The data points with error bars represent the results of numerical experiments at $N = 4000$ (10 samples per point).

to the non convergence of the gradient descent simulations before the cutoff of 2000 optimization epochs.

In Appendix C, we also provide an in-depth analysis of teacher and student training losses and the mean squared error between teacher and student activations/preactivations around α_I and α_S , in order to further clarify the described phenomena.

7. Conclusions

In this work we developed a statistical physics framework for analysing knowledge distillation in high-dimensional models solvable with the replica method. The framework yields a deterministic description of the typical properties of the studied model, via a set of fixed point equations that track the behavior of the relevant order parameters.

We applied our framework to a prototypical case of knowledge distillation in the presence of mismatch between teacher and student networks. In particular, we considered two linear classifiers with different support (a stronger teacher network and a weaker student network) trained over a binary classification problem with data generated according to a Gaussian mixture.

We were able to highlight the inheritance properties of KD, showing that learning from properly regularized teachers can effectively transfer the good generalization properties to the student, with little fine-tuning at the level of the distillation loss. In our model, we also showed that, by distilling the knowledge of an optimal teacher, a model-agnostic KD student can approach the Bayes-optimal

generalization bound (relative to its sparsity level), whereas usual regularized logistic regression remains clearly sub-optimal. In order to validate the theoretical predictions, we offered a comparison with the results of finite size numerical experiments.

Finally, we analyzed the peculiar double-descent phenomenology that can appear in a KD setting, hybridizing between the interpolation peaks in regression problems and the separability threshold peak in classification problems.

The presented analytic results were obtained through the non-rigorous (yet exact) replica method, but the convex nature of the studied optimization problems suggests the possibility of an independent rigorous derivation via the Gordon minimax theorem (Gordon (1985)), along the lines of Mignacco et al. (2020). We leave this technical goal for future work.

Another natural but challenging research direction is to consider the case of more realistic data generative models (e.g. Goldt et al. (2019); Gerace et al. (2020)) and, more importantly, more complex neural network architectures (e.g., a random features Mei and Montanari (2019), or one hidden-layer networks Aubin et al. (2018)) that could allow for a more realistic mismatch between teacher and student models.

Acknowledgments

We thank Stéphane d’Ascoli for his inspiring presentation of the knowledge distillation problem. We acknowledge funding from the ERC under the European Union’s Horizon 2020 Research and Innovation Programme Grant Agreement 714608-SMiLe.

References

- Rohan Anil, Gabriel Pereyra, Alexandre Passos, Róbert Ormándi, George E. Dahl, and Geoffrey E. Hinton. Large scale distributed neural network training through online distillation. In *6th International Conference on Learning Representations, ICLR 2018*. OpenReview.net, 2018.
- Benjamin Aubin, Antoine Maillard, Florent Krzakala, Nicolas Macris, Lenka Zdeborová, et al. The committee machine: Computational to statistical gaps in learning a two-layers neural network. In *Advances in Neural Information Processing Systems*, pages 3223–3234, 2018.
- Carlo Baldassi, Federica Gerace, Carlo Lucibello, Luca Saglietti, and Riccardo Zecchina. Learning may need only a few bits of synaptic precision. *Physical Review E*, 93(5):052313, 2016.
- Carlo Baldassi, Federica Gerace, Hilbert J Kappen, Carlo Lucibello, Luca Saglietti, Enzo Tartaglione, and Riccardo Zecchina. Role of synaptic stochasticity in training low-precision neural networks. *Physical review letters*, 120(26):268103, 2018.
- Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019.
- Z Berkay Celik, David Lopez-Paz, and Patrick McDaniel. Patient-driven privacy control through generalized distillation. In *2017 IEEE Symposium on Privacy-Aware Computing (PAC)*, pages 1–12. IEEE, 2017.

- Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. In *Advances in Neural Information Processing Systems*, pages 742–751, 2017.
- Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Darkrank: Accelerating deep metric learning via cross sample similarities transfer. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- Andreas Engel and Christian Van den Broeck. *Statistical mechanics of learning*. Cambridge University Press, 2001.
- Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *ICLR*. OpenReview.net, 2019.
- Silvio Franz and Giorgio Parisi. Phase diagram of coupled glassy systems: A mean-field study. *Physical review letters*, 79(13):2486, 1997.
- Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1607–1616. PMLR, 2018.
- Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mezard, and Lenka Zdeborova. Generalisation error in learning with random features and the hidden manifold model. In *Proceedings of the 37th International Conference on Machine Learning*, pages 3452–3462. PMLR, 2020.
- Sebastian Goldt, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. Modelling the influence of data structure on learning in neural networks. *arXiv preprint arXiv:1909.11500*, 2019.
- Yehoram Gordon. Some inequalities for gaussian processes and applications. *Israel Journal of Mathematics*, 50(4):265–289, 1985.
- Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. In *Advances in neural information processing systems*, pages 1135–1143, 2015.
- Song Han, Huizi Mao, and William J. Dally. Deep compression: Compressing deep neural network with pruning, trained quantization and huffman coding. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.
- Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.
- Donald Olding Hebb. *The organization of behavior: A neuropsychological theory*. Psychology Press, 2005.
- Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015.

- Haiping Huang, KY Michael Wong, and Yoshiyuki Kabashima. Entropy landscape of solutions in the binary perceptron problem. *Journal of Physics A: Mathematical and Theoretical*, 46(37): 375002, 2013.
- Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2704–2713, 2018.
- Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, 2016.
- Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.
- Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.
- Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.
- Marc Mézard, Giorgio Parisi, and Miguel Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company, 1987.
- Francesca Mignacco, Florent Krzakala, Yue Lu, Pierfrancesco Urbani, and Lenka Zdeborova. The role of regularization in classification of high-dimensional noisy Gaussian mixture. In *Proceedings of the 37th International Conference on Machine Learning*, pages 6874–6883. PMLR, 2020.
- Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems*, pages 4694–4703, 2019.
- Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. In *ICLR (Workshop)*, 2015.
- Mary Phuong and Christoph Lampert. Towards understanding knowledge distillation. In *International Conference on Machine Learning*, pages 5142–5151, 2019.
- Arman Rahbar, Ashkan Panahi, Chiranjib Bhattacharyya, Devdatt P. Dubhashi, and Morteza Haghir Chehreghani. On the unreasonable effectiveness of knowledge distillation: Analysis in the kernel regime. *CoRR*, 2020.
- 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.
- Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

- Jiaxi Tang, Rakesh Shivanna, Zhe Zhao, Dong Lin, Anima Singh, Ed H Chi, and Sagar Jain. Understanding and improving knowledge distillation. *arXiv preprint arXiv:2002.03532*, 2020.
- Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4133–4141, 2017.
- Ruichi Yu, Ang Li, Vlad I Morariu, and Larry S Davis. Visual relationship detection with internal and external linguistic knowledge distillation. In *Proceedings of the IEEE international conference on computer vision*, pages 1974–1982, 2017.
- Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisit knowledge distillation: a teacher-free framework. *arXiv preprint arXiv:1909.11723*, 2019.
- Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *ICLR*, 2017.

Appendix A. Replica Computations

A.1. Typical learning in the GMM

We are interested in evaluating the following average free-entropy:

$$\Phi = \lim_{\beta \rightarrow \infty} \frac{1}{\beta N} \left\langle \log \int d\mathbf{w} e^{-\frac{\beta\lambda}{2} \|\mathbf{w}\|_2^2} \int db \prod_{\mu} e^{-\beta \ell(y^{\mu}, \sigma(\sum_{i=1}^N \frac{w_i x_i^{\mu}}{\sqrt{N}} + b))} \right\rangle_{\{\mathbf{x}^{\mu}, y^{\mu}\}} \quad (25)$$

where $\ell(\cdot)$ is the cross-entropy loss and the training data $\{\mathbf{x}^{\mu}, y^{\mu}\}$ is distributed according to a Gaussian Mixture model:

$$\mathbf{x}^{\mu} = (2y^{\mu} - 1) \frac{\mathbf{v}}{\sqrt{N}} + \mathbf{z} \quad (26)$$

with $v_i, z_i \sim \mathcal{N}(0, \Delta)$ and $y^{\mu} \sim \rho \delta(y^{\mu} - 1) + (1 - \rho) \delta(y^{\mu})$. Note that, because of the isotropy of the data model, instead of integrating over the possible realizations for the signal vector \mathbf{v} it is possible to fix the gauge $\mathbf{v} = (1, 1, \dots, 1)^T$ (e.g., as in [Engel and Van den Broeck \(2001\)](#)).

In the $\beta \rightarrow \infty$ limit, the statistical measure on the weights and the bias $\{\mathbf{w}, b\}$ focuses over the minimizer of the training loss, which yields the logistic regression optimization problem we want to characterize through this calculation. In order to evaluate the quenched average appearing in Eq. (25), we resort to the non-rigorous Replica Method, introduced in the context of Disordered Systems [Mézard et al. \(1987\)](#) and based on the identity:

$$\log(x) = \lim_{n \rightarrow 0} \frac{x^n - 1}{n}. \quad (27)$$

So, instead of evaluating the average of the free-energy directly, we introduce n interacting replicas of the original system and in the end we will recover the original expression by extrapolating the limit $n \rightarrow 0$.

We can thus focus on the calculation of the replicated volume:

$$\Omega^n = \int \prod_a d\mathbf{w}^a e^{-\frac{\beta\lambda}{2} \|\mathbf{w}\|_2^2} \int \prod_a db^a \prod_{\mu} \prod_a \left\langle e^{-\beta \ell(y^{\mu}, \sigma(\sum_{i=1}^N \frac{w_i^a x_i^{\mu}}{\sqrt{N}} + b^a))} \right\rangle_{\{\mathbf{x}^{\mu}, y^{\mu}\}} \quad (28)$$

$$= \int \prod_a d\mathbf{w}^a e^{-\frac{\beta\lambda}{2} \|\mathbf{w}\|_2^2} \int \prod_a db^a \int \prod_{\mu} \prod_a \frac{d\lambda_{\mu}^a d\hat{\lambda}_{\mu}^a}{2\pi} e^{i\hat{\lambda}_{\mu}^a (\lambda_{\mu}^a - \sum_{i=1}^N \frac{w_i^a x_i^{\mu}}{\sqrt{N}})} \times, \quad (29)$$

$$\prod_{\mu} \prod_a \left\langle e^{-\beta \ell(y^{\mu}, \sigma(\lambda_{\mu}^a + b^a))} \right\rangle_{\{\mathbf{x}^{\mu}, y^{\mu}\}}$$

where in the second line we isolated the dependency on each training data point $\mathbf{x}^{\mu}, y^{\mu}$ by introducing the preactivation variables λ_{μ}^a via Dirac's δ functions, allowing us to take the disorder average:

$$\mathbb{E}_{\mathbf{x}^{\mu}} e^{-i \sum_a \hat{\lambda}_{\mu}^a \frac{\mathbf{x}^{\mu} \cdot \mathbf{w}^a}{\sqrt{N}}} = e^{-i(2y^{\mu}-1) \sum_a \hat{\lambda}_{\mu}^a \frac{\sum w_i^a v_i}{N}} e^{-\frac{\Delta}{2} \sum_{ab} \hat{\lambda}_{\mu}^a \hat{\lambda}_{\mu}^b \frac{\sum w_i^a w_i^b}{N}} + \mathcal{O}(N^{-3/2}). \quad (30)$$

Now, we can introduce the overlap order parameters:

- $m^a = \frac{\sum w_i^a v_i}{N}$, representing the magnetization, i.e. the overlap between the learned weight configuration and the true signal v .

- $q^{ab} = \frac{\sum_i w_i^a w_i^b}{N}$, representing the overlap between two configurations sampled from the measure Eq. (25).

Then, we can rewrite the replicated volume as:

$$\Omega^n = \int \prod_a \frac{dm^a d\hat{m}^a}{2\pi/N} \int \prod_{ab} \frac{dq^{ab} d\hat{q}^{ab}}{2\pi/N} \int \prod_{ab} db^a G_I (G_S)^N (G_E)^{\alpha N} \quad (31)$$

where we separated the action in three different contributions: an interaction term, containing a trace over the order parameters and their conjugates

$$G_I = \exp \left(-N \left(\sum_a \hat{m}^a m^a + \sum_{ab} \hat{q}^{ab} q^{ab} \right) \right), \quad (32)$$

an entropic term, factorized over the components of the weight vector and containing the information about the regularization term in the loss function

$$G_S = \int \prod_a dw^a e^{-\frac{\beta\lambda}{2}(w^a)^2} \exp \left(\sum_a \hat{m}^a w^a + \sum_{ab} \hat{q}^{ab} w^a w^b \right), \quad (33)$$

and an energetic term, factorized over the training data points and containing the cross-entropy term:

$$G_E = \int \prod_a \left(\frac{d\lambda^a d\hat{\lambda}^a}{2\pi} e^{i\lambda^a \hat{\lambda}^a} \right) e^{-\frac{\Delta}{2} \sum_{ab} \hat{\lambda}_a \hat{\lambda}_b q^{ab}} \left\langle \prod_a e^{-\beta \ell(y, \sigma(\lambda^a + (2y-1)m^a + b^a))} \right\rangle_y. \quad (34)$$

A.1.1. REPLICAS SYMMETRIC ANSATZ

In order to proceed in the computation, we have to make an assumption on the structure of the order parameters and the geometric organization of the n replicas of the original system. Because of the convexity of the optimization problem we are considering we are justified in adopting the simplest possible assumption, the so-called Replica Symmetric ansatz, posing:

- $m^a = m$ for all $a = 1, \dots, n$, and same for their conjugates.
- $q^{ab} = q$ for all $a > b$, $q^{ab} = Q$ for all $a = b$, and same for their conjugates.
- $b^a = b$ for all $a = 1, \dots, n$.

Then, one can substitute the RS ansatz in the interaction term and easily obtain:

$$\frac{\log G_I}{nN} = g_I = - \left(\hat{m}m + \frac{\hat{Q}Q}{2} - \frac{\hat{q}q}{2} \right) \quad (35)$$

In the entropic term, after the substitutions one gets:

$$G_S = \int \prod_a dw^a e^{-\frac{\beta\lambda}{2}(w^a)^2} \exp \left(\hat{m} \sum_a w^a + \frac{1}{2} (\hat{Q} - \hat{q}) \sum_a (w^a)^2 + \frac{1}{2} \hat{q} \left(\sum_a w^a \right)^2 \right) \quad (36)$$

$$= \int \mathcal{D}z_0 \left\{ \int dw e^{-\frac{\beta\lambda}{2}(w)^2} \exp\left(\frac{1}{2}(\hat{Q} - \hat{q})w^2 + (\hat{m} + \sqrt{\hat{q}}z_0)w\right) \right\}^n \quad (37)$$

where in the second line a Hubbard-Stratonovich transformation, introducing the auxiliary variable $z_0 \sim \mathcal{N}(0, 1)$, allowed factorization over the replica index. Now we can take the logarithm in the $n \rightarrow 0$ limit, obtaining:

$$\frac{\log G_S}{nN} = g_S = \int \mathcal{D}z_0 \log \int dw \exp\left(\frac{1}{2}(\hat{Q} - \hat{q} - \beta\lambda)w^2 + (\hat{m} + \sqrt{\hat{q}}z_0)w\right). \quad (38)$$

Similarly, one can obtain the RS energetic contribution:

$$G_E = \mathbb{E}_y \int \prod_a \left(\frac{d\lambda^a d\hat{\lambda}^a}{2\pi} e^{i\lambda^a \hat{\lambda}^a} \right) e^{-\frac{\Delta}{2}(Q-q)\sum_a (\hat{\lambda}^a)^2 - \frac{\Delta}{2}q(\sum_a \hat{\lambda}^a)^2} \prod_a e^{-\beta \ell(y, \sigma(\lambda^a + (2y-1)m+b))} \quad (39)$$

$$= \mathbb{E}_y \int \mathcal{D}z_0 \left\{ \int \frac{d\lambda}{\sqrt{2\pi}} \frac{1}{\sqrt{\Delta(Q-q)}} e^{-\frac{1}{2}\frac{(\lambda + \sqrt{\Delta q}z_0)^2}{\Delta(Q-q)}} e^{-\beta \ell(y, \sigma(\lambda + (2y-1)m+b))} \right\}^n \quad (40)$$

$$= \mathbb{E}_y \int \mathcal{D}z_0 \left\{ \int \mathcal{D}\lambda e^{-\beta \ell(y, \sigma(\sqrt{\Delta(Q-q)}\lambda + \sqrt{\Delta q}z_0 + (2y-1)m+b))} \right\}^n \quad (41)$$

so taking the logarithm:

$$\frac{\log G_E}{n} = g_E = \mathbb{E}_y \int \mathcal{D}z_0 \log \int \mathcal{D}\lambda e^{-\beta \ell(y, \sigma(\sqrt{\Delta(Q-q)}\lambda + \sqrt{\Delta q}z_0 + (2y-1)m+b))}. \quad (42)$$

A.1.2. ZERO TEMPERATURE LIMIT

Now we can focus on the $\beta \rightarrow \infty$ limit, in which we recover the origin empirical risk minimization problem. Because of convexity, as we lower the temperature the overlap q between two different configurations sampled from the statistical measure will approach the typical norm Q , suggesting the scaling:

$$(Q - q) = \delta q / \beta \quad (43)$$

Moreover, the conjugate parameters also need to be properly rescaled:

$$(\hat{Q} - \hat{q}) = -\beta \delta \hat{q}, \quad \hat{q} \sim \beta^2 \hat{q}, \quad \hat{m} \sim \beta \hat{m}. \quad (44)$$

which gives, for the interaction term:

$$g_i = -\beta \left(\hat{m}m + \frac{1}{2}(\hat{q}\delta q - \delta\hat{q}q) \right) \quad (45)$$

In the entropic term one gets an integration over a one-dimensional optimization problem:

$$g_s = \beta \int \mathcal{D}z_0 \max_w \left(-\frac{\lambda + \delta\hat{q}}{2}w^2 + (\hat{m} + \sqrt{\hat{q}}z_0)w \right) \quad (46)$$

where the maximum is located at:

$$w^* = \frac{(\hat{m} + \sqrt{\hat{q}}z_0)}{(\lambda + \delta\hat{q})} \quad (47)$$

and the expression can be evaluated analytically, giving:

$$g_s = \beta \frac{\hat{m}^2 + \hat{q}}{2(\lambda + \delta \hat{q})}. \quad (48)$$

Finally, in the energetic contribution we get:

$$g_E = \mathbb{E}_y \int \mathcal{D}z M_E \quad (49)$$

$$M_E = \max_u -\frac{u^2}{2} - \ell \left(y, \sigma \left(\sqrt{\Delta \delta q} u + \sqrt{\Delta} z + (2y - 1) m + b \right) \right) \quad (50)$$

and one obtains the free-entropy:

$$\Phi = - \left(\hat{m} m + \frac{1}{2} (\hat{q} \delta q - \delta \hat{q} q) \right) + g_S + \alpha g_E. \quad (51)$$

The fixed-point equation that characterize the logistic regression problem in the high dimensional limit are obtained by extremizing the free-entropy with respect to the order parameters and their conjugates, which is nothing but a saddle-point condition for the action in the Statistical Physics framework.

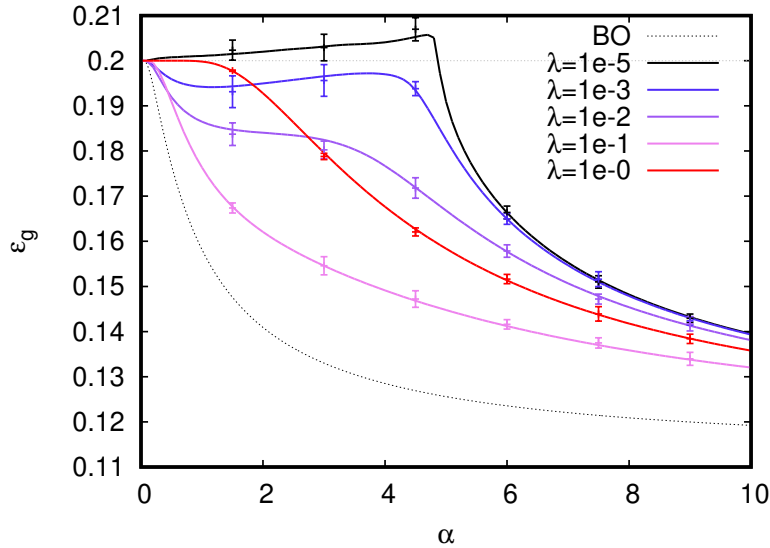


Figure 6: Classification in the GM model with $\rho = 0.2$ and $\Delta = 1$. Colored lines: linear classifier trained with logistic regression and ridge regularization. Dashed black line: plug-in estimator achieving the Bayes optimal performance.

In Fig. 6 we compare the generalization performance for a model trained with L_2 -regularized logistic regression with the Bayes-optimal performance, in the $\rho = 0.2$, $\Delta = 1$ setting. This will represent the baseline teacher model in the study of the distillation process. Note that the optimal value for the regularization parameter is of order $\lambda \sim 1e - 1$ and that higher values will hinder the generalization performance.

A.2. Distillation in the GMM

We will now present the derivation of the free energy expression for the distillation framework analyzed in the main text. In the following, all the parameters that refer to the teacher linear classifier will be denoted with an additional tilde.

As mentioned, in order to avoid the trivial scenario where the student converges exactly to the weight configuration of the teacher, we will consider a weaker “student” model, with a $0 < \eta < 1$ fraction of weights set to zero throughout the learning process. Thus, we aim to evaluate the expected value (in high dimensions) of the following free-entropy:

$$\Phi = \lim_{\beta \rightarrow \infty} \frac{1}{\beta N} \left\langle \left\langle \log \int \prod_{i=1}^{\eta N} dw_i e^{-\frac{\beta \lambda}{2} \|w\|_2^2} \int db \prod_{\mu} e^{-\frac{\beta}{2} \ell \left(\sigma \left(\frac{\tilde{w} \cdot x^\mu}{\sqrt{N}} + \tilde{b} \right), \sigma \left(\frac{w \cdot x^\mu}{\sqrt{N}} + b \right) \right)} \right\rangle_{\tilde{w}} \right\rangle_{\{x^\mu, y^\mu\}} \quad (52)$$

where the internal brackets represent an average over the teacher $\{\tilde{w}, \tilde{b}\}$ measure.

In order to evaluate the internal average we will use a different version of the replica trick, based on the following identity:

$$\langle f(\tilde{w}) \rangle_{\tilde{w}} = \frac{\int d\mu(\tilde{w}) f(\tilde{w})}{\int d\mu(\tilde{w})} = \lim_{\tilde{n} \rightarrow 0} \int \prod_{c=0}^{\tilde{n}} d\mu(\tilde{w}^c) f(\tilde{w}^1) \quad (53)$$

introducing \tilde{n} replicas of the teacher configuration but coupling only the first one to the student system. In this way, in the $\tilde{n} \rightarrow 0$ limit we can recover the expectation over the teacher measure. This type of formalism is closely related to the seminal work [Franz and Parisi \(1997\)](#) and was more recently applied for example in [Huang et al. \(2013\)](#); [Baldassi et al. \(2016, 2018\)](#).

As in the previous computation, the quenched disorder average can be taken only by replacing the logarithm in the definition of the free-entropy with the $n \rightarrow 0$ limit of the replicated system, so we will focus on the evaluation of:

$$\begin{aligned} & \frac{1}{N} \lim_{n, \tilde{n} \rightarrow 0} \partial_n \left\langle \lim_{\tilde{\beta} \rightarrow \infty} \lim_{\beta \rightarrow \infty} \int \prod_{c=1}^{\tilde{n}} d\tilde{w}^c e^{-\frac{\beta \tilde{\lambda}}{2} \|\tilde{w}^c\|_2^2} \int \prod_{c=1}^{\tilde{n}} d\tilde{b}^c \prod_{\mu, c} e^{-\frac{\tilde{\beta}}{2} \ell \left(y^\mu, \sigma \left(\sum_{i=1}^N \frac{\tilde{w}_i^c x_i^\mu}{\sqrt{N}} + \tilde{b}^c \right) \right)} \right. \\ & \times \left. \int \prod_{a=1}^n dw^a e^{-\frac{\beta \lambda}{2} \|w^a\|_2^2} \int \prod_{a=1}^n db^a \prod_{\mu, a} e^{-\frac{\beta}{2} \ell \left(\sigma \left(\sum_{i=1}^N \frac{w_i^a x_i^\mu}{\sqrt{N}} + b^a \right), \sigma \left(\sum_{i=1}^N \frac{w_i^a x_i^\mu}{\sqrt{N}} + b^a \right) \right)} \right\rangle_{\{x^\mu, y^\mu\}} \quad (54) \end{aligned}$$

Following the same steps as above, we introduce the Dirac’s δ (integral representation) for teacher and student preactivations, but in this case we will also separate the first η components of the teacher (where the student weights are non-zero) from the rest:

$$\begin{aligned} 1 = & \int \prod_{\mu, c} \frac{d\tilde{u}_\mu^c d\hat{u}_\mu^c}{2\pi} e^{i\hat{u}_\mu^c \left(\tilde{u}_\mu^c - \sum_{i=\eta N+1}^N \frac{\tilde{w}_i^c x_i^\mu}{\sqrt{N}} \right)} \int \prod_{\mu, c} \frac{d\tilde{\lambda}_\mu^c d\hat{\lambda}_\mu^c}{2\pi} e^{i\hat{\lambda}_\mu^c \left(\lambda_\mu^c - \sum_{i=1}^{\eta N} \frac{\tilde{w}_i^c x_i^\mu}{\sqrt{N}} \right)} \times \\ & \int \prod_{\mu, a} \frac{d\lambda_\mu^a d\hat{\lambda}_\mu^a}{2\pi} e^{i\hat{\lambda}_\mu^a \left(\lambda_\mu^a - \sum_{i=1}^{\eta N} \frac{w_i^a x_i^\mu}{\sqrt{N}} \right)}. \quad (55) \end{aligned}$$

Now we separately take the disorder average over the non-zeros components:

$$\prod_{i=1}^{\eta N} \mathbb{E}_{x_i^\mu} e^{-i \left(\sum_c \hat{\lambda}_c^\mu \frac{\tilde{w}_i^c}{\sqrt{N}} + \sum_a \hat{\lambda}_a^\mu \frac{w_i^a}{\sqrt{N}} \right) x_i^\mu} = \quad (56)$$

$$= e^{-i(2y^\mu - 1) \left(\sum_c \hat{\lambda}_c^\mu \frac{\sum_{i=1}^{\eta N} \tilde{w}_i^c v_i}{N} + \sum_a \hat{\lambda}_a^\mu \frac{\sum_{i=1}^{\eta N} w_i^a v_i}{N} \right)} \times \quad (57)$$

$$e^{-\frac{\Delta}{2} \left(\sum_{cd} \hat{\lambda}_c^\mu \hat{\lambda}_d^\mu \frac{\sum_{i=1}^{\eta N} \tilde{w}_i^c \tilde{w}_i^d}{N} + \sum_{ab} \hat{\lambda}_a^\mu \hat{\lambda}_b^\mu \frac{\sum_{i=1}^{\eta N} w_i^a w_i^b}{N} + 2 \sum_{ac} \hat{\lambda}_a^\mu \hat{\lambda}_c^\mu \frac{\sum_{i=1}^{\eta N} w_i^a \tilde{w}_i^c}{N} \right)}, \quad (58)$$

and the zeros components:

$$\prod_{i=\eta N+1}^N \mathbb{E}_{x_i^\mu} e^{-i \left(\sum_c \hat{u}_c^\mu \frac{\tilde{w}_i^c}{\sqrt{N}} \right) x_i^\mu} = \quad (59)$$

$$= e^{-i(2y^\mu - 1) \left(\sum_c \hat{u}_c^\mu \frac{\sum_{i=\eta N+1}^N \tilde{w}_i^c v_i}{N} \right) - \frac{\Delta}{2} \sum_{cd} \hat{u}_c^\mu \hat{u}_d^\mu \frac{\sum_{i=\eta N+1}^N \tilde{w}_i^c \tilde{w}_i^d}{N}}. \quad (60)$$

Finally, we introduce the overlap parameters:

- $\tilde{m}^c = \frac{\sum_{i=1}^{\eta N} \tilde{w}_i^c v_i}{N}$ and $\tilde{m}_0^c = \frac{\sum_{i=\eta N+1}^N \tilde{w}_i^c v_i}{N}$, representing the magnetization of the teacher in the direction of the true signal v .
- $m^a = \frac{\sum_{i=1}^{\eta N} w_i^a v_i}{N}$, representing the magnetization of the student in the direction of v .
- $\tilde{q}^{cd} = \frac{\sum_{i=1}^{\eta N} \tilde{w}_i^c \tilde{w}_i^d}{N}$ and $\tilde{q}_0^{cd} = \frac{\sum_{i=\eta N+1}^N \tilde{w}_i^c \tilde{w}_i^d}{N}$, representing the typical overlap between different teacher configurations.
- $q^{ab} = \frac{\sum_{i=1}^{\eta N} w_i^a w_i^b}{N} = \frac{\sum_{i=1}^{\eta N} w_i^a w_i^b}{N}$, representing the typical overlap between different student configurations.
- $S^{ac} = \frac{\sum_{i=1}^{\eta N} w_i^a \tilde{w}_i^c}{N}$, representing the typical overlap between teacher and student.

We can now rewrite the replicated volume as:

$$\begin{aligned} \Omega^n &= \lim_{\tilde{n} \rightarrow 0} \int \prod_c d\tilde{b}^c \int \prod_a db^a \int \prod_c \frac{d\tilde{m}^c d\tilde{m}_0^c}{2\pi/N} \int \prod_c \frac{d\tilde{m}_0^c d\tilde{m}_0^c}{2\pi/N} \int \prod_a \frac{dm^a d\hat{m}^a}{2\pi/N} \int \prod_{cd} \frac{d\tilde{q}^{cd} d\hat{q}^{cd}}{2\pi/N} \times \\ &\times \int \prod_{cd} \frac{d\tilde{q}_0^{cd} d\hat{q}_0^{cd}}{2\pi/N} \int \prod_{ab} \frac{dq^{ab} d\hat{q}^{ab}}{2\pi/N} \int \prod_{ac} \frac{dS^{ac} d\hat{S}^{ac}}{2\pi/N} G_I (G_S)^{\eta N} (G_{S_0})^{(1-\eta)N} (G_E)^{\alpha N} \quad (61) \end{aligned}$$

with the definitions for the interaction term:

$$G_I = \exp \left(-N \left(\sum_c (\hat{m}^c \tilde{m}^c + \hat{m}_0^c \tilde{m}_0^c) + \sum_a \hat{m}^a m^a + \sum_{cd} (\hat{q}^{cd} \tilde{q}^{cd} + \hat{q}_0^{cd} \tilde{q}_0^{cd}) + \sum_{ab} \hat{q}^{ab} q^{ab} + \sum_{ca} \hat{S}^{ca} S^{ca} \right) \right) \quad (62)$$

the two entropic terms:

$$G_S = \int \prod_c d\tilde{w}^c e^{-\frac{1}{2}\tilde{\beta}\tilde{\lambda}(\tilde{w}^c)^2} \int \prod_a dw^a e^{-\frac{1}{2}\beta\lambda(w^a)^2} \times \\ \times \exp\left(\sum_c \hat{m}^c \tilde{w}^c + \sum_a \hat{m}^a w^a + \sum_{cd} \hat{q}^{cd} \tilde{w}^c \tilde{w}^d + \sum_{ab} \hat{q}^{ab} w^a w^b + \sum_{ca} \hat{S}^{ca} w^a \tilde{w}^c\right), \quad (63)$$

$$G_{S_0} = \int \prod_c d\tilde{w}^c e^{-\frac{1}{2}\tilde{\beta}\tilde{\lambda}(\tilde{w}^c)^2} \exp\left(\sum_c \hat{m}_0^c \tilde{w}^c + \sum_{cd} \hat{q}_0^{cd} \tilde{w}^c \tilde{w}^d\right), \quad (64)$$

and the energetic term:

$$G_E = \mathbb{E}_y \int \prod_c \frac{d\tilde{u}^c d\hat{u}^c}{2\pi} e^{i\tilde{u}^c \tilde{u}^c} \int \prod_c \frac{d\tilde{\lambda}^c d\hat{\lambda}^c}{2\pi} e^{i\tilde{\lambda}^c \tilde{\lambda}^c} \int \prod_a \frac{d\lambda^a d\hat{\lambda}^a}{2\pi} e^{i\lambda^a \hat{\lambda}^a} \times \\ \times e^{-\frac{\Delta}{2}(\sum_{cd} \hat{u}_c \hat{u}_d \tilde{q}_0^{cd} + \sum_{cd} \hat{\lambda}_c \hat{\lambda}_d \tilde{q}^{cd} + \sum_{ab} \hat{\lambda}_a \hat{\lambda}_b q^{ab} + 2\sum_{ac} \hat{\lambda}_a \hat{\lambda}_c S^{ac})} \times \\ \times \prod_c e^{-\frac{\tilde{\beta}}{2}\ell(y, \tilde{u}^c + \tilde{\lambda}^c + (2y-1)(\tilde{m}_0^c + \tilde{m}^c) + \tilde{b}^c)} \prod_a e^{-\frac{\beta}{2}\ell(\tilde{u}^1 + \tilde{\lambda}^1 + (2y-1)(\tilde{m}_0^1 + \tilde{m}^1) + \tilde{b}^1, \lambda^a + (2y-1)m^a + b^a)} \quad (65)$$

A.2.1. REPLICA SYMMETRIC ANSATZ

Since also the student learning problem entails a convex optimization, we can safely assume Replica Symmetry to be realized also at the level of its order parameters. Moreover, as in the previous calculation, we should average over the realizations of the Gaussian signal \mathbf{v} but we can also exploit the isotropy and fix the gauge $\mathbf{v} = \mathbf{1}^T$. The two magnetizations characterizing the overlap between the teacher vector and the signal, \tilde{m} and \tilde{m}_0 (along the first ηN components and the complementary $(1-\eta)N$ components), will typically give $\tilde{m}/\tilde{m}_0 = \eta/(1-\eta)$. Thus we can pose:

- For the teacher magnetizations $\tilde{m}^c = \eta \tilde{m}$, $\tilde{m}_0^c = (1-\eta) \tilde{m}$. We can also set the conjugates to be equal: $\hat{m}^c = \hat{m}_0^c = \hat{m}$.
- Similarly, $\tilde{q}^{cd} = \eta \tilde{q}$, $\tilde{q}_0^{cd} = (1-\eta) \tilde{q}$ for $c \neq d$; and $\tilde{q}^{cd} = \eta \tilde{Q}$, $\tilde{q}_0^{cd} = (1-\eta) \tilde{Q}$ for $c = d$.
- For the student magnetization and self overlap $m^a = m$, $q^{ab} = q$ for $a \neq b$; $q^{ab} = Q$ for $a = b$.
- Since the student is coupled only to the first replica of the teacher, in general we will have two distinct teacher-student overlaps: $S^{ca} = S$ for $c = 1$, $S^{ca} = \tilde{S}$ for $c \neq 1$.

It is easy to see that the order parameters referred to the teacher are determined by the same saddle point-equations obtained in the previous section, since at finite \tilde{n} we can send $n \rightarrow 0$ and the $\mathcal{O}(1)$ term corresponds to the action Eq. (51). In order to get the fixed point equations that characterize the student, instead, we can substitute the $\tilde{n} \rightarrow 0$ limit explicitly and keep the $\mathcal{O}(n)$ terms in the series expansion at small n of the action in (61).

In this limit, the normalized logarithm of the interaction term yields:

$$g_I = - \left(\hat{m}m + \frac{\hat{Q}Q}{2} - \frac{1}{2}\hat{q}q + \hat{S}S - \hat{\tilde{S}}\tilde{S} \right). \quad (66)$$

In the $\tilde{\eta} \rightarrow 0$ limit, the entropic term G_{S_0} does not contribute to the saddle-point, so it will be ignored in the following. The calculations for G_S , instead, are a bit more involved than in the previous case because of the double average characterizing the distillation framework. After some manipulation and three separate Hubbard-Stratonovich, introducing the Gaussian variables x, z and \tilde{z} , one gets:

$$g_S = \frac{1}{n} \lim_{\tilde{n} \rightarrow 0} \log \int \mathcal{D}x \int \mathcal{D}z \int \mathcal{D}\tilde{z} \times \quad (67)$$

$$\times \int \prod_c d\tilde{w}^c \exp \left(\frac{1}{2} (\hat{Q} - \hat{q} - \tilde{\lambda}) \sum_c (\tilde{w}^c)^2 + \left(\hat{m} + \sqrt{\hat{S}}x + \sqrt{\hat{q} - \hat{\tilde{S}}}z \right) \sum_c \tilde{w}^c \right) \quad (68)$$

$$\times \int \prod_a dw^a \exp \left(\frac{1}{2} (\hat{Q} - \hat{q} - \lambda) \sum_a (w^a)^2 + \left(\hat{m} + \sqrt{\hat{S}}x + (\hat{S} - \hat{\tilde{S}})w_1 + \sqrt{\hat{q} - \hat{\tilde{S}}}z \right) \sum_a w^a \right) \quad (69)$$

$$= \int \mathcal{D}x \int \mathcal{D}z \int \mathcal{D}\tilde{z} \frac{\int d\tilde{w} \exp(\tilde{A}) \log(\int dw \exp(A))}{\int d\tilde{w} \exp(\tilde{A})} \quad (70)$$

where by sending $\tilde{n} \rightarrow 0$ we reestablished the expectation appearing in Eq. (52) and where we defined:

$$\tilde{A} = \frac{1}{2} (\hat{Q} - \hat{q} - \tilde{\lambda}) (\tilde{w})^2 + \left(\hat{m} + \sqrt{\hat{S}}x + \sqrt{\hat{q} - \hat{\tilde{S}}}z \right) \tilde{w} \quad (71)$$

$$A = \frac{1}{2} (\hat{Q} - \hat{q} - \lambda) (w)^2 + \left(\hat{m} + \sqrt{\hat{S}}x + (\hat{S} - \hat{\tilde{S}})w_1 + \sqrt{\hat{q} - \hat{\tilde{S}}}z \right) w. \quad (72)$$

After a couple rotations between the introduced Gaussian variables, it is possible to perform the $\int \mathcal{D}x$ integral analytically and obtain:

$$g_s/n = \int \mathcal{D}z \int \mathcal{D}\tilde{z} \frac{\int d\tilde{w} \exp(\tilde{A}) \log(\int dw \exp(A))}{\int d\tilde{w} \exp(\tilde{A})} \quad (73)$$

with:

$$\tilde{A}' = \frac{1}{2} (\hat{Q} - \hat{q} - \tilde{\lambda}) (\tilde{w})^2 + \left(\hat{m} + \sqrt{\hat{q}\tilde{z}} \right) \tilde{w} \quad (74)$$

$$A' = \frac{1}{2} (\hat{Q} - \hat{q} - \lambda) (w)^2 + \left(\hat{m} + (\hat{S} - \hat{\tilde{S}})w_1 + \frac{\hat{S}}{\sqrt{\hat{q}}}\tilde{z} + \sqrt{\hat{q} - \frac{\hat{S}^2}{\hat{q}}}z \right) w. \quad (75)$$

The calculation follows the same lines also for the energetic term, where after the introduction of four x, \tilde{z}, \hat{z} and z , one can factorize over the replica indices and send $\tilde{n} \rightarrow 0$, yielding:

$$g_E = \mathbb{E}_y \int \mathcal{D}x \int \mathcal{D}\tilde{z} \int \mathcal{D}\hat{z} \int \mathcal{D}z \times \frac{\int \frac{d\tilde{u}d\hat{u}}{2\pi} \int \frac{d\tilde{\lambda}d\hat{\lambda}}{2\pi} e^{\tilde{B}_0 + \tilde{B} - \frac{\beta}{2}\ell(y, \sigma(\tilde{u} + \tilde{\lambda} + (2y-1)\tilde{m} + \tilde{b}))} \log \int \frac{d\lambda d\hat{\lambda}}{2\pi} e^{B - \frac{\beta}{2}\ell(\sigma(\tilde{u} + \tilde{\lambda} + (2y-1)\tilde{m} + \tilde{b}), \sigma(\lambda + (2y-1)m + b))}}{\int \frac{d\tilde{u}d\hat{u}}{2\pi} \int \frac{d\tilde{\lambda}d\hat{\lambda}}{2\pi} e^{\tilde{B}_0 + \tilde{B} - \frac{\beta}{2}\ell(y, \sigma(\tilde{u} + \tilde{\lambda} + (2y-1)\tilde{m} + \tilde{b}))}} \quad (76)$$

where we defined:

$$\tilde{B}_0 = -\frac{\Delta}{2} (1 - \eta) (\tilde{Q} - \tilde{q}) \hat{u}^2 + i\hat{u} (\tilde{u} + \sqrt{\Delta(1-\eta)} \tilde{q}\hat{z}) \quad (77)$$

$$\tilde{B} = -\frac{\Delta}{2} \eta (\tilde{Q} - \tilde{q}) (\hat{\lambda})^2 + i\hat{\lambda} \left(\tilde{\lambda} + \sqrt{\Delta\tilde{S}x} + \sqrt{\Delta(\eta\tilde{q} - \tilde{S})} \tilde{z} \right) \quad (78)$$

$$B = -\frac{\Delta}{2} (Q - q) \hat{\lambda}^2 + i\hat{\lambda} \left(\lambda + \sqrt{\Delta\tilde{S}x} + \sqrt{\Delta(q - \tilde{S})} z + i\Delta (S - \tilde{S}) \hat{\lambda} \right) \quad (79)$$

In order to disentangle the Gaussian integrations and allow us to land on an expression where we can explicitly perform a few of them we start by shifting $z' = z + i\hat{\lambda} \frac{\Delta(S-\tilde{S})}{\sqrt{\Delta(q-\tilde{S})}}$. Now, we can proceed and simplify the $d\hat{\lambda}, d\tilde{\lambda}, d\hat{u}$ integrals and get, after shifting and rescaling $d\lambda, d\tilde{\lambda}, d\tilde{u}$:

$$g_E = \mathbb{E}_y \int \mathcal{D}x \int \mathcal{D}\tilde{z} \int \mathcal{D}\hat{z} \frac{\int \mathcal{D}z \int \mathcal{D}\tilde{u} \int \mathcal{D}\tilde{\lambda} e^{-\frac{\beta}{2}\ell(y, \sigma(\tilde{h}'))} \log \int \mathcal{D}\lambda e^{-\frac{\beta}{2}\ell(\sigma(\tilde{h}'), \sigma(h))}}{\int \mathcal{D}\tilde{u} \int \mathcal{D}\tilde{\lambda} e^{-\frac{\beta}{2}\ell(y, \sigma(\tilde{h}))}}, \quad (80)$$

with:

$$\tilde{h}' = \sqrt{\Delta(1-\eta)} (\tilde{Q} - \tilde{q}) \tilde{u} - \sqrt{\Delta(1-\eta)} \tilde{q}\hat{z} + \sqrt{\Delta \left(\eta (\tilde{Q} - \tilde{q}) - \frac{(S - \tilde{S})^2}{q - \tilde{S}} \right)} \tilde{\lambda} + \left(\sqrt{\Delta\tilde{S}x} + \sqrt{\Delta(\eta\tilde{q} - \tilde{S})} \tilde{z} + \frac{\Delta(S - \tilde{S})}{\sqrt{\Delta(q - \tilde{S})}} z \right) + (2y-1) \tilde{m} + \tilde{b}, \quad (81)$$

$$\tilde{h} = \sqrt{\Delta(1-\eta)} (\tilde{Q} - \tilde{q}) \tilde{u} - \sqrt{\Delta(1-\eta)} \tilde{q}\hat{z} + \sqrt{\Delta\eta} (\tilde{Q} - \tilde{q}) \tilde{\lambda} - \left(\sqrt{\Delta\tilde{S}x} + \sqrt{\Delta(\eta\tilde{q} - \tilde{S})} \tilde{z} \right) + (2y-1) \tilde{m} + \tilde{b} \quad (82)$$

$$h = \sqrt{\Delta(Q-q)} \lambda - \left(\sqrt{\Delta\tilde{S}x} + \sqrt{\Delta(q - \tilde{S})} z \right) + (2y-1) m + b. \quad (83)$$

After a few rotations between $\tilde{\lambda}, \tilde{u}, z, \hat{z}, \tilde{z}$ and x , one can perform the $\mathcal{D}\tilde{u}, \mathcal{D}\hat{z}, \mathcal{D}x$ integrals and get the final expression:

$$g_E/n = \left\langle \int \mathcal{D}z \int \mathcal{D}\hat{z} \frac{\int \mathcal{D}\tilde{\lambda} e^{-\frac{\beta}{2}\ell(y, \sigma(\tilde{h}))} \log \int \mathcal{D}\lambda e^{-\frac{\beta}{2}\ell(\sigma(\tilde{h}), \sigma(h))}}{\int \mathcal{D}\tilde{\lambda} e^{-\frac{\beta}{2}\ell(y, \sigma(\tilde{h}))}} \right\rangle_y \quad (84)$$

with the definitions:

$$\tilde{h} = \sqrt{\Delta (\tilde{Q} - \tilde{q})} \tilde{\lambda} + \sqrt{\Delta \tilde{q}} \tilde{z} + (2y - 1) \tilde{m} + \tilde{b}, \quad (85)$$

$$h = \sqrt{\Delta (Q - q)} \lambda + \frac{\Delta (S - \tilde{S}) \tilde{\lambda}}{\sqrt{\Delta (\tilde{Q} - \tilde{q})}} + \frac{\sqrt{\Delta \tilde{S}} \tilde{z}}{\sqrt{\tilde{q}}} + \sqrt{\Delta \left(q - \frac{\tilde{S}^2}{\tilde{q}} - \frac{(S - \tilde{S})^2}{(\tilde{Q} - \tilde{q})} \right)} z + (2y - 1) m + b. \quad (86)$$

A.2.2. ZERO TEMPERATURE LIMIT

Finally, we can recover the nested optimization characterizing the distillation framework by successively taking the two limits $\tilde{\beta} \rightarrow \infty$ and $\beta \rightarrow \infty$. As in the previous calculation, we have to introduced rescaled overlap order parameters:

$$(\tilde{Q} - \tilde{q}) = \delta \tilde{q} / \tilde{\beta}, \quad (Q - q) = \delta q / \beta, \quad S - \tilde{S} = \delta S / \tilde{\beta} \quad (87)$$

and rescale also all the conjugate parameters:

- $\hat{Q} \rightarrow \tilde{\beta}^2 \hat{q} + \mathcal{O}(\tilde{\beta}), \hat{q} \rightarrow \tilde{\beta}^2 \hat{q}, (\hat{Q} - \hat{q}) \rightarrow -\tilde{\beta} \delta \hat{q}, \hat{m} \rightarrow \tilde{\beta} \hat{m},$
- $\hat{Q} \rightarrow \beta^2 \hat{q} + \mathcal{O}(\beta), \hat{q} \rightarrow \beta^2 \hat{q}, (\hat{Q} - \hat{q}) \rightarrow -\beta \delta \hat{q}, \hat{m} \rightarrow \beta \hat{m},$
- $\hat{S} \rightarrow \tilde{\beta} \beta \hat{S} + \mathcal{O}(\beta), \hat{S} \rightarrow \tilde{\beta} \beta \hat{S}, (\hat{S} - \hat{S}) \rightarrow \beta \delta \hat{S}.$

With these scalings, the interaction term reads:

$$g_I = -\beta \left(\hat{m} m + \frac{1}{2} (\hat{q} \delta q - \delta \hat{q} q) + \left(\hat{S} \delta S + \delta \hat{S} \tilde{S} \right) \right) + \mathcal{O}(1). \quad (88)$$

In the entropic term, in the zero-temperature limit the integrals over the teacher and student weights become extremum operations:

$$g_S = \lim_{\beta \rightarrow \infty} \beta \int \mathcal{D}z \int \mathcal{D}\tilde{z} M_s^* \quad (89)$$

where:

$$\begin{aligned} M_s^* &= \max_w \left\{ -\frac{1}{2} (\lambda + \delta \hat{q}) w^2 + \left(\hat{m} + \delta \hat{S} \tilde{w}^* + \frac{\hat{S}}{\sqrt{\hat{q}}} \tilde{z} + \sqrt{\frac{\hat{q} \hat{q} - \hat{S}^2}{\hat{q}}} z \right) w \right\} \\ &= \frac{1}{2} \frac{\left(\hat{m} + \delta \hat{S} \tilde{w}^* + \frac{\hat{S}}{\sqrt{\hat{q}}} \tilde{z} + \sqrt{\frac{\hat{q} \hat{q} - \hat{S}^2}{\hat{q}}} z \right)^2}{\lambda + \delta \hat{q}} \end{aligned} \quad (90)$$

and where the teacher weight configuration, as in Eq. (47), maximizes the action:

$$\tilde{w}^* = \operatorname{argmax}_{\tilde{w}} \left\{ -\frac{1}{2}(\tilde{\lambda} + \delta\hat{q})\tilde{w}^2 + (\hat{m} + \sqrt{\hat{q}\tilde{z}})\tilde{w} \right\} = \frac{\hat{m} + \sqrt{\hat{q}\tilde{z}}}{\tilde{\lambda} + \delta\hat{q}}. \quad (91)$$

With an analytic expression for the maxima, the $\int \mathcal{D}z \int \mathcal{D}\tilde{z}$ integrations can be carried out, giving:

$$g_S = \frac{\beta}{2} \frac{\left(\hat{m} + \hat{m} \frac{\delta\hat{S}}{\lambda + \delta\hat{q}} \right)^2 + \left(\frac{\hat{S}}{\sqrt{\hat{q}}} + \frac{\delta\hat{S}\sqrt{\hat{q}}}{\lambda + \delta\hat{q}} \right)^2 + \frac{\hat{q}\hat{q} - \hat{S}^2}{\hat{q}}}{\lambda + \delta\hat{q}}. \quad (92)$$

Lastly, in the $\tilde{\beta}, \beta \rightarrow \infty$, also the $\mathcal{D}\lambda, \mathcal{D}\tilde{\lambda}$ integrals in the energetic term become one-dimensional extremum operations, and we have:

$$g_E = \beta \mathbb{E}_y \int \mathcal{D}z \int \mathcal{D}\tilde{z} M_E^* \quad (93)$$

where:

$$M_E^* = \max_{\lambda} \left\{ -\frac{1}{2} \frac{\lambda^2}{\Delta \delta q} - \frac{1}{2} \ell \left(\sigma \left(\tilde{h}(\lambda^*) \right), \sigma \left(h(\lambda) \right) \right) \right\} \quad (94)$$

with:

$$\tilde{h}(\tilde{\lambda}) = \tilde{\lambda} + (2y - 1) \tilde{m} + \tilde{b} + \sqrt{\Delta \tilde{q} \tilde{z}} \quad (95)$$

$$h(\lambda) = \lambda + (2y - 1) m + b + \frac{\delta S}{\delta \tilde{q}} \tilde{\lambda} + \sqrt{\Delta} \frac{S}{\sqrt{\tilde{q}}} \tilde{z} + \sqrt{\Delta \left(q - \frac{S^2}{\tilde{q}} + \mathcal{O}\left(\frac{1}{\tilde{\beta}}\right) \right)} z \quad (96)$$

and

$$\tilde{\lambda}^* = \operatorname{argmax}_{\tilde{\lambda}} -\frac{1}{2} \frac{\tilde{\lambda}^2}{\Delta \delta \tilde{q}} - \frac{1}{2} \ell \left(y, \sigma \left(\tilde{h}(\tilde{\lambda}) \right) \right). \quad (97)$$

The free-entropy of Eq. (16) is thus recovered after dividing the various contributions by β :

$$\Phi = - \left(\hat{m}m + \frac{1}{2} (\hat{q}\delta q - \delta\hat{q}q) + \left(\hat{S}\delta S + \delta\hat{S}S \right) \right) + \eta g_S + \alpha g_E \quad (98)$$

A.3. Distillation with optimal teacher

The replica calculation for the distillation setting with optimal teacher is very similar to the one presented in the first section of the Appendix (for the typical logistic regression framework), since we make an explicit assumption on the statistical measure for the teacher weight vector and the double average (as in the previous section) is not needed.

As described in the main text, we assume the teacher to be represented by a noisy version of the signal $\mathbf{w} = \mathbf{v} + \sqrt{\frac{\Delta}{\alpha}} \mathbf{h}$, where each component of the noise is independent and normal distributed $h_i \sim \mathcal{N}(0, 1)$. This choice induces an average magnetization $\tilde{m} = 1 + \mathcal{O}(N^{-1/2})$ and a norm $\tilde{q} = 1 + \frac{\Delta}{\alpha}$, same as in the case of the optimal plug-in estimator of Eq. (13). When the bias is set to $\tilde{b} = \Delta \frac{(1+\Delta/\alpha)}{2} \log\left(\frac{\rho}{1-\rho}\right)$ the teacher achieves a generalization performance matching the Bayes optimal generalization, and this justifies our modeling choice for characterizing distillation from an optimal teacher.

We thus want to evaluate the free-entropy for an η -sparse student learning from the outputs produced by this optimal teacher:

$$\Phi = \lim_{\beta \rightarrow \infty} \frac{1}{\beta N} \left\langle \log \int d\mathbf{w}_\eta e^{-\frac{\lambda}{2} \|\mathbf{w}\|_\eta^2} \int db \prod_\mu e^{-\beta \ell \left(\sigma \left(\frac{(\mathbf{v} + \mathbf{h} \sqrt{\Delta/\alpha}) \cdot \mathbf{w}^\mu}{\sqrt{N}} + \frac{\Delta(1+\Delta/\alpha)}{2} \log \frac{\rho}{1-\rho} \right), \sigma \left(\frac{\mathbf{w} \cdot \mathbf{w}^\mu}{\sqrt{N}} + b \right) \right)} \right\rangle_{\{\mathbf{x}^\mu, \mathbf{y}^\mu, \mathbf{v}, \mathbf{h}\}} \quad (99)$$

As usual, instead of actually averaging over \mathbf{v} we will set $\mathbf{v} = \mathbf{1}^T$. We can isolate the dependency over the training set by introducing the preactivation variables l^μ (for the teacher) and λ_a^μ (for the student) via Dirac's delta functions and then perform the disorder average:

$$\mathbb{E}_{\mathbf{x}^\mu} e^{-i \sum_a \hat{\lambda}_a^\mu \frac{\mathbf{x}^\mu \cdot \mathbf{w}^a}{\sqrt{N}} - i \hat{l}_\mu \frac{(\mathbf{v}_i + \mathbf{h}_i \sqrt{\Delta/\alpha}) \cdot \mathbf{x}^\mu}{\sqrt{N}}} = \quad (100)$$

$$= e^{-i(2y^\mu - 1) \left(\sum_a \hat{\lambda}_a^\mu \frac{\sum_{i=1}^{\eta N} w_i^a}{N} + \hat{l}_\mu \left(1 + \sqrt{\Delta/\alpha} \frac{\sum_{i=1}^N h_i}{N} \right) \right)} \times \quad (101)$$

$$\times e^{-\frac{\Delta}{2} \left(\sum_{ab} \hat{\lambda}_a^\mu \hat{\lambda}_b^\mu \frac{\sum_{i=1}^{\eta N} w_i^a w_i^b}{N} + 2 \hat{l}_\mu \sum_a \hat{\lambda}_a^\mu \left(\frac{\sum_{i=1}^{\eta N} w_i^a}{N} + \sqrt{\Delta/\alpha} \frac{\sum_{i=1}^N w_i^a h_i}{N} \right) + \hat{l}_\mu^2 \left(1 + \Delta/\alpha \frac{\|\mathbf{h}\|^2}{N} + 2 \sqrt{\Delta/\alpha} \frac{\sum_{i=1}^N h_i}{N} \right) \right)},$$

which leads us to introducing the overlap order parameters:

$$m^a = \frac{\sum w_i^a}{N}, \quad S^a = \frac{\sum w_i^a h_i}{N}, \quad q^{ab} = \frac{\sum_i w_i^a w_i^b}{N}. \quad (102)$$

On the other hand, we know that for a Gaussian i.i.d. normal vector \mathbf{h} the magnetization in the direction of the signal and the norm will be $\tilde{m} = \frac{\sum h_i}{N} = 0$ and $\tilde{q} = \frac{\sum h_i^2}{N} = 1$.

The replicated volume thus reads:

$$\Omega^n = \int \prod_a \frac{d\tilde{S}^a d\hat{S}^a}{2\pi/N} \int \prod_a \frac{dm^a d\hat{m}^a}{2\pi/N} \int \prod_{ab} \frac{dq^{ab} d\hat{q}^{ab}}{2\pi/N} \int \prod_a db^a G_I (G_S)^{\eta N} (G_E)^{\alpha N} \quad (103)$$

with the definitions:

$$G_I = \exp \left(-N \left(\sum_a \hat{m}^a m^a + \sum_a \hat{S}^a \tilde{S}^a + \sum_{ab} \hat{q}^{ab} q^{ab} \right) \right) \quad (104)$$

$$G_S = \int \mathcal{D}h \int \prod_a d\mu(w^a) \exp \left(\sum_a \hat{m}^a w^a + \sum_a \hat{S}^a w^a h + \sum_{ab} \hat{q}^{ab} w^a w^b \right) \quad (105)$$

and (after a little algebra):

$$G_E = \int \mathcal{D}l \int \prod_a \left(\frac{d\lambda^a d\hat{\lambda}^a}{2\pi} e^{i\lambda^a \hat{\lambda}^a} \right) \times \\ \times e^{-\frac{\Delta}{2} \sum_{ab} \hat{\lambda}_a \hat{\lambda}_b \left(q^{ab} - \frac{(m^a + \sqrt{\Delta/\alpha} S^a)(m^b + \sqrt{\Delta/\alpha} S^b)}{1 + \Delta/\alpha} \right) - i l \sqrt{\frac{\Delta}{(1+\Delta/\alpha)}} \sum_a \hat{\lambda}_a (m^a + \sqrt{\Delta/\alpha} S^a)} \\ \times \mathbb{E}_y \prod_a e^{-\beta \ell \left(\sigma \left(l \sqrt{\Delta(1+\Delta/\alpha)} + (2y-1) + \frac{\Delta(1-\Delta/\alpha)}{2} \log \frac{\rho}{1-\rho} \right), \sigma(\lambda^a + (2y-1)m^a + b^a) \right)} \quad (106)$$

A.3.1. RS ANSATZ AND ZERO TEMPERATURE LIMIT

The rest of the computation is basically identical to that presented in the first section of the Appendix, so we will only report the final expressions for the free-entropy:

$$\Phi = g_I + \eta g_S + \alpha g_E. \quad (107)$$

with the following definitions for the interaction, entropic and energetic terms:

$$g_I = - \left(\hat{m}m + \hat{S}S + \frac{1}{2} (\hat{q}\delta q - \delta\hat{q}q) \right), \quad (108)$$

$$g_S = \frac{\hat{m}^2 + \hat{S}^2 + \hat{q}}{2(\lambda + \delta\hat{q})}, \quad (109)$$

$$g_E = \mathbb{E}_y \int \mathcal{D}\tilde{z} \int \mathcal{D}z M_E, \quad (110)$$

where:

$$M_E = \max_u - \frac{u^2}{2} - \ell \left(\sigma(\tilde{h}(\tilde{z})), \sigma(h(u, z, \tilde{z})) \right), \quad (111)$$

and:

$$\tilde{h}(\tilde{z}) = (2y - 1) + \frac{\Delta'}{2} \log \frac{\rho}{1 - \rho} + \sqrt{\Delta(1 - \Delta/\alpha)}\tilde{z}, \quad (112)$$

$$h(u, z, \tilde{z}) = \sqrt{\Delta\delta q}u + \sqrt{\Delta \left(q - \frac{(m + \sqrt{\Delta/\alpha}S)^2}{1 + \Delta/\alpha} \right)}z + \sqrt{\frac{\Delta}{(1 + \Delta/\alpha)}}(m + \sqrt{\Delta/\alpha}S)\tilde{z} + (2y - 1)m + b. \quad (113)$$

Appendix B. On the inheritance of the teacher regularization

B.1. Direct and inherited student regularization

We here show additional experiments on the different effects of direct and indirect (inherited through KD) L_2 regularization on the student generalization performance.

In Fig. 7, we display the distillation generalization curves (full lines) and the direct regularization curves (dashed lines) at fixed values of the regularizer intensity. It is evident that the optimal regularization for the student (with sparsity $\eta = 0.5$) is of the same order of the optimal value for the teacher $\lambda \simeq 0.1$ (cfr. Fig. 6). If we look at the low α regime, it is also clear that only a direct ridge regularization on the student can grant him a good generalization performance (in the pure distillation setting we see the α_I and α_S interpolation peaks). This observation strongly motivates a mixed approach, where the distillation student is also regularized with an L_2 penalty (we will consider this setting in the following). However, in the large α regime we observe an interesting advantage of distillation: in the case of an overshoot in the regularization intensity ($\lambda \simeq 1$), the distillation student performance is less hindered than the directly regularized student. This provides an indication that the knowledge distillation process may require less fine-tuning than usual empirical risk minimization with cross-entropy loss.

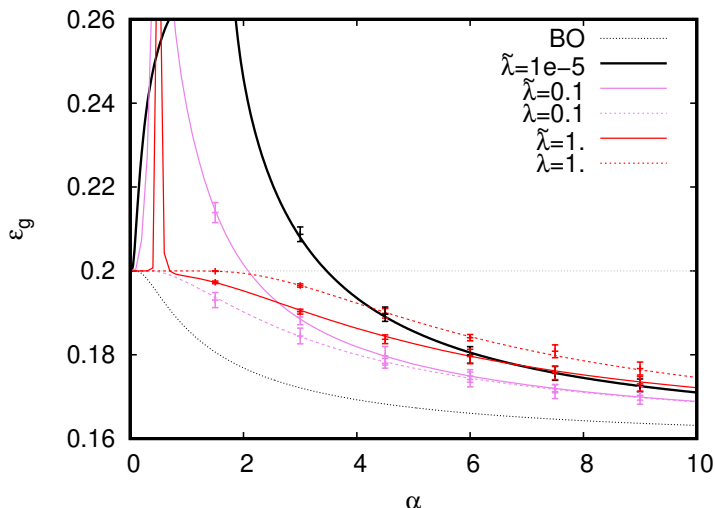


Figure 7: Comparison between the pure distillation generalization curves (blue full curves), as in Fig. 3, and the performance of a ridge regularized student learning from the labels (red dashed curves), at $\rho = 0.2$, $\Delta = 1$. (Grey curve) Optimal generalization, achieved with the sparsified plug-in estimator. The data points with error bars represent the results of numerical experiments at $N = 4000$ (10 samples per point).

B.2. On the KD mixing parameter

We here show some more details on the effect of the mixing parameter χ in the distillation loss.

In Fig. 8 we fix $\tilde{\lambda}, \lambda = 0.15$ and vary the value of the mixing parameter χ . As expected from the results reported in the previous paragraphs, we can see that in the low α regime it is better not to rely upon the teacher outputs in order to avoid the interpolation cusp (around $\alpha = \eta$): lower values of χ (red, pink curves) yield better generalization. As α increases, an appropriate value of the mixing parameter can increase the overall regularization felt by the student up to the optimal amount, guaranteeing an improved performance (purple curve). Interestingly, through the tuning of χ KD can be made to match the performance achieved with optimal regularization, however a reduction in the performance gap with respect to the plug-in estimator bound is never observed.

B.3. Regularization through uniform label smoothing

We consider a different type of regularization scheme called uniform label smoothing [Szegedy et al. \(2016\)](#); [Müller et al. \(2019\)](#). We introducing a smoothing parameter ϵ and replacing the "hard" ground truth labels with their softer counterpart $y \rightarrow y(1 - \epsilon) + (1 - y)\epsilon$ at training. This type of regularization strategy is known to be effective in preventing overconfidence of the trained model, especially in the case of noisy data.

In Fig. 9, we compare two different scenarios: in the first case, the student learns from the smoothed labels directly (full colored lines); in the second case, we consider pure distillation from a teacher that learned the smoothed labels (black dots). In both settings the ridge regularization is fixed at the baseline level $\lambda, \tilde{\lambda} = 1e-5$ in order to isolate the regularization effect of the softer labels.

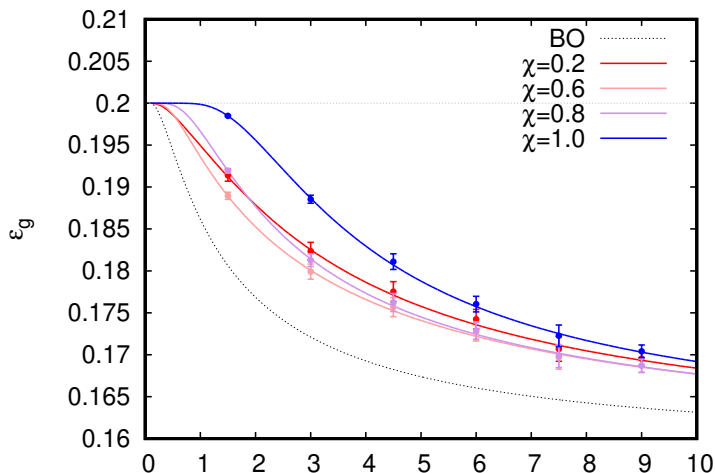


Figure 8: Generalization performance achieved by a regularized distillation student ($\eta = 0.5$, $\lambda = 1e - 1$, $T = 2$) learning from a optimally regularized teacher ($\tilde{\lambda} = 0.15$), at varying values of the mixing parameter χ and with $\rho = 0.2$, $\Delta = 1$. (Dashed black curve) Generalization bound, achieved by the sparsified plug-in estimator. The data points with error bars represent the results of numerical experiments at $N = 4000$ (10 samples per point).

Surprisingly, the student generalization performance obtained in the two scenarios is practically indistinguishable, implying that the KD process can perfectly transfer this type of regularization. Note that, because of the simple nature of the GM generative model, we find it is very beneficial to have learn with higher ϵ at smaller values of α .

B.4. Varying the Knowledge Distillation temperature

We consider again a pure distillation setting with $\chi = 1$, but now explore the effect of changing the distillation temperature T . In particular, in the cross-entropy term in the knowledge distillation loss, the usual outputs will be replaced by $\sigma(h) \rightarrow \sigma(h/T)$, where T can increase (lower) the difference between the probabilities of assigning each label.

The general idea behind the introduction of this temperature is that after training, in a multi-class problem, the *softmax* output function will typically produce small probabilities in correspondence of the incorrect categories and the difference in the assigned weight will be flattened because of the Boltzmann-like form of the activation. Introducing a high temperature can instead reweight the output probability distribution, accentuating the differences in probabilities assigned to incorrect labels [Hinton et al. \(2015\)](#); [Tang et al. \(2020\)](#). Of course, this effect cannot be explored in the simple binary classification setting. However, it is still possible to observe a positive effect of a high T in the low α regime.

In [Fig.10](#) we vary the distillation temperature in the small training set regime, considering an unregularized student that learns from a teacher with ridge regularization intensity $\tilde{\lambda} = 0.15$ (nearly optimal setting). It is clear that raising the distillation temperature can mitigate the overfitting

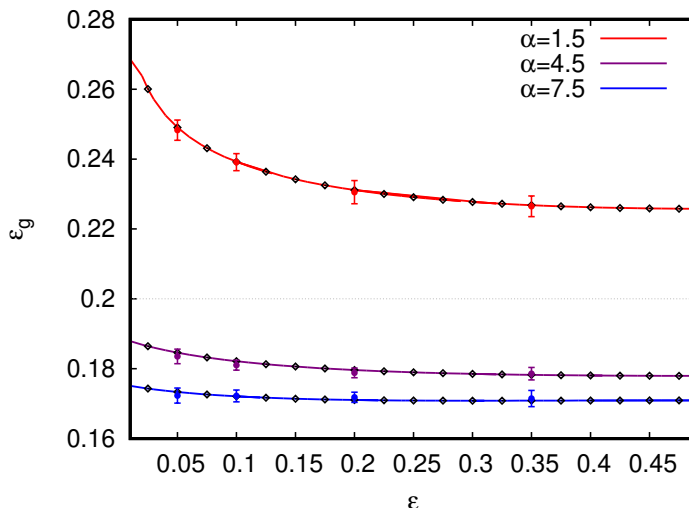


Figure 9: Comparison of the generalization performance at fixed values of α of a $\eta = 0.5$ sparse student in two scenarios: either the student learns directly the smoothed labels (colored lines), or it learns through pure distillation from a teacher trained on the same smoothed labels (black dots). The data points with error bars represent the results of numerical experiments in the second scenario at $N = 4000$ (10 samples per point).

phenomenon observed around the $\alpha = \eta$ interpolation peak. Note that, since the magnitude of the learned preactivations is effectively decreased, the saturating regime of the $\sigma(\cdot)$ activation function is avoided and this yields larger differences between the teacher outputs on different patterns, allowing for a better transfer of knowledge.

Appendix C. On the Double-descent phenomena

We have seen in section 6 the appearance of sharp interpolation peak at $\alpha_I = \eta$, when the number of parameters of the student model equals the size of the training set. Moreover, note that the peak becomes more pronounced when the teacher regularization is close to the optimal value $\tilde{\lambda} = 0.1$.

While such a location for the interpolation peak is uncharacteristic of logistic regression, a similar cusp is typically observed in the weak regularization regime when the classifier is trained with a Mean Squared Error (MSE) loss function [Hastie et al. \(2019\)](#); [Mignacco et al. \(2020\)](#). What one would expect in the case of logistic regression is instead a less pronounced peak, located in correspondence of the separability threshold $\alpha_S(\rho, \Delta)$ for the training dataset. Note that, in the mismatched distillation framework, we expect two distinct separability thresholds of this type $\alpha_S < \tilde{\alpha}_S$, one for the student and one for the teacher.

In order to understand the origin of the unusual interpolation peak at $\alpha = \eta$, in Fig. 11, we display the behavior of a series of relevant quantities: the average cross-entropy-per-pattern (top left), the student norm (top right), the average MSE-distance between teacher and student outputs (bottom left) and the average MSE between teacher and student preactivations. In all the plots

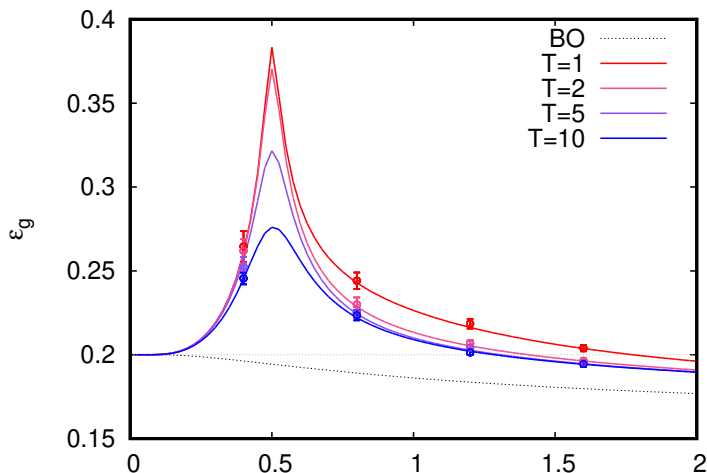


Figure 10: Generalization performance in the low α regime of a pure distillation student ($\chi = 1$) learning from a optimally regularized teacher ($\tilde{\lambda} = 0.15$) with increasing values of the distillation temperature, at $\rho = 0.2$, $\Delta = 1$, $\eta = 0.5$. (Dashed black curve) Generalization bound for the student, achieved by the sparsified plug-in estimator. The data points with error bars represent the results of numerical experiments at $N = 4000$ (10 samples per point).

we mark the three introduced thresholds, that in our parameter setting are located at $\alpha_I = 0.5$, $\alpha_S \simeq 1.75$ and $\tilde{\alpha}_S \simeq 4.75$.

In the top left plot we can see that, when the student learns from a teacher with vanishing regularization (black curve), the optimal cross-entropy remains close to zero if the dataset is separable $\alpha < \alpha_S$, and jumps to finite values otherwise (similar to what usually happens in logistic regression: indeed, the weakly regularized teacher replicates the original labels almost exactly while $\alpha < \tilde{\alpha}_S$). The associated generalization error peak (cfr. the grey curve in Fig. 2) is thus caused by the explosion of the student norm (as expected with unregularized logistic regression at the separability threshold $\alpha = \alpha_S$).

On the other hand, at finite regularizations (violet and red curves) the teacher outputs are no longer binary and the minimum achievable cross-entropy becomes strictly greater zero, as can be seen in top right plot. While the number of associated linear constraints is lower than the number of parameters $\alpha < \alpha_I$, the student is able to faithfully reproduce the non-polarized teacher outputs (bottom left plot). However, by doing so the student overfits the noisy data and the increase in its norm (top right plot) gives rise to a sharp generalization error peak. These observations seem to be consistent with the scenario observed in [Phuong and Lampert \(2019\)](#), where it was shown that below the interpolation threshold the student converges to the projection of the teacher’s weight vector onto the data span and reproduces the teacher preactivations. Note that, when the teacher is over-regularized $\tilde{\lambda} = 1$, teacher and student maximum norms are lower, partially tempering the generalization cusp.

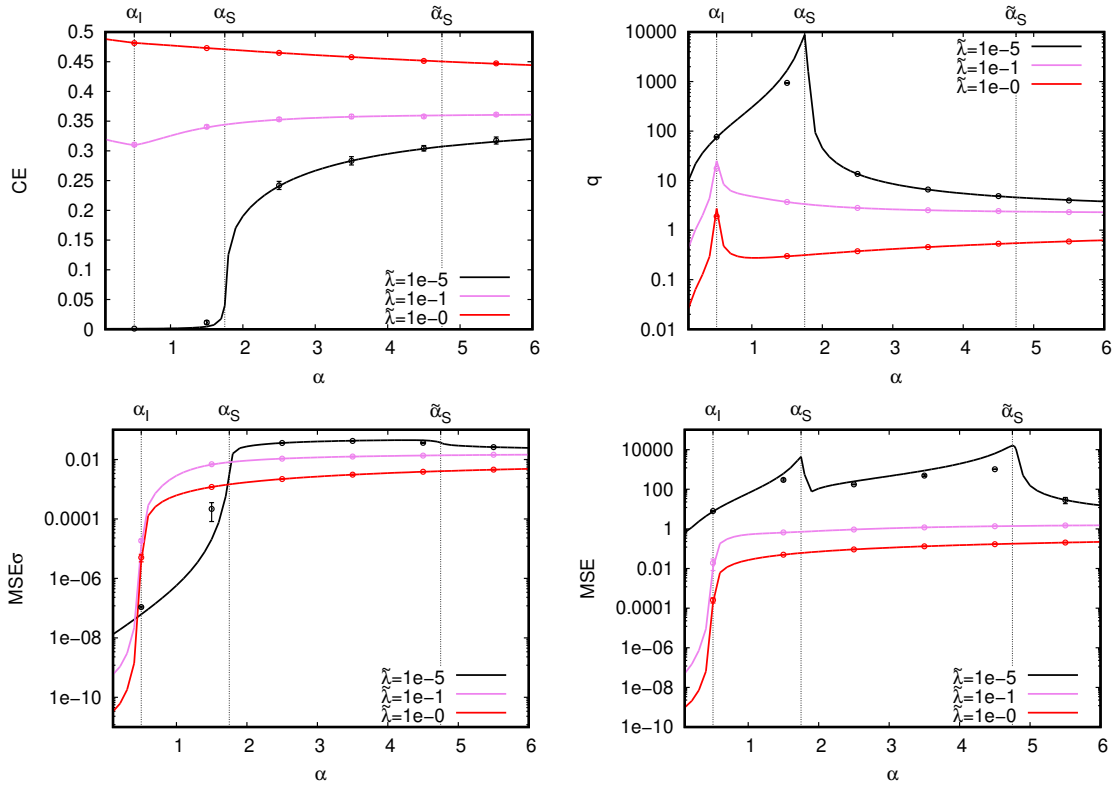


Figure 11: *Top left plot*: Typical student cross-entropy loss, averaged over the training set. *Top right plot*: Typical student norm. *Bottom left plot*: Typical MSE-distance between student and teacher outputs. *Top right plot*: Typical MSE-distance between student and teacher preactivations. All the results are collected in the pure distillation setting ($\rho = 0.2$, $\Delta = 1$, $\chi = 1$), at teacher regularizations $\tilde{\lambda} = 1e - 5$ (black), $\tilde{\lambda} = 1e - 1$ (violet), $\tilde{\lambda} = 1$ (red). The data points with error bars represent the results of numerical experiments at $N = 4000$ (10 samples per point). Due to the explosion of teacher and student norms (around the interpolation peak and the separability thresholds) in several occasions the employed numerical optimization routine (Adam optimizer [Kingma and Ba \(2015\)](#)) couldn't converge before the imposed hard cutoff of 2000 epochs. This can explain the discrepancies with the theoretical predictions.

Finally, in the bottom right plot, we see the average MSE-distance between teacher and student preactivations. When the teacher is weakly regularized (black curve) the deviation between teacher and student preactivations sharply increases around the student interpolation threshold α_S (where the student norm is greater than the teacher's) and around the teacher interpolation threshold $\tilde{\alpha}_S$ (where the teacher norms reaches its maximum). At higher regularization levels this behavior disappears, since the cross-entropy minimization no longer induces great spikes in the student norm.

Appendix D. Numerical results for the balanced case

We here report the results obtained in the pure distillation setting when the two datapoints clusters are balanced $\rho = 0.5$. As mentioned (and reported in Mignacco et al. (2020)), in this special case the ridge regularization is able to approach the Bayes optimal performance, in the $\lambda \rightarrow \infty$ limit.

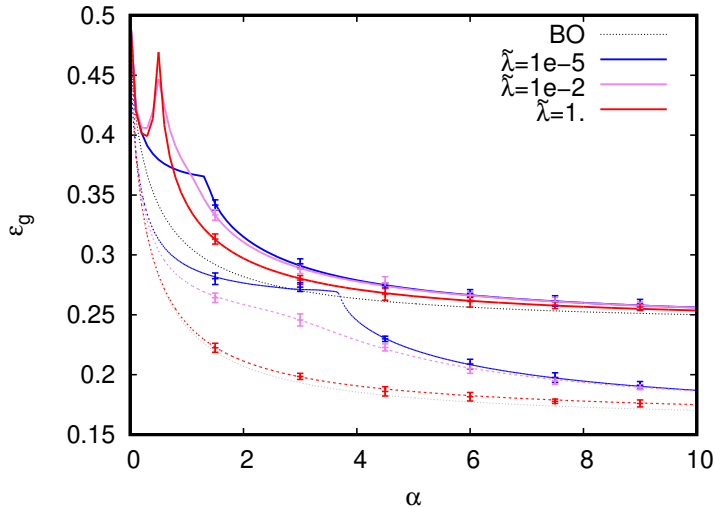


Figure 12: Balanced clusters setting ($\rho = 0.5$). (Black dashed line) Bayes optimal performance (e.g., achieved by the sparsified plug-in estimator). (Colored dashed lines) Teacher generalization performance, after training with ridge regularized logistic regression (colors indicate the regularizer intensity). (Full colored lines) Student generalization performance, in the pure distillation setting ($\lambda = 1e - 5$, $\chi = 1$, $T = 1$).

In Fig. 12, we display the teacher (dashed lines) and student (full lines) generalization performances at varying ridge regularization intensity in the teacher loss, and compare them with the Bayes optimal performance (dashed). Clearly, higher regularization induces better generalization for the teacher and this property is directly inherited also by the student, when α is above the interpolation peak at $\alpha = \eta$ (more details in the main text).