ILLUMINATING DARK KNOWLEDGE VIA RANDOM MATRIX ENSEMBLES

Anonymous authors
Paper under double-blind review

ABSTRACT

It is all but certain that machine learning models based on deep neural networks will soon feature ubiquitously in a wide variety of critical products and services that people rely on. This should be a major cause of concern given that we still lack a rigorous understanding of the failure modes of these systems, and can hardly make guarantees about the conditions under which the models are expected to work. In particular, we would like to understand how these models manage to generalize so well, even when seemingly overparametrized, effectively evading many of the intuitions expected from statistical learning theory. We argue that Distillation (Caruana et al., 2006, Hinton et al., 2014) provides us with a rich playground for understanding what enables generalization in a concrete setting. We carry out a precise high-dimensional analysis of generalization under distillation in a real-world setting, eschewing ad hoc assumptions, and instead consider models actually encountered in the wild.

1 INTRODUCTION

Statistical folklore suggests that models with considerably more parameters than required to fit the training data should lead to a significant level of overfitting. Yet deep neural networks with ever increasing capacities seem to generalize remarkably well in practice. This suggests revisiting the foundations of statistical learning theory in order to develop a deeper understanding about how these models actually work. While there's currently a rapidly growing body of numerically-grounded theoretical work contributing important insights to unraveling the puzzle, there are still many basic questions left unanswered.

We will present a compelling case arguing that understanding how generalization works in the context of Knowledge Distillation (Bucilua & Niculescu-Mizil (2006), Hinton et al. (2015)) is an ideal stepping stone for understanding generalization in the general case. We advance this program by developing a theoretical framework (based on recent developments in Random Matrix Theory) to probe the conditions required for effective knowledge transfer between models via distillation. For alternative theoretical treatments addressing generalization in the distillation context, we refer the reader to the more formal approach in (Lopez-Paz et al. (2016)) and an analysis that studies the dynamics of deep linear models under distillation (Phuong & Lampert (2019)). Our work differs from the latter in that we are able to bypass the intricacies of the effects of dynamics by considering the asymptotic high-dimensional regime appropriately scaled to mimic the conditions under which universality emerges in ensembles of random matrices of extremely large size.

For the sake of concreteness, we will focus our attention to answering the following question ...

Let $\mathfrak{T}$ denote a model with a very large number of parameters $n_{\mathfrak{T}}$, which will serve as our teacher. Similarly, $\mathfrak{S}$, playing the role of a student, denotes a model with a parameter count $n_{\mathfrak{S}} \ll n_{\mathfrak{T}}$.

Suppose that, when both models are trained on a given dataset to solve a particular supervised classification task, the teacher achieves a generalization performance $\texttt{acc}_{\mathfrak{T}}$ while the student achieves a baseline generalization performance $\texttt{acc}_{\mathfrak{S}} < \texttt{acc}_{\mathfrak{T}}$.

What is the expected generalization performance when the student is trained using a combination of ground truth labels and the teacher's knowledge via distillation?

2 THEORETICAL UNDERPINNINGS

In what follows, we gather and develop the tools necessary to answer the question as posed, and proceed to carry out a precise high-dimensional asymptotic analysis of generalization within the distillation setting. We also present experimental results showing that our theoretical approach produces results that match what one observes empirically, essentially providing a theory of distillation based on universal properties of ensembles of large random matrices. The general theory is naturally enough, very complicated, but certain special cases are found to admit a fairly complete discussion.

Originally framed as a "knowledge transfer" technique (Bucilua & Niculescu-Mizil (2006)) and "rediscovered" in Hinton et al. (2015), distillation compresses the knowledge in a large capacity model (the teacher) into a smaller student model by encouraging the student to match the predictions of the teacher in addition to the usual training signal, the idea being that the probabilities that the teacher assigns to incorrect class label contain a lot of valuable information about the way in which the teacher generalizes.

As it turns out, the idea does actually work in many cases, but the results obtained depend not only on the hyperparameters used in carrying out the distillation, but also on the dataset size, "impedance matching" between the student and teacher architectures, and the characteristics of the distribution from which the data is drawn.

We would like to develop a better understanding of the conditions under which distillation is guaranteed to work so as to determine when distillation would be beneficial, without blindly carrying out hyperparameter sweeps.

2.1 GENERALITIES

Our analytical framework is almost entirely driven by the concentration of measure phenomenon – or what is essentially equivalent – statistical mechanical arguments which hold when replica symmetry is unbroken. This phenomenon is at the core of our current theoretical understanding of high-dimensonal statistical models, spanning an impressive range of disciplines. [1]

In what follows,

- $\mathbb{Y}$ denotes the one-hot encoded ground truth labels,
- $\mathbb{P}_T$ and $\hat{\mathbb{P}}_T$ denote the softmax outputs for the student and teacher respectively, after scaling the logits by T^{-1} where T is the temperature,
- W denotes the weights in the penultimate layer of the student's model, and
- $M^\dagger$ denotes the transpose of any matrix M.

In keeping with the original prescription for knowledge distillation (Hinton et al. (2015)), we will parametrize our distillation loss as follows:

$$\mathcal{L} = \frac{\mathcal{L}_{CE}(\mathbb{P}_1|\mathbb{Y}) + \alpha(T)\mathcal{L}_{KL}(\mathbb{P}_T|\hat{\mathbb{P}}_T)}{1+\alpha(T)/T} \tag{1}$$

where $\alpha(T)$ is some "nicely behaved" function of the temperature T, and $\mathcal{L}_{CE}$/$\mathcal{L}_{KL}$ denote the cross-entropy/relative entropy functions respectively.

To recover the original formulation, one simply sets $\alpha(T) = \frac{1-\alpha_H}{\alpha_H}T^2$ where $\alpha_H \in [0,1]$ is the interpolation parameter in Hinton et al. (2015). The reason to consider a more general formulation is mainly to avoid the heuristics used in the original formulation.[2]

[1] see, e.g. [Mézard et al. (1987); Talagrand (2011); Mézard & Montanari (2009)] for a sampling of physics problems and [Montanari & Shah (2006); Panchenko (2014); Bapst & Coja-Oghlan (2016)] for a sampling of computer science applications.

[2] While the heuristics work just fine in practice, the original justification given in Hinton et al. (2015) depends strongly on the magnitude of the student's logits. More details can be found in Anonymized (2020)

Lemma 1 *The classification accuracy of any model with softmax outputs $\mathbb{P}_T$ when trained with one-hot labels $\mathbb{Y}$ is given by*

$$\text{accuracy} = \lim_{\gamma \to 0} \frac{\text{Tr}(\mathbb{P}_{\gamma T} \mathbb{Y}^\dagger)}{N_{\text{data}}} \tag{2}$$

Lemma 2 *The SGD dynamics of a student trained via distillation using the loss (1) is identical to that of a dual student trained without distillation via the following prescription:*

1. Both the distilled student and the dual model are initialized in exactly the same manner.
2. The dual student is trained using an interpolation of "hard" and "soft" labels via

$$\mathbb{Y}_{\text{dual}} = \frac{T\mathbb{Y} + \alpha(T)\hat{\mathbb{P}}_T}{T + \alpha(T)} \tag{3}$$

3. The dual student's softmax outputs are given by the interpolation

$$\mathbb{P}_{\text{dual}} = \frac{T\mathbb{P}_1 + \alpha(T)\mathbb{P}_T}{T + \alpha(T)} \tag{4}$$

The previous lemmata equip us with the tools required to cast the problem of analyzing the generalization performance of models under distillation as a problem about random matrix ensembles. More precisely,

1. $\mathbb{P}_{\gamma T}$ is a real-valued non-symmetric matrix whose elements have a distribution induced from the coupling of the distribution of the training data with the distribution of the weight parameters under SGD.

   Similarly, $\mathbb{Y}$ is a non-symmetric matrix whose elements are determined by the distribution of the labels. At this point, we have assumed nothing about these distributions.

   As a technical aside, we mention that, much as in conventional commutative statistics, it is a lot easier to deal with distributions that are centered/standardized. We thus define

$$\delta\mathbb{P}_{\gamma T} := \mathbb{P}_{\gamma T} - \langle \mathbb{P}^\dagger_{\gamma T} \rangle, \qquad \delta\mathbb{Y}_{\gamma T} := \mathbb{Y} - \langle \mathbb{Y} \rangle,$$

   where $\langle \cdot \rangle$ denotes taking expectations.
2. It follows that the matrix product $\delta\mathbb{P}_{\gamma T}\delta\mathbb{Y}^\dagger$ is also a real-valued non-symmetric random matrix matrix whose distribution is determined by the distributions of $\mathbb{P}_{\gamma T}$ and $\mathbb{Y}^\dagger$
3. From (2), the accuracy depends only on the trace of $\mathbb{P}_{\gamma T}\mathbb{Y}^\dagger$ and hence is determined uniquely by the eigenvalues of the matrix. This essentially couches the generalization problem as a problem of identifying the random matrix ensemble to which $\mathbb{P}_{\gamma T}\mathbb{Y}^\dagger$ belongs. As is usual in random matrix theory, the properties of the appropriate ensemble acquire universal characteristics in the asymptotic regime $N_{\text{data}} \to \infty$, $N_{\text{classes}} \to \infty$ with $q = N_{\text{classes}}/N_{\text{data}}$ fixed.
4. A further simplification occurs since the trace is invariant under symmetrization, allowing us to take advantage of the fact that a lot more is known about Hermitian random matrix ensembles relative to their non-Hermitian counterparts. Thus, we will henceforth work with the pair of Hermitian matrices

$$H_{\gamma T} := \frac{\delta\mathbb{Y}^\dagger\delta\mathbb{P}_{\gamma T} + \delta\mathbb{P}^\dagger_{\gamma T}\delta\mathbb{Y}}{2N_{\text{data}}} \tag{5}$$

   and

$$\hat{H}_{\gamma T} := \frac{\delta\hat{\mathbb{P}}^\dagger\delta\mathbb{P}_{\gamma T} + \delta\mathbb{P}^\dagger_{\gamma T}\delta\hat{\mathbb{P}}}{2N_{\text{data}}} \tag{6}$$

   so that the expression for the generalization performance reads

$$\text{accuracy} = \lim_{\gamma \to 0} H_{\gamma T} + \text{Tr}\left[\frac{\langle \delta \mathbb{Y} \rangle \langle \delta \mathbb{P}^{\dagger}_{\gamma T} \rangle}{N_{\text{data}}}\right] \tag{7}$$

Proposition 3 *For any dataset whose labels are uniformly distributed across classes, we have that*

$$\text{Tr}\left[\frac{\langle \delta \mathbb{Y} \rangle \langle \delta \mathbb{P}^{\dagger}_{\gamma T} \rangle}{N_{\text{data}}}\right] = \frac{1}{N_{\text{classes}}} \tag{8}$$

Thus, the accuracy (7) reads

$$\text{accuracy} = \lim_{\gamma \to 0} H_{\gamma T} + \frac{1}{N_{\text{classes}}} \tag{9}$$

Proposition 4 *Under distillation, the generalization performance of the distilled student is obtained from the accuracy of the dual model (2, 3, 4) via*

$$\text{accuracy} = \lim_{\gamma \to 0} \frac{\text{Tr}(\mathbb{P}^{(\text{dual})}_{\gamma T} \mathbb{Y}^{\dagger}_{\text{dual}})}{N_{\text{data}}} \tag{10}$$

Using the definitions in (5, 6), and writing $\mathbf{1}$ for the symmetric matrix whose entries are all unity, a direct computation yields

$$\frac{\mathbb{P}^{(\text{dual})}_{\gamma T} \mathbb{Y}^{\dagger}_{\text{dual}}}{N_{\text{data}}} \quad = \quad \frac{\mathbf{1}}{N_{\text{classes}}} + \frac{T^2 H_{\gamma} + \alpha(T)(H_{\gamma T} + \hat{H}_{\gamma}) + \alpha(T)^2 \hat{H}_{\gamma T}}{[T + \alpha(T)]^2} \tag{11}$$

Once we identify the random matrix ensembles to which the individual summands in (11) belong, we can estimate the *expected accuracy* of the dual model as precisely as we would like, and consequently the expected accuracy of the distilled student.

In keeping with the folk wisdom in random matrix theory, any estimator based on (11) will yield very precise results in the regime where one has a large number of classes and a large number of training samples. Empiricallly, we found that the estimator works best for students with fully connected layers trained on whitened inputs, and that very deep models with residual connections require accounting for corrections to the dynamics which can be done, but requires substantially more technical effort.

2.2 Drilling Down

So far, our analysis has been completely general and we can hardly identify the relevant random matrix ensemble without additional constraints/information. It should be obvious that this has to be the case, otherwise one would be able to predict the generalization performance of a neural network *a priori*, prior to training, rendering supervised learning a solved problem. That would indeed be a coup.

As stated in the introduction, the only reason why we are able to make some progress in understanding generalization under distillation is because we have access to two additional pieces of information:

1. the baseline performance of the student when trained only on the "hard" labels $\mathbb{Y}$, and
2. the performance of the teacher when trained only on the "hard" labels.

These two additional pieces of information make the problem tractable in the following sense: since we have access to the student's performance without distillation, we can run a large number of experiments to generate samples of H_{γ}, diagonalize the samples to obtain their corresponding

eigenvalues, and use these observations to design a random matrix ensemble that reproduces the observed statistics. We can also do the same for the teacher, to obtain an estimate of samples of $\hat{H}_{\gamma T}$.

There is overwhelming empirical evidence that, after training, the covariance matrix of the softmax outputs, *viz.* $\mathbb{P}_\gamma^\dagger \mathbb{P}_\gamma / N_{\rm data}$ has diagonals that are essentially identically distributed and at least an order of maginitude larger in size relative to the off-diagonals. This is not unexpected since it is known, theoretically and empirically, that deep linear networks trained on iid inputs from an arbitrary distribution have precisely this property (Ndirango & Lee (2019)). Whenever this "concentration phenomenon" occurs, we can precisely idenfity the ensemble to which $\mathbb{P}_{\gamma T}$ belongs thanks to the beautiful body of work around the Random Energy Model initiated in (Derrida (1981))and made mathematically rigorous in the tour de force (Talagrand (2011)). In the latter, we find the following gem:

Proposition 5 *[Section 13.1 of Talagrand (2011)] If the covariance matrix $\mathbb{P}_\gamma^\dagger \mathbb{P}_\gamma / N_{\rm data}$ has dominant iid diagonal elements, then as $\gamma \to 0$, the distribution of the entries of $\mathbb{P}_\gamma$ converges to the Poisson-Dirichlet distribution with parameter γ*

Lemma 6 *The one-hot encoded labels $\mathbb{Y}$ obey the identity* $\lim_{\gamma\to 0} \texttt{softmax}(\mathbb{Y}/\gamma) = \mathbb{Y}$. [3]

Putting the last two statements together identifies the appropriate ensemble as one associated with the so-called Free-Bessel laws (Banica et al. (2011)) which are intimately tied to the Wishart ensemble.

Proposition 7 *[Essentially Bouchaud & Potters (2020)] If $\mathbb{P}_\gamma$ and $\mathbb{Y}$ are random matrices with iid entries and have covariances proportional to the identity matrix, then the matrices H_γ as defined above belong to a generalization of the Wishart ensemble. Furthermore, the spectral density of this ensemble can be explicitly calculated to any desired degree of accuracy.*

For our purposes, the one-hot labels $\mathbb{Y}$ trivially have a covariance proportional to the identity matrix whenever the labels are equally represented among the classes to be categorized. While the matrices $\mathbb{P}_\gamma$ do not have covariances proportional to the identity, their covariances are very precisely estimated by

$$\frac{\mathbb{P}_\gamma^\dagger \mathbb{P}_\gamma}{N_{\rm data}} \simeq {\rm Tr}\left[\frac{\mathbb{P}_\gamma^\dagger \mathbb{P}_\gamma}{N_{\rm data}}\right]\left(\mathbb{I} - \frac{1}{N_{\rm classes}}\mathbf{1}\right)$$

as demonstrated in (Ndirango & Lee (2019)). For the purposes of constructing an estimator, this is more than adequate since the matrix $\mathbb{I} - \frac{1}{N_{\rm classes}}\mathbf{1}$ has a spectrum that matches the identity matrix save one zero eigenvalue.

We finally have all the pieces needed to construct a model for $\lim_{\gamma\to 0} H_\gamma$.

Proposition 8

(i) Under the hypothesis that SGD dynamics does not introduce long range correlations between the parameters of the model whenever the model's parameters are initialized randomly, the matrices $\lim_{\gamma\to 0} H_\gamma$ belong to the generalized Wishart ensemble, corresponding to the cross-covariance matrix associated with matrices whose elements are independently drawn from the Poisson-Dirichlet distribution.

(ii) The corresponding generalized Marcenko-Pastur law is determined by

$$\lim_{\gamma\to 0}\langle {\rm Tr}\, H_\gamma\rangle = \overline{\rm accuracy}_{\rm baseline} - \frac{1}{N_{\rm classes}}$$

where $\overline{\rm accuracy}_{\rm baseline}$ is obtained by training an ensemble of student models and taking the average of their generalization performance.

[3] Proof: direct computation

An example of the theoretical prediction is shown in Figure 1.

As shown in the appendices, by exploiting results from free probability theory, (Nica & Speicher (1998))[4] we arrive at our formula for generalization under distillation:

$$\frac{\overline{\text{accuracy}}_{\text{distilled}} - 1/N_{\text{classes}}}{\overline{\text{accuracy}}_{\text{baseline}} - 1/N_{\text{classes}}} = T\frac{[1+\alpha(T)g(a_-,a_+,T)][1+\alpha(T)\chi]}{[T+\alpha(T)]^2} \tag{12}$$

where the quantities $a_\pm$ denote the upper and lower bounds for the support of the eigenvalue distribution of the generalized Wishart ensemble, g is a function of the support and the tempertature, and χ is a factor that measures the performance of the student when trained using only hard labels, versus how the student would perform if trained only on soft labels from the teacher. All of these quantities are entirely independent of the distillation process.

3 NUMERICS

The purpose of this section is to establish that some kind of universality emerges naturally as a bona fide phenomenon under SGD dynamics, and not just an artifact, suggested, however strongly, by the modeling assumptions introduced above to derive our results. The necessary Python code to replicate all the following experiments will be made available at `http://github.com/to_be_revealed_when_allowed`.

The first set of results establishes the validity of our model for the asymptotic eigenvalue distribution for the matrix $\lim_{\gamma\to 0} H_\gamma$.

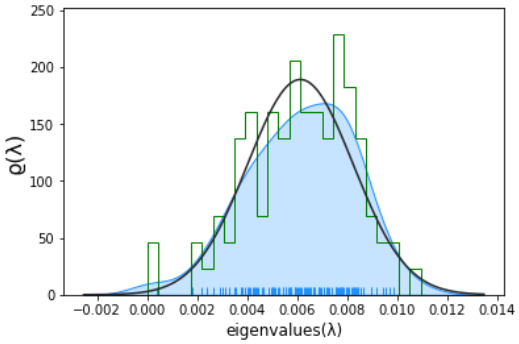

Figure 1: Theoretical prediction (black curve) for the distribution of eigenvalues of H_γ vs. the observed historgram for a resnet8 student trained on CIFAR 100. The shaded region is the result of a kernel density estimator. Note that the theoretical curve is symmetric. This is an artifact of the small correlation coefficient that occurs in this particular problem.

The second set of results establish the validity of the expression (12). To obtain the empirical results, we trained an ensemble of resnet8 students distilled from a resnet18 teacher for a sampling of choices of T and the function $\alpha(T)$. These were then checked against the theoretical predictions from (12), and we observe excellent agreement between theory and experiment.

[4] which, in the crudest terms is a generalization of probability theory to the non-commutative setting. In other words, non-commutative "free" random variables are the analog of independent random variables in the commutative case.

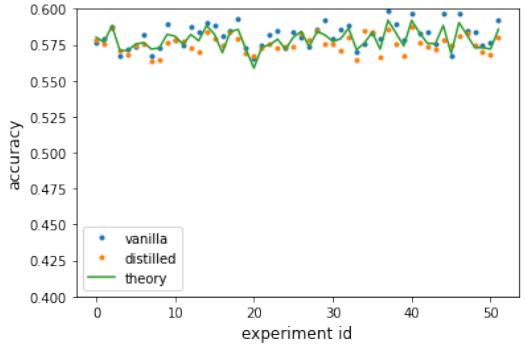

Figure 2: The generalization accuracy of a resnet8 student ($\sim$89k parameters) trained on CIFAR 100 in the "vanilla" case and distilled from a resnet18 (1.1M parameters). The teacher's baseline performance on CIFAR100 is 0.62. The figure shows empirical results of 50 experimental runs with various choices for T and $\alpha(T)$ superimposed against the predictions from the calculations detailed above.

REFERENCES

Authors Anonymized. Softmax classifiers as generalized rems. *eprint*, 2020.

T. Banica, S. T. Belinschi, M. Capitaine, and B. Collins. Free bessel laws. *Canadian Journal of Mathematics*, 63(1):3–37, 2011.

V. Bapst and A. Coja-Oghlan. Harnessing the Bethe free energy. *Random Structures & Algorithms*, 49(4):694–741, 2016.

Jean-Philippe Bouchaud and Marc Potters. On a generalisation of the marcenko-pastur problem. *https://arxiv.org/abs/2009.0711*, 2020.

Caruana Rich Bucilua, Cristian and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pp. 535–541. Association for Computing Machinery, 2006.

Bernard Derrida. Random-energy model: An exactly solvable model of disordered systems. *Phys. Rev. B*, 24:2613–2626, Sep 1981. doi: 10.1103/PhysRevB.24.2613. URL https://link.aps.org/doi/10.1103/PhysRevB.24.2613.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *https://arxiv.org/abs/1503.02531*, 2015.

D. Lopez-Paz, B. Schölkopf, L. Bottou, and V. Vapnik. Unifying distillation and privileged information. In *International Conference on Learning Representations (ICLR)*, November 2016.

M. Mézard and A. Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

M. Mézard, G. Parisi, and M. Virasoro. *Spin glass theory and beyond: an introduction to the replica method and its applications*, volume 9. World Scientific Publishing Company, 1987.

A. Montanari and D. Shah. Counting good truth assignments of random k-sat formulae. *arXiv preprint cs/0607073*, 2006.

Anthony Ndirango and Tyler Lee. Generalization in multitask deep neural classifiers: a statistical physics approach. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 32*, pp. 15862–15871. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/9715-generalization-in-multitask-deep-neural-classifiers-a-statistical-physics-app pdf.

A. Nica Nica and R. Speicher. Commutators of free random variables. In *C Duke Math. J. 92, no. 3, 553–592, 1998*, 1998.

D. Panchenko. On the replica symmetric solution of the k-sat model. *Electronic Journal of Probability*, 19, 2014.

Mary Phuong and Christoph Lampert. Towards understanding knowledge distillation. volume 97 of *Proceedings of Machine Learning Research*, pp. 5142–5151, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/phuong19a.html.

Michel Talagrand. *Mean Field Models for Spin Glasses: Volumes I and II*, volume 15. Springer, 2011.

A Appendix