A Statistical Perspective on Distillation

Aditya Krishna Menon [1] **Ankit Singh Rawat** [1] **Sashank J. Reddi** [1] **Seungyeon Kim** [1] **Sanjiv Kumar** [1]

Abstract

Knowledge distillation is a technique for improving a "student" model by replacing its one-hot training labels with a label *distribution* obtained from a "teacher" model. Despite its broad success, several basic questions — e.g., Why does distillation help? Why do more accurate teachers not necessarily distill better? — have received limited formal study. In this paper, we present a statistical perspective on distillation which sheds light on these questions. Our core observation is that a "Bayes teacher" providing the true class-probabilities can *lower the variance* of the student objective, and thus improve performance. We then establish a *bias-variance* tradeoff that quantifies the utility of teachers that approximate the Bayes class-probabilities. This provides a formal criterion as to what constitutes a "good" teacher, namely, the quality of its probability estimates. Finally, we illustrate how our statistical perspective facilitates novel applications of distillation to bipartite ranking and multiclass retrieval.

1. Introduction

Distillation is the process of using a "teacher" model to improve the performance of a "student" model (Bucilă et al., 2006; Ba & Caruana, 2014; Hinton et al., 2015). In its simplest form, one trains the student to fit the teacher's *soft distribution* over labels, rather than one-hot training labels. While originally devised as a means of model compression, distillation has proven successful in improving students with the *same* architecture as the teacher (Rusu et al., 2016; Furlanello et al., 2018), and found several broader uses (Papernot et al., 2016; Yim et al., 2017; Liu et al., 2019).

Given its empirical success, it is natural to ask: why does distillation help? Answering this requires confronting a number of puzzling empirical observations, e.g., that improving teacher accuracy can *harm* distillation performance (Müller et al., 2019). One commonly accepted intuition from Hinton et al. (2015) is that the teacher's soft labels provide "dark knowledge" via weights on the "wrong" labels $y' \neq y$ for an example (x, y). But what are the formal statistical benefits of using soft over one-hot labels, and what is the "ideal" set of soft labels? While several works have studied various aspects of distillation (Lopez-Paz et al., 2016; Phuong & Lampert, 2019; Mobahi et al., 2020; Ji & Zhu, 2020; Zhang & Sabuncu, 2020; Zhou et al., 2021; Hsu et al., 2021) (cf. §2.2), precise answers to these questions remain elusive.

In this paper, we present a statistical perspective on distillation which sheds light on why it can aid performance, explicate the statistical value of "dark knowledge", and provide a formal criterion as to what constitutes a "good" teacher. Our key observation is that a teacher providing the *true (Bayes) class-probabilities* can *lower the variance* of the student objective, and thus improve performance. Further, teachers that reliably approximate these probabilities — for which merely being accurate does *not* suffice — possess a *bias-variance* tradeoff quantifying how they may improve generalisation. Beyond providing conceptual insight, this perspective facilitates novel applications of distillation to problems such as bipartite ranking and multiclass retrieval.

In sum, our contributions are:

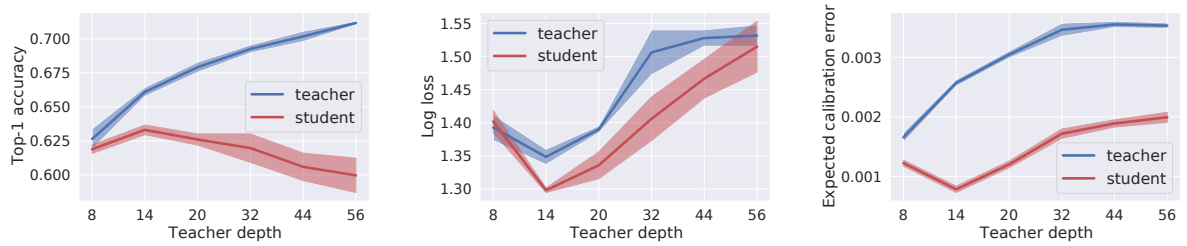
(i) We present a statistical view of distillation, by establishing that the student's expected loss inherently smooths labels with the *Bayes class-probabilities* (§3). We then quantify the benefit of using these Bayes probabilities in place of one-hot labels. This gives a statistical perspective on "dark knowledge": for an example (x, y), the logits on "wrong" labels $y' \neq y$ encode information about the underlying class distribution. This helps the student minimise a better approximation to the true generalisation error, which can improve performance.

(ii) We quantify a *bias-variance* tradeoff for teachers providing an approximation to the Bayes probabilities (§4). This gives a concrete criterion for assessing if a teacher is "good", i.e., the quality of its *probability estimates* as estimated, e.g., by the log-loss. This does *not* necessarily correspond to a more accurate teacher; see Figure 1.

(iii) We illustrate how our statistical perspective facilitates novel practical applications of distillation, e.g., to bipartite ranking and multiclass retrieval problems (§5).

[1]Google Research, New York. Correspondence to: Aditya Krishna Menon <adityakmenon@google.com>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.



(a) Top-1 accuracy. (b) Log-loss. (c) Expected calibration error.

Figure 1. Illustration of how teachers with better test set class-probability estimates generally yield better students, in keeping with our bias-variance bound (Proposition 3). On CIFAR-100, we train teachers that are ResNets of varying depths, and distill these to a student ResNet of fixed depth 8. Figure 1(a) reveals that as its depth increases, the teacher gets increasingly more accurate on the test set. *However*, the teacher's probability estimates on the test set become progressively poorer approximations of the Bayes class-probabilities $\mathbf{p}^*(x)$ beyond depth 14: the teacher's test set log-loss and calibration error (Guo et al., 2017) (both measures of the quality of the probability estimates) increase beyond this depth (Figure 1(b) and 1(c)). Intuitively, the depth 14 provides an optimal balance between the bias and variance in the teacher's predictions. In line with Proposition 3, the depth 14 teacher with the best probability estimates produces the most accurate student (Figure 1(a)). See §4.3 for more details, and §4.1 for the formal bias-variance tradeoff.

Our findings are verified for linear models, neural networks, and decision trees, on both controlled synthetic and real-world datasets. This illustrates a broader point of our statistical perspective: distillation can be understood as a basic tool which has utility *not* just limited to neural networks.

More broadly, our simple measure of a teacher's utility for distillation — the quality of its probability estimates, which may be estimated by the test set log-loss, square-loss, or calibration error — gives one means of reasoning about various empirical findings; e.g., temperature scaling can be seen as improving the teacher's probability calibration (cf. §4.2), while teachers that are merely *accurate* but provide poor probability estimates may distill poorly (Figure 1, 3).

2. Background and Notation

2.1. Multiclass Classification

In multiclass classification, we are given a training sample $S \doteq \{(x_n, y_n)\}_{n=1}^N \sim \mathbb{P}^N$, for distribution $\mathbb{P}$ over instances $\mathcal{X}$ and labels $\mathcal{Y} = [L] \doteq \{1, 2, \ldots, L\}$. Our goal is to learn a predictor $\mathbf{f} \colon \mathcal{X} \to \mathbb{R}^L$ with minimal *risk*:

$$R(\mathbf{f}) \doteq \mathop{\mathbb{E}}_{(x,y)\sim\mathbb{P}} \big[\ell(y, \mathbf{f}(x))\big]. \tag{1}$$

Here, $\ell(y, \mathbf{f}(x))$ is the loss of predicting $\mathbf{f}(x) \in \mathbb{R}^L$ when the true label is $y \in [L]$. A canonical example is the softmax cross-entropy, or the log-loss with softmax activation:

$$\ell(y, \mathbf{f}(x)) = -f_y(x) + \log\Big[\sum\nolimits_{y' \in [L]} e^{f_{y'}(x)}\Big]. \tag{2}$$

We may approximate the risk $R(\mathbf{f})$ via the *empirical risk*

$$\hat{R}(\mathbf{f}; S) \doteq \frac{1}{N} \sum\nolimits_{n \in [N]} \mathbf{e}_{y_n}^\top \ell(\mathbf{f}(x_n)), \tag{3}$$

for one-hot encoding $\mathbf{e}_y \in \{0,1\}^L$ of a given label $y \in [L]$, and vector of losses for each possible label $\ell(\mathbf{f}(x)) \doteq \big(\ell(1, \mathbf{f}(x)), \ldots, \ell(L, \mathbf{f}(x))\big) \in \mathbb{R}_+^L$.

2.2. Knowledge Distillation

Distillation involves using a "teacher" model to improve a "student" model. Classically, this may be done via matching the two models' logits (Bucilă et al., 2006). We follow the generalisation in Hinton et al. (2015), wherein one computes *teacher class-probability estimates* $\mathbf{p}^{\mathrm{t}}(x) \doteq [p^{\mathrm{t}}(y|x)]_{y \in [L]}$, where $p^{\mathrm{t}}(y|x)$ estimates how likely x is to be classified as y. These are used by a student model which replaces the empirical risk (3) with the *distilled risk*

$$\tilde{R}(\mathbf{f}; S) = \frac{1}{N} \sum\nolimits_{n \in [N]} \mathbf{p}^{\mathrm{t}}(x_n)^\top \ell(\mathbf{f}(x_n)), \tag{4}$$

so that the one-hot encoding of labels is replaced with the teacher's distribution over labels. To construct (4), we simply require $\mathbf{p}^{\mathrm{t}}$ be a valid label distribution. For a neural network teacher trained to minimise the softmax cross-entropy, $\mathbf{p}^{\mathrm{t}}$ can be the softmax of the *temperature scaled* teacher logits (Hinton et al., 2015, Equation 2). This recovers logit matching as a special case (Hinton et al., 2015). Extensions of this setup, such as matching intermediate layers, have also been considered (Romero et al., 2015; Zagoruyko & Komodakis, 2017; Kim et al., 2018; Jain et al., 2020).

2.3. Why Does Distillation Help?

While it is well-accepted that distillation is empirically valuable, there is less consensus as to *why* this is the case. Hinton et al. (2015) attributed its success to the encoding of "dark knowledge" in the probabilities the teacher assigns to the "wrong" labels $y' \neq y$ for example (x, y). This plausibly

aids the student by weighting samples differently (Furlanello et al., 2018; Tang et al., 2020). However, formalisations of this intuition are limited. One elegant exception is Lopez-Paz et al. (2016), who showed that distillation can be helpful *assuming* soft labels speed up learning; however, this does not establish when and why this can be the case.

The impact of distillation on the optimisation of the student model has been explored for specific model classes (Phuong & Lampert, 2019; Rahbar et al., 2020; Ji & Zhu, 2020). Distillation may also be seen as a data-dependent *regulariser* on the student model (Dong et al., 2019; Yuan et al., 2020), as explicated by Mobahi et al. (2020); Zhang & Sabuncu (2020) for *self-distillation* (wherein the student and teacher have the same model architecture). Foster et al. (2019) provided a generalisation bound for students constrained to learn a model close to the teacher. Such works do not explicate what constitutes an "ideal" teacher, nor quantify how an approximation to this ideal teacher affects generalisation.

Recently, Zhou et al. (2021) provided a bias-variance perspective on distillation, which is similar in spirit to the analysis of §4. However, they did not quantify the variance-reduction benefits of employing the Bayes probabilities (cf. Lemma 1), formalise how these benefits translate to approximate Bayes probabilities (cf. Proposition 4), nor provide broader examples of the value of distillation (cf. §5 and Appendix B) that exploit this view.

3. Distillation: a Class-Probability View

We now present a statistical perspective on distillation, which gives insight into why it can aid generalisation. Central to our perspective are two observations:

(i) the risk in (1) we seek to minimise inherently smooths labels by the class-probabilities $\mathbb{P}(y \mid x)$; and,
(ii) such smoothing yields a *lower variance* objective compared to using one-hot labels $\mathbf{e}_y$.

3.1. Bayes Knows Best: Distilling Class-Probabilities

We begin with the following elementary observation: the population risk $R(\mathbf{f})$ for $\mathbf{f} \colon \mathcal{X} \to \mathbb{R}^L$ is

$$R(\mathbf{f}) = \mathop{\mathbb{E}}_{x}\Big[\mathop{\mathbb{E}}_{y|x}\big[\ell(y, \mathbf{f}(x))\big]\Big] = \mathop{\mathbb{E}}_{x}\big[\mathbf{p}^*(x)^\top \ell(\mathbf{f}(x))\big], \quad (5)$$

where $\mathbf{p}^*(x) \doteq \big[\mathbb{P}(y|x)\big]_{y\in[L]}$ is the *Bayes class-probability distribution* over the labels. Intuitively, $\mathbb{P}(y|x)$ is the suitability of y for x: when $\mathbf{p}^*(x)$ is not concentrated on a single label, there is an *inherent confusion* amongst the labels. The risk involves drawing $x \sim \mathbb{P}(x)$, and averaging the loss of $\mathbf{f}(x)$ over all $y \in [L]$, weighted by $\mathbb{P}(y \mid x)$.

Given an $(x_n, y_n) \sim \mathbb{P}$, the empirical risk (3) approximates $\mathbf{p}^*(x_n)$ with the one-hot $\mathbf{e}_{y_n}$, which is only supported on one label. While $\mathbf{e}_{y_n}$ is an unbiased estimate of $\mathbf{p}^*(x_n)$, it is a significant reduction in granularity. By contrast, consider the following *Bayes-distilled risk* on a sample $S \sim \mathbb{P}^N$:

$$\hat{R}_*(\mathbf{f}; S) \doteq \frac{1}{N}\sum\nolimits_{n=1}^{N} \mathbf{p}^*(x_n)^\top \ell(\mathbf{f}(x_n)). \quad (6)$$

This is a distillation objective (cf. (4)) using a *Bayes teacher*, which provides the student with the true class-probabilities. Rather than fitting to a single label realisation $y_n \sim \mathrm{Discrete}(\mathbf{p}^*(x_n))$, a student minimising (6) considers all *alternate* label realisations, weighted by their likelihood. When ℓ is the log-loss of (2), (6) is the KL divergence between $\mathbf{p}^*$ and $\mathrm{softmax}(\mathbf{f}(x_n))$ plus a constant.

Both the standard empirical risk $\hat{R}(\mathbf{f}; S)$ in (3) and Bayes-distilled risk $\hat{R}_*(\mathbf{f}; S)$ in (6) are unbiased estimates of the population risk $R(\mathbf{f})$ in (1). Intuitively, however, we expect that a student minimising (6) ought to generalise better. We can make this intuition precise in the following.

Lemma 1. *Let $\mathbb{V}$ denote the variance of a random variable. For any fixed predictor $\mathbf{f} \colon \mathcal{X} \to \mathbb{R}^L$,*

$$\mathbb{V}_{S\sim\mathbb{P}^N}\big[\hat{R}_*(\mathbf{f}; S)\big] \leq \mathbb{V}_{S\sim\mathbb{P}^N}\big[\hat{R}(\mathbf{f}; S)\big],$$

where equality holds iff $\forall x \in \mathcal{X}$, the loss values $\ell(\mathbf{f}(x))$ are constant on the support of $\mathbf{p}^*(x)$, i.e.,

$$(\forall x \in \mathcal{X})\,(\forall y, y' \in \mathrm{supp}(p^*(x)))\,\ell(y, f(x)) = \ell(y', f(x)).$$

Proof of Lemma 1. By definition,

$$\begin{aligned}\mathrm{LHS} &= \frac{1}{N}\cdot\mathbb{V}\Big[\mathbf{p}^*(x)^\top \ell(\mathbf{f}(x))\Big] \\ &= \frac{1}{N}\cdot\mathbb{V}\left[\mathop{\mathbb{E}}_{y|x}\big[\ell(y, \mathbf{f}(x))\big]\right] \\ &= \frac{1}{N}\cdot\mathop{\mathbb{E}}_{x}\left[\mathop{\mathbb{E}}_{y|x}\big[\ell(y, \mathbf{f}(x))\big]\right]^2 - \frac{1}{N}\cdot\left[\mathop{\mathbb{E}}_{x}\mathop{\mathbb{E}}_{y|x}\big[\ell(y, \mathbf{f}(x))\big]\right]^2.\end{aligned}$$

$$\begin{aligned}\mathrm{RHS} &= \frac{1}{N}\cdot\mathbb{V}\left[\ell(y, \mathbf{f}(x))\right] \\ &= \frac{1}{N}\cdot\mathop{\mathbb{E}}_{x}\mathop{\mathbb{E}}_{y|x}\big[\ell(y, \mathbf{f}(x))^2\big] - \frac{1}{N}\cdot\left[\mathop{\mathbb{E}}_{x}\mathop{\mathbb{E}}_{y|x}\big[\ell(y, \mathbf{f}(x))\big]\right]^2.\end{aligned}$$

In both cases, the second term simply equals $R(\mathbf{f})^2$ since both estimates are unbiased. For fixed $x \in \mathcal{X}$, the result follows by Jensen's inequality applied to the random variable $Z(x) \doteq \ell(y, \mathbf{f}(x))$. Equality occurs iff each $Z(x)$ is constant, which requires ℓ to be constant on $\mathrm{supp}(\mathbf{p}^*(x))$. □

The condition on equality of variance is intuitive: the two risks trivially agree when, $\forall x$, f is non-discriminative (attaining equal loss on all labels), or when a label is inherently deterministic ($\mathbf{p}^*(x)$ is concentrated on one label). For discriminative predictors and non-deterministic labels, however, the Bayes-distilled risk can have much *lower variance*.

The reward of reducing variance is better generalisation: a student minimising (6) better minimises the population risk (1) compared to using one-hot labels. Leveraging Maurer & Pontil (2009), we may quantify how the Bayes-distilled loss' *empirical* variance influences generalisation.

Proposition 2. *Pick any bounded loss ℓ.* [1] *Fix a hypothesis class $\mathcal{F}$ of predictors $\mathbf{f}\colon \mathcal{X} \to \mathbb{R}^L$, with induced class $\mathcal{H}^* \subset [0,1]^{\mathcal{X}}$ of functions $h(x) \doteq \mathbf{p}^*(x)^\top \ell(\mathbf{f}(x))$. Suppose $\mathcal{H}^*$ has uniform covering number $\mathcal{N}_\infty$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$ over $S \sim \mathbb{P}^N$,*

$$R(\mathbf{f}) \leq \hat{R}_*(\mathbf{f}; S) + \mathcal{O}\Big(\sqrt{\mathbb{V}^*_N(\mathbf{f})/N} \cdot \sqrt{\log\left(\mathcal{M}^*_N/\delta\right)} + \log\left(\mathcal{M}^*_N/\delta\right)/N\Big),$$

where $\mathcal{M}^*_N \doteq \mathcal{N}_\infty(\frac{1}{N}, \mathcal{H}^*, 2N)$ and $\mathbb{V}^*_N(\mathbf{f})$ is the empirical variance of $\{\mathbf{p}^*(x_n)^\top \ell(\mathbf{f}(x_n))\}_{n=1}^N$.

Proof of Proposition 2. This a simple consequence of Maurer & Pontil (2009, Theorem 6), which is a uniform convergence version of Bennet's inequality (Bennett, 1962). □

By contrast, the bound achievable for the standard empirical risk using one-hot labels will depend on the variance of the one-hot loss values (Maurer & Pontil, 2009, Theorem 6). Combined with Lemma 1, the Bayes-distilled empirical risk results in a lower variance penalty. Thus, the Bayes-distilled risk yields a generalisation bound with a more favourable rate of convergence with increased sample size N. For a more refined generalisation analysis accounting for teacher under- and over-fitting, see Dao et al. (2021) which builds on an earlier version of this work.

To summarise the above, a student should ideally have access to the underlying class-probabilities $\mathbf{p}^*(x)$, rather than a single realisation $y \sim \text{Discrete}(\mathbf{p}^*(x))$: these probabilities result in a lower-variance student objective, which aids generalisation. This provides a statistical perspective on the value of "dark knowledge": for the "Bayes teacher" $\mathbf{p}^*$, the logits on alternate labels $y' \neq y$ for an example (x, y) provide information about the Bayes class-probabilities.

3.2. Illustration: the Value of a Bayes Teacher

We illustrate our statistical perspective in a controlled setting where the Bayes $\mathbf{p}^*(x)$ is known, and show that distilling with such a "Bayes teacher" benefits learning. We generate a (binary) labelled training sample $S = \{(x_n, y_n)\}_{i=1}^N$ from a distribution $\mathbb{P}$ comprising 10-dimensional isotropic Gaussian class-conditionals with means $\pm(0.25, 0.25, \ldots, 0.25)$. We may explicitly compute the Bayes $\mathbf{p}^*(x) = (\mathbb{P}(y = 0 \mid x), \mathbb{P}(y = 1 \mid x))$ as $\mathbb{P}(y = 1 \mid x) = \sigma((\theta^*)^\top x)$, for $\theta^* \doteq (0.5, 0.5, \ldots, 0.5)$ and sigmoid function $\sigma(\cdot)$.

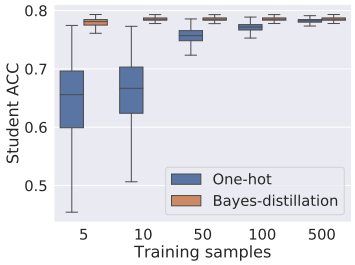


Figure 2. Illustration that distilling from a "Bayes teacher" aids generalisation. We consider learning a linear model student on a synthetic problem with known Bayes probabilities $\mathbf{p}^*$. Distilling using these probabilities significantly improves accuracy over training with one-hot labels, particularly in the small-sample regime.

We compare standard logistic regression on S and *Bayes-distilled* logistic regression using $\mathbf{p}^*(x_n)$ per (6). Figure 2 compares these approaches for varying training set sizes N, where for each N we perform 100 independent trials and measure accuracy on a test set of 10^4 samples. For small N, Bayes-distillation offers a noticeable gain over one-hot training, in line with our theory of the former ensuring lower variance (Lemma 1). Both methods see improved performance with larger N, but one-hot encoding has greater gains: thus, the variance reduction offered by Bayes-distillation can compensate for having only a few student samples. For additional experiments, see Appendix C.3.

4. Distilling from an Imperfect Teacher

The previous section explicates how an idealised "Bayes teacher" can benefit a student. We now study how this translates to the more realistic setting of using an imperfect, "approximate Bayes teacher" learned from data.

4.1. A Bias-Variance Bound for Distillation

Our first observation is that a teacher's predictor $\mathbf{p}^{\mathrm{t}}$ is typically an *imperfect estimate* of the true $\mathbf{p}^*$. For example, a teacher minimising softmax cross-entropy effectively minimises $\mathbb{E}[\mathrm{KL}(\mathbf{p}^*(x) \| \mathbf{p}^{\mathrm{t}}(x))]$. Of course, in practice a teacher is unlikely to learn $\mathbf{p}^{\mathrm{t}} = \mathbf{p}^*$, as its model class may not be rich enough to capture the true $\mathbf{p}^*$. Further, even if the teacher can represent $\mathbf{p}^*$, it may not be able to *learn* this perfectly given a finite sample, owing to both statistical (e.g., overfitting) and optimisation (e.g., non-convexity) issues.

Will such an imperfect estimate of $\mathbf{p}^*$ still improve generalisation? To answer this, we establish a *bias-variance* tradeoff for distillation. Specifically, we show the difference between the distilled risk $\tilde{R}(\mathbf{f}; S)$ (cf. (4)) and population risk $R(\mathbf{f})$ (cf. (5)) depends on how well the teacher esti-

[1] Note that the boundedness assumption on the loss is standard (Boucheron et al., 2005, Theorem 4.1), and may be enforced in practice with some form of regularisation (e.g., weight decay).

mates the Bayes class-probability distribution $\mathbf{p}^*$ in a mean squared-error (MSE) sense. This, in turn, admits a classic bias-variance decomposition for the teacher estimates. (In the following, for simplicity we assume that the student and teacher are trained on separate samples.)

Proposition 3. *Pick any bounded loss ℓ. Suppose we have a teacher model $\mathbf{p}^{\mathrm{t}}$ with corresponding distilled empirical risk in (4). For any predictor $\mathbf{f}\colon \mathcal{X} \to \mathbb{R}^L$,*

$$\mathbb{E}\big[(\tilde{R}(\mathbf{f};S) - R(\mathbf{f}))^2\big] \leq \frac{1}{N} \cdot \mathbb{V}\left[\mathbf{p}^{\mathrm{t}}(x)^\top \ell(\mathbf{f}(x))\right] + \mathcal{O}\left(\mathbb{E}\big[\|\mathbf{p}^{\mathrm{t}}(x) - \mathbf{p}^*(x)\|_2\big]\right)^2. \tag{7}$$

Proof of Proposition 3. Let $\Delta \doteq \tilde{R}(\mathbf{f};S) - R(\mathbf{f})$. Then,

$$\mathbb{E}\big[(\tilde{R}(\mathbf{f};S) - R(\mathbf{f}))^2\big] = \mathbb{E}\big[\Delta^2\big] = \mathbb{V}[\Delta] + \mathbb{E}\big[\Delta\big]^2.$$

Observe that

$$\begin{aligned}
\mathbb{E}\big[\Delta\big] &= \mathop{\mathbb{E}}_{x}\big[(\mathbf{p}^{\mathrm{t}}(x) - \mathbf{p}^*(x))^\top \ell(\mathbf{f}(x))\big] \\
&\leq \mathop{\mathbb{E}}_{x}\big[\|\mathbf{p}^{\mathrm{t}}(x) - \mathbf{p}^*(x)\|_2 \cdot \|\ell(\mathbf{f}(x))\|_2\big] \\
&\leq \mathop{\mathbb{E}}_{x}\big[\|\mathbf{p}^{\mathrm{t}}(x) - \mathbf{p}^*(x)\|_2 \cdot c \cdot \|\ell(\mathbf{f}(x))\|_\infty\big] \\
&\leq c \cdot \mathop{\mathbb{E}}_{x}\big[\|\mathbf{p}^{\mathrm{t}}(x) - \mathbf{p}^*(x)\|_2\big],
\end{aligned}$$

where the second line is by the Cauchy-Schwartz inequality, and the third line by the equivalence of norms with a constant c. Now, (7) follows since $R(\mathbf{f})$ is a constant, implying

$$\mathbb{V}[\Delta] = \mathbb{V}[\tilde{R}(\mathbf{f};S)] = \frac{1}{N} \cdot \mathbb{V}\left[\mathbf{p}^{\mathrm{t}}(x)^\top \ell(\mathbf{f}(x))\right]. \quad \square$$

Unpacking the Proposition 3, the fidelity of the distilled risk depends on two factors: how variable the expected loss is for a random instance; and how well the teacher estimates $\mathbf{p}^{\mathrm{t}}$ approximates the true $\mathbf{p}^*$ in an MSE sense. The latter dominates in the large N regime, and admits a classic bias-variance tradeoff: per Appendix A.2, we may write (7) as

$$\mathbb{E}\big[(\tilde{R}(\mathbf{f};S) - R(\mathbf{f}))^2\big] \leq \frac{1}{N} \cdot \mathbb{V}\left[\mathbf{p}^{\mathrm{t}}(x)^\top \ell(\mathbf{f}(x))\right] + \mathcal{O}\Big(\|\mathbb{E}\big[\mathbf{p}^{\mathrm{t}}(x)\big] - \mathbf{p}^*(x)\|_2^2 + \mathbb{V}\left[\mathbf{p}^{\mathrm{t}}(x)\right]\Big). \tag{8}$$

The final two terms reflect a classic phenomena: increasing the teacher complexity can lower bias (yield better approximations of $\mathbf{p}^*$), *but* this is traded off with higher variance (more unstable predictions). Balancing the delicate tradeoff between these quantities yields a teacher with low MSE against $\mathbf{p}^*$. Proposition 3 establishes that such a teacher yields improved bounds on the student's generalisation error. Indeed, per the previous section, we may deduce that

$$R(\mathbf{f}) \leq \tilde{R}(\mathbf{f};S) + \mathfrak{C}(\mathcal{F},N) + \mathcal{O}\left(\mathbb{E}\|\mathbf{p}^{\mathrm{t}}(x) - \mathbf{p}^*(x)\|_2\right), \tag{9}$$

where $\mathfrak{C}$ is the penalty term for the hypothesis class $\mathcal{F}$ from Proposition 2. As is intuitive, using an imperfect teacher invokes an additional penalty depending on how far the predictions are from the Bayes, in a squared-error sense. See Proposition 4 (Appendix A) for a formal statement of (9), and a comparison to existing bounds.

4.2. Discussion and Implications

We have provided a statistical perspective on distillation, resting on the observation that a "Bayes teacher" can reduce variance in the student objective, and that a teacher approximating the Bayes probabilities offers a bias-variance tradeoff. Our results follow readily from this perspective, but the resulting implications and conceptual insights into distillation are subtle, and merit discussion.

What makes a teacher "good"? The above specifies a concrete means of assessing the quality of a teacher's soft labels: it is beneficial for them to be "good" probability estimates, in the sense of having low squared error against the Bayes probabilities $\mathbf{p}^*$. Indeed, Proposition 3 establishes that in such cases, the resulting student can generalise better.

We make three qualifying remarks. First, in practice, exactly assessing the squared error to $\mathbf{p}^*$ is infeasible (since $\mathbf{p}^*$ is often unknown), but one may *estimate* the quality of the teacher probabilities on a holdout set, e.g. by computing the log-loss, square-loss, or calibration error (Guo et al., 2017) on the one-hot labels. While such estimates are necessarily imperfect, they can detect poor teacher probabilities.

Second, Proposition 3 may be loose, and further is *not* a lower bound. A comprehensive theory of distillation would require specifying such *necessary* conditions. Nonetheless, the qualitative trend of our bounds can still hold in practice. For example, Figure 1 (see also §4.3) illustrates how increasing the depth of a ResNet model may increase accuracy, but degrade quality of probability estimates.

Third, the focus on approximating the Bayes probabilities $\mathbf{p}^*$ suggests that the teacher's predictions ought to primarily reflect aleatoric, rather than epistemic, uncertainty (Senge et al., 2014; Hüllermeier & Waegeman, 2021).

Why can more accurate teachers distill worse? A curious empirical observation is that more accurate teachers may lead to *worse* students (Müller et al., 2019). Our statistical view offers one possible insight on this behaviour: recall that we show that good probability modelling can aid student generalisation. However, while models producing good probabilities will also be accurate, models that are accurate do *not* necessarily offer good probabilities (Devroye et al., 1996, Section 6.7). Indeed, despite being accurate, deep networks often make over-confident, poorly calibrated predictions (Guo et al., 2017; Rothfuss et al., 2019).

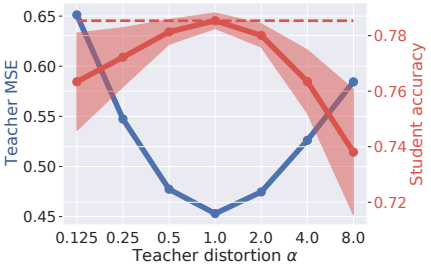


Figure 3. Illustration that teacher accuracy is insufficient to predict distillation performance. On a synthetic problem, we construct a family of teachers (parameterised by α) with the *same accuracy* (red dashed line), but differing *probability estimate* quality (blue line), as measured by MSE against the Bayes $\mathbf{p}^*$. Distilling each teacher to linear student models yields notably different accuracies (red solid line), with best performance for the teacher with best probability estimates (i.e., $\alpha = 1.0$).

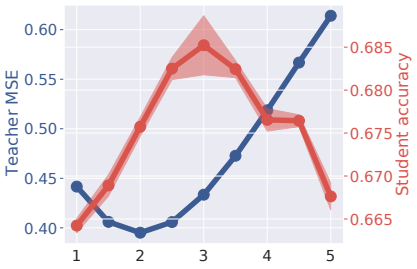


Figure 4. Illustration that suitable temperature scaling can improve the quality of teacher probabilities. Distilling from a ResNet-32 to ResNet-14 on CIFAR-100, when the temperature T is too high or too low, the teacher distribution is too flat or spiky respectively, and thus a poor reflection of the true $\mathbf{p}^*$. Tuning T improves the teacher probability quality, as measured by the square-loss on the test set. This also tracks the student accuracy, particularly at the extremes, consistent with our perspective.

Why does temperature scaling help? Temperature scaling (Hinton et al., 2015) is a common trick wherein one smooths overly confident teacher logits $\mathbf{f}^{\mathrm{t}}(x)$ via $\mathbf{p}^{\mathrm{t}}(y \mid x) \propto \exp(f_y^{\mathrm{t}}(x)/T)$ for $T > 0$. In line with our statistical view, such scaling can help the student target be closer to the true label distribution $\mathbf{p}^*$ (and thus aid generalisation), rather than just conveying the most accurate label.

Trading off bias for variance: model complexity. Following (8), to obtain a tighter bound on student generalisation, one may favour a higher-bias teacher (i.e., $\mathbf{p}^{\mathrm{t}}$ is a poorer approximation to $\mathbf{p}^*$) if it has lower variance (i.e., $\mathbf{p}^{\mathrm{t}}$ varies less across samples). Such a bias-variance tradeoff can be obtained, e.g., by tuning the teacher complexity.

In the context of neural networks, recent works have challenged the conventional wisdom (Geman et al., 1992) of increased model complexity (e.g., increased network width and/or depth) implying a higher variance (Neal et al., 2018; Yang et al., 2020). Note that Proposition 3 does not assume or impose a particular bias-variance tradeoff; it specifies how a *given* tradeoff affects student generalisation. In particular, it suggests that models simultaneously achieving low bias *and* variance ought to distill well.

Trading off bias for variance: label smoothing. Label smoothing (Szegedy et al., 2016) mixes the student labels with uniform predictions, i.e., uses $\mathbf{p}^{\mathrm{t}}(x) = (1-\alpha) \cdot \mathbf{e}_y + \frac{\alpha}{L} \cdot \mathbf{1}$ for $\alpha \in (0, 1]$. From the perspective of modelling $\mathbf{p}^*(x)$, this introduces a *bias* over using the observed labels, but lowers *variance* owing to the $(1-\alpha)$ scaling. Provided the bias is not too large, smoothing can thus aid generalisation. Recent work has also studied how smoothing induces variance-reduction in *optimisation* (Xu et al., 2020).

On training sample re-use. Our analysis requires that the teacher and student employ distinct training samples. This holds in the common setting where the teacher labels a large unlabelled pool (Radosavovic et al., 2018; Yalniz et al., 2019). Recently, Dao et al. (2021) showed how the teacher and student can use the sample training sample with a more careful "cross-fitting" procedure. This reduces the risk of overfitting in the teacher's probability estimates, at the expense of a slightly increased computational cost.

4.3. Illustration: the Value of Teacher Probabilities

We illustrate the above points with empirical results on both controlled synthetic as well as real-world problems, using linear models, neural networks, and decision trees. These highlight that statistical reasoning about distillation is not necessarily limited to their use in deep neural networks.

Teacher accuracy does not suffice (Figure 3). We first illustrate that teacher accuracy does not always translate to good student performance. For the Gaussian data of §3.2 with Bayes probabilities $\mathbf{p}^*$, we construct teacher probabilities $\bar{\mathbf{p}}(x) = \Psi_\alpha(\mathbf{p}^*(x))$. Here, Ψ_α is such that changing α does not affect accuracy, *but* makes $\mathbf{p}^*$ more concentrated, and thus degrades their probabilistic quality (e.g., MSE). Concretely, we let $\Psi_\alpha(u) = \frac{1}{2} + \frac{1}{2} \cdot (2u-1)^\alpha \cdot \mathrm{sgn}(2u-1)$ for $\alpha \in \{2^{-3}, 2^{-2}, \ldots, 2^3\}$. (See Appendix C.2 for a plot.)

Figure 3 confirms that despite *the teacher accuracy being the same* for all α, the student accuracy is *systematically harmed* when $\alpha \neq 1$. Further, the student's accuracy closely tracks the quality of the teacher's probability estimates, as measured by MSE $\|\bar{\mathbf{p}} - \mathbf{p}^*\|_2^2$. This is in keeping with our analysis on the value of good probability estimates.

Temperature scaling and probability quality (Figure 4). We verify that temperature scaling improves the quality of the teacher probabilities, and that this tracks the distilled student performance. We perform distillation from a ResNet-32 to ResNet-14 on CIFAR-100, and apply temperature scaling with $T \in \{1.0, 1.5, 2.0, \ldots, 5.0\}$. For each T, we

measure the test set MSE of the teacher probabilities (against the one-hot labels), as well as the accuracy of the student model. Note that, unlike the synthetic setting where we had access to $\mathbf{p}^*$, the former is necessarily an *estimate* of the actual quality of the teacher probabilities.

Figure 4 shows that as T is varied, the MSE of the teacher probabilities varies smoothly, with optimal temperature $T^* = 2$. The optimal student accuracy is achieved with a slightly higher temperature $T^* = 3$ — reflecting that we have provided *upper* bounds on the student risk — but as the probability quality degrades beyond this point, student accuracy rapidly declines, consistent with our perspective.

Trading off bias for variance: decision trees (Figure 5). We illustrate that in choosing amongst different teachers, *one may favour a higher-bias teacher* — i.e., one with worse probability estimates — *if it has lower variance*. We show this in a controlled setting, wherein we train a series of increasingly complex teacher models on a synthetic problem. Here, the data is sampled from a marginal $\mathbb{P}(x)$ which is uniform on $[0, 1]^2$, and $\mathbb{P}(y = 1 \mid x)$ has a "checkerboard" pattern, so that positives and negatives are sprinkled in alternating squares; see Appendix C.1 for an illustration.

We consider *decision tree* teachers of depth $d \in \{4, 5, 6, 7, 8\}$. Increasing d reduces teacher *bias*, but increases *variance* (since deeper trees can better approximate $\mathbf{p}^*$, but are more complex). For fixed d, we train a teacher on a training sample S (with $N = 5000$). We distill its predictions to a depth 4 student tree, and compute the test set teacher MSE and student AUC-ROC over 100 trials.

Figure 5 shows that at depth $d = 6$, the teacher achieves the best MSE approximation of $\mathbf{p}^*$. In keeping with our analysis, this also corresponds to the teacher whose resulting student generalises the best. Note that at $d > 6$, the teacher has lower bias, but higher variance; the higher-bias $d = 6$ teacher achieves a better tradeoff in terms of MSE. See Appendix C.4 for additional results on a distinct problem.

Trading off bias for variance: ResNet (Figure 1). Recall that in Figure 1, we train teacher ResNets of varying depths on CIFAR-100, and distill these to a student ResNet of fixed depth 8. We see that teachers with better probabilities (in an MSE sense) generally yield better students. Further, even though the teacher model gets increasingly more accurate as its depth increases, improved accuracy does *not* correspond to improved MSE. Prior work has observed that mismatch between the sizes of the student and teacher can also affect distillation (Cho & Hariharan, 2019; Mirzadeh et al., 2020). To mitigate such confounders, in Figure 10 (Appendix), we extend Figure 1 to include students with depth 14 and 20, and find the general trends for depth 8 hold.

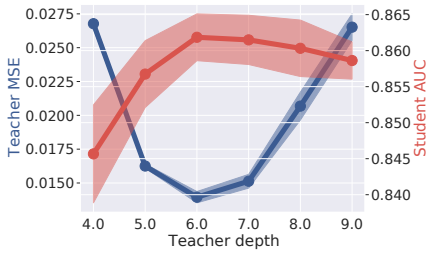


Figure 5. Illustration of bias-variance tradeoff in a controlled setting: one may favour a higher-bias teacher — i.e., one with comparatively worse probability estimates — if it has lower variance. On a synthetic problem, we consider a family of decision tree teachers of various depths, which are distilled to a depth 4 student. Increasing the teacher depth reduces the teacher bias, at the expense of increased variance. There is an optimal teacher depth $d = 6$ that balances these terms, and minimises the teacher MSE. This teacher MSE closely tracks student performance, per our analysis.

5. Applications of the Statistical View

Our statistical view has given conceptual insight into how distillation can improve classification. We now show a potential practical benefit of this view: it gives a simple, generic way to apply distillation to settings beyond classification. Specifically, by expressing objectives in terms of the Bayes class-probabilities, one may derive empirical estimates based on teacher model outputs. The results diverge from a naïve application of distillation, as we now illustrate.

5.1. Distillation for Bipartite Ranking

Given a distribution $\mathbb{P}$ over $\mathcal{X} \times \{\pm 1\}$, the bipartite ranking problem (Agarwal & Niyogi, 2009) involves learning a scorer $f \colon \mathcal{X} \to \mathbb{R}$ that minimises the *pairwise disagreement*,

$$\mathrm{PD}(f) \doteq \mathop{\mathbb{E}}_{x|y=+1} \mathop{\mathbb{E}}_{x'|y=-1} [\![f(x) < f(x')]\!],$$

i.e., the probability that a randomly drawn positive scores lower than a randomly drawn negative. This is equivalently one minus the area under the ROC curve of f (Agarwal et al., 2005), and measures how well f distinguishes the positive from negative samples. Given a training sample $S = \{(x_n, y_n)\}_{n=1}^N \sim \mathbb{P}^N$, an empirical estimate of PD is

$$\widehat{\mathrm{PD}}(f) \propto \sum_{i \in S_+} \sum_{j \in S_-} [\![f(x_i) < f(x_j)]\!], \tag{10}$$

where S_+, S_- are the subset of positive and negative samples. Intuitively, we consider pairs of samples with positive and negative labels, and assess the difference in their scores.

In a distillation context, given a powerful teacher model, one can construct a tighter approximation to the original risk. Indeed, since $\mathbb{P}(x \mid y) \propto \mathbb{P}(y \mid x) \cdot \mathbb{P}(x)$ we may write

$$\mathrm{PD}(f) \propto \mathop{\mathbb{E}}_{x \sim \mu} \mathop{\mathbb{E}}_{x' \sim \mu} \Big[p^*(x) \cdot (1 - p^*(x')) \cdot [\![f(x) < f(x')]\!] \Big],$$

where μ is the marginal distribution over instances, and $p^*(x) \doteq \mathbb{P}(y = +1 \mid x)$. Per the previous section, given a teacher model, we may use its probabilities p^{t} in place of p^* to obtained the *distilled bipartite risk*,

$$\widetilde{\mathrm{PD}}(f) \propto \sum_{i \in S, j \in S - \{i\}} p^{\mathrm{t}}(x_i) \cdot (1 - p^{\mathrm{t}}(x_j)) \cdot [\![f(x_i) < f(x_j)]\!].$$

Observe that we consider *all* pairs of samples, as opposed to partitioning them into groups of "positives" and "negatives" in (10). Further, each term involves *two* applications of smoothing with the teacher probabilities, as opposed to the standard single application in classification (cf. (4)). This point also arises in the following.

5.2. Distillation for Multiclass Retrieval

Given a distribution $\mathbb{P}$ over $\mathcal{X} \times [L]$, the multiclass retrieval problem (Lapin et al., 2018; Reddi et al., 2019) involves learning a predictor $\mathbf{f} \colon \mathcal{X} \to \mathbb{R}^L$ such that for an example (x, y), the *top-ranked* labels in $\mathbf{f}(x) \in \mathbb{R}^L$ contain the true label y. One popular choice of loss for this task is the family of *binary decoupled losses* (Reddi et al., 2019),

$$\ell(y, \mathbf{f}(x)) = \phi(f_y(x)) + \sum\nolimits_{y' \neq y} \phi(-f_{y'}(x)), \qquad (11)$$

where $\phi \colon \mathbb{R} \to \mathbb{R}_+$ is a margin loss for binary classification, e.g., the hinge loss $\phi(z) = [1 - z]_+$. Intuitively, the first term encourages high score for the "positive" label y, while the second encourages low score for the "negatives" $y' \neq y$.

For fixed $x \in \mathcal{X}$, the *expected* loss over draws of y is:

$$\begin{aligned} \mathbf{p}^*(x)^\top \ell(\mathbf{f}(x)) &= \mathbf{p}^*(x)^\top (\phi(\mathbf{f}(x)) - \phi(-\mathbf{f}(x))) + \mathbf{1}^\top \phi(-\mathbf{f}(x)) \\ &= \mathbf{p}^*(x)^\top \phi(\mathbf{f}(x)) + (\mathbf{1} - \mathbf{p}^*(x))^\top \phi(-\mathbf{f}(x)) \\ &= \sum\nolimits_{y \in [L]} \mathbb{P}(y \mid x) \cdot \phi(f_y(x)) + \\ &\quad \sum\nolimits_{y' \in [L]} (1 - \mathbb{P}(y' \mid x)) \cdot \phi(-f_{y'}(x)), \end{aligned} \qquad (12)$$

for all-ones vector $\mathbf{1} \in \mathbb{R}^L$, and $\mathbf{p}^*(x)$ is the vector of $\mathbb{P}(y \mid x)$ as before. Compared to (11), the first term considers all labels $y \in [L]$ as "smoothed positives", akin to the standard application of distillation. Interestingly, the second term considers all labels $y' \in [L]$ as "smoothed negatives", with variable weights depending on $\mathbb{P}(y' \mid x)$. Intuitively, the loss pays more attention to those negatives y' which are plausible alternate explanations for x.

Following our statistical perspective, on a finite sample $S = \{(x_n, y_n)\}_{n=1}^N$, one may replace these Bayes probabilities with teacher model estimates $\mathbf{p}^{\mathrm{t}}$, yielding:

$$\begin{aligned} \tilde{R}(\mathbf{f}; S) &= \frac{1}{N} \sum_{n \in [N]} \sum_{y \in [L]} p^{\mathrm{t}}(y \mid x_n) \cdot \Big[\phi(f_y(x_n)) + z(x_n) \Big] \\ z(x_n) &\doteq \sum\nolimits_{y' \in [L]} \alpha_{y'}(x_n) \cdot \phi(-f_{y'}(x_n)) \\ \alpha_{y'}(x_n) &\doteq 1 - p^{\mathrm{t}}(y' \mid x_n). \end{aligned} \qquad (13)$$

Akin to the standard classification case, applying such differential weighting on the positive *and* negative labels can lead to a lower-variance estimate of the expected loss. We may extend this to the popular softmax cross-entropy (cf. (2)),

$$\ell(y, \mathbf{f}(x)) = -f_y(x) + \log \left[\sum_{y' \in [L]} e^{f_{y'}(x)} \right].$$

Here, the inner summation may be regarded as penalising high scores for "negative" labels $y' \neq y$. Inspired by (12) we may differentially weight the negatives, yielding the loss:

$$\begin{aligned} \tilde{R}(\mathbf{f}; S) &= \frac{1}{N} \sum_{n \in [N]} \sum_{y \in [L]} p^{\mathrm{t}}(y \mid x_n) \cdot [-f_y(x_n) + z(x_n)] \\ z(x_n) &\doteq \log \left(\sum\nolimits_{y' \in [L]} \alpha_{y'}(x_n) \cdot e^{f_{y'}(x_n)} \right) \\ \alpha_{y'}(x_n) &\doteq 1 - [\![y' \neq y_n]\!] \cdot p^{\mathrm{t}}(y' \mid x_n). \end{aligned} \qquad (14)$$

Note here that $\alpha_{y_n}(x_n) = 1$, unlike in (13); this choice ensures the non-negativity of $\ell(y, \mathbf{f}(x))$, since the loss can be seen as log-loss under a weighted softmax distribution.

We reiterate that the above uses the teacher in two ways: the first is the standard use of distillation to smooth the positive labels. The second is a novel use of distillation to smooth the *negative* labels. Compared to the standard loss, for any candidate label y, we apply (instance- and label-dependent) weights to negative labels $y' \neq y$ when computing the loss. We thus term this objective *negative-aware distillation*.

Empirically, we have found superior performance by weighting negatives using "unnormalised" teacher probabilities, rather than p^{t} directly. Specifically, one may use $1 - \sigma(a \cdot s^{\mathrm{t}}_{y'}(x))$, where $s^{\mathrm{t}}_{y'}(x)$ is the teacher *logit*, $\sigma(\cdot)$ is the sigmoid function, and $a > 0$ is a scaling parameter which may be tuned. Intuitively, compared to using p^{t}, such a weighting allows for multiple y' to have high (or low) weights simultaneously. This is useful in retrieval scenarios, where there may be *multiple* relevant labels for a given $x \in \mathcal{X}$.

5.3. Empirical Illustration

We defer an empirical analysis of our bipartite ranking formulation to Appendix C.6. We now assess the negative-aware distillation objective on benchmark datasets for multiclass retrieval, AMAZONCAT-13K and AMAZONCAT-670K (McAuley & Leskovec, 2013; Bhatia et al., 2015). We use a feedforward "teacher" model with a single (linear) hidden layer of width 512, trained to minimise the softmax cross-entropy. For the "student", we make the hidden layer width 8 for AMAZONCAT-13K and 64 for AMAZONCAT-670K (since the latter has many more labels).

We compare training the student with one-hot labels, teacher logits (distillation), and teacher logits with additional negative smoothing in the softmax per (14). Our aim in doing so is to confirm that the core idea of negative-aware distilla-

	AMAZONCAT-13K			AMAZONCAT-670K		
Method	**P@1**	**P@3**	**P@5**	**P@1**	**P@3**	**P@5**
Teacher	0.8495	0.7412	0.6109	0.3983	0.3598	0.3298
Student	0.7913	0.6156	0.4774	0.3307	0.3004	0.2753
Student + SD	0.8131	0.6363	0.4918	0.3461	0.3151	**0.2892**
Student + NAD	**0.8560**	**0.7148**	**0.5715**	**0.3480**	**0.3161**	0.2865

Table 1. Precision@k of negative-aware distillation (NAD), standard distillation (SD) and student baseline on multiclass retrieval task. NAD improves performance over both techniques.

tion (14) – using a teacher to smooth *negatives* in addition to positives – can improve upon standard distillation.

We compare all methods based on the precision@k for $k \in \{1, 3, 5\}$, averaged over multiple runs. Table 1 summarises our findings. Distillation offers a small but consistent performance bump over the student baseline. Negative-aware distillation further improves upon this, especially for $k = 1$ and 3, confirming the value of weighing negatives differently. The gains are significant on AMAZONCAT-13K, where negative-aware distillation even *improves* upon the teacher model. Finally, we note that our use of (negative-aware) distillation here is subject to the same principles as the previous sections: e.g., temperature scaling on our teacher models improves their probabilistic calibration, and this tracks the student performance; see Appendix C.7.

5.4. Discussion and Implications

The above are two examples of how the statistical view of distillation facilitates its applications to problems beyond multi-class classification. The key ingredient in each application is expressing the true expected loss in terms of the Bayes class-probabilities. The resulting objectives involve non-standard uses of the teacher compared to classic distillation, e.g., to weight both positive *and* negative labels.

More broadly, one may involve a similar procedure for other problems with complex objectives; see Appendix B for additional examples, including robustness to label noise.

6. Concluding Remarks

Our statistical perspective on distillation builds on a simple observation: *distilling with the Bayes class-probabilities yields a better estimate of the population risk*. This provides conceptual insight into why distillation can help, and provides a simple, principled means of using distillation in settings beyond classification.

There are several potential directions for future study. For example, combining our analysis with study of the other effects of distillation (e.g., on optimisation (Phuong & Lampert, 2019)) would be of interest. It is also of interest to study more fine-grained notions of bias and variance, e.g., the notion of "regularisation samples" in Zhou et al. (2021).

References

Agarwal, S. and Niyogi, P. Generalization bounds for ranking algorithms via algorithmic stability. *J. Mach. Learn. Res.*, 10:441–474, June 2009. ISSN 1532-4435.

Agarwal, S., Graepel, T., Herbrich, R., Har-Peled, S., and Roth, D. Generalization bounds for the area under the roc curve. *Journal of Machine Learning Research*, 6(14): 393–425, 2005.

Ba, J. and Caruana, R. Do deep nets really need to be deep? In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 27*, pp. 2654–2662. Curran Associates, Inc., 2014.

Bennett, G. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45, 1962.

Bhatia, K., Jain, H., Kar, P., Varma, M., and Jain, P. Sparse local embeddings for extreme multi-label classification. In *Advances in Neural Information Processing Systems*, pp. 730–738, 2015.

Boucheron, S., Bousquet, O., and Lugosi, G. Theory of classification: a survey of some recent advances. *ESAIM: Probability and Statistics*, 2005.

Bucilă, C., Caruana, R., and Niculescu-Mizil, A. Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pp. 535–541, New York, NY, USA, 2006. ACM.

Cho, J. H. and Hariharan, B. On the efficacy of knowledge distillation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4793–4801, 2019.

Dao, T., Kamath, G. M., Syrgkanis, V., and Mackey, L. Knowledge distillation as semiparametric inference. In *International Conference on Learning Representations*, 2021.

Devroye, L., Györfi, L., and Lugosi, G. *A Probabilistic Theory of Pattern Recognition*, volume 31 of *Stochastic Modelling and Applied Probability*. Springer, 1996. ISBN 978-1-4612-0711-5.

Dong, B., Hou, J., Lu, Y., and Zhang, Z. Distillation $\approx$ early stopping? harvesting dark knowledge utilizing anisotropic information retrieval for overparameterized neural network, 2019.

Duchi, J., Hashimoto, T., and Namkoong, H. Distributionally robust losses for latent covariate mixtures, 2020.

Foster, D. J., Greenberg, S., Kale, S., Luo, H., Mohri, M., and Sridharan, K. Hypothesis set stability and generalization. In *Advances in Neural Information Processing Systems 32*, pp. 6726–6736. Curran Associates, Inc., 2019.

Furlanello, T., Lipton, Z. C., Tschannen, M., Itti, L., and Anandkumar, A. Born-again neural networks. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, pp. 1602–1611, 2018.

Geman, S., Bienenstock, E., and Doursat, R. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, 1992.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1321–1330, 2017.

Hinton, G. E., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015.

Hsu, D., Ji, Z., Telgarsky, M., and Wang, L. Generalization bounds via distillation. In *International Conference on Learning Representations*, 2021.

Hüllermeier, E. and Waegeman, W. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110(3):457–506, 2021.

Jain, H., Gidaris, S., Komodakis, N., Pérez, P., and Cord, M. Quest: Quantized embedding space for transferring knowledge. In Vedaldi, A., Bischof, H., Brox, T., and Frahm, J.-M. (eds.), *Computer Vision – ECCV 2020*, pp. 173–189, Cham, 2020. Springer International Publishing.

Ji, G. and Zhu, Z. Knowledge distillation in wide neural networks: Risk bound, data efficiency and imperfect teacher. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, 2020.

Kim, J., Park, S., and Kwak, N. Paraphrasing complex network: Network compression via factor transfer. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 2765–2774, Red Hook, NY, USA, 2018. Curran Associates Inc.

Lapin, M., Hein, M., and Schiele, B. Analysis and optimization of loss functions for multiclass, top-k, and multilabel classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(7):1533–1554, July 2018.

Liu, Y., Jia, X., Tan, M., Vemulapalli, R., Zhu, Y., Green, B., and Wang, X. Search to distill: Pearls are everywhere but not the eyes, 2019.

Lopes, R. G., Fenu, S., and Starner, T. Data-free knowledge distillation for deep neural networks. *CoRR*, abs/1710.07535, 2017.

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

Lukasik, M., Bhojanapalli, S., Menon, A. K., and Kumar, S. Does label smoothing mitigate label noise?, 2020.

Maurer, A. and Pontil, M. Empirical bernstein bounds and sample-variance penalization. In *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*, 2009.

McAuley, J. and Leskovec, J. Hidden factors and hidden topics: Understanding rating dimensions with review text. In *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys '13, pp. 165–172, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450324090.

Mirzadeh, S.-I., Farajtabar, M., Li, A., and Ghasemzadeh, H. Improved knowledge distillation via teacher assistant: Bridging the gap between student and teacher. In *AAAI*, 2020.

Mobahi, H., Farajtabar, M., and Bartlett, P. L. Self-distillation amplifies regularization in hilbert space. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33*, 2020.

Müller, R., Kornblith, S., and Hinton, G. E. When does label smoothing help? In *Advances in Neural Information Processing Systems 32*, pp. 4696–4705, 2019.

Natarajan, N., Dhillon, I. S., Ravikumar, P. D., and Tewari, A. Learning with noisy labels. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 1196–1204, 2013.

Nayak, G. K., Mopuri, K. R., Shaj, V., Babu, R. V., and Chakraborty, A. Zero-shot knowledge distillation in deep networks. *CoRR*, abs/1905.08114, 2019.

Neal, B., Mittal, S., Baratin, A., Tantia, V., Scicluna, M., Lacoste-Julien, S., and Mitliagkas, I. A modern take on the bias-variance tradeoff in neural networks. *CoRR*, abs/1810.08591, 2018. URL http://arxiv.org/abs/1810.08591.

Papernot, N., McDaniel, P. D., Wu, X., Jha, S., and Swami, A. Distillation as a defense to adversarial perturbations against deep neural networks. In *IEEE Symposium on Security and Privacy, SP 2016, San Jose, CA, USA, May 22-26, 2016*, pp. 582–597, 2016.

Phuong, M. and Lampert, C. Towards understanding knowledge distillation. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 5142–5151, 2019.

Radosavovic, I., Dollár, P., Girshick, R. B., Gkioxari, G., and He, K. Data distillation: Towards omni-supervised learning. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 4119–4128, 2018.

Rahbar, A., Panahi, A., Bhattacharyya, C., Dubhashi, D., and Chehreghani, M. H. On the unreasonable effectiveness of knowledge distillation: Analysis in the kernel regime, 2020.

Reddi, S. J., Kale, S., Yu, F., Holtmann-Rice, D., Chen, J., and Kumar, S. Stochastic negative mining for learning with large output spaces. In *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1940–1949. PMLR, 16–18 Apr 2019.

Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., and Bengio, Y. Fitnets: Hints for thin deep nets. In Bengio, Y. and LeCun, Y. (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Rothfuss, J., Ferreira, F., Boehm, S., Walther, S., Ulrich, M., Asfour, T., and Krause, A. Noise regularization for conditional density estimation, 2019.

Rudin, C. The P-Norm Push: A simple convex ranking algorithm that concentrates at the top of the list. *Journal of Machine Learning Research*, 10:2233–2271, Oct 2009.

Rusu, A. A., Colmenarejo, S. G., Gülçehre, Ç., Desjardins, G., Kirkpatrick, J., Pascanu, R., Mnih, V., Kavukcuoglu, K., and Hadsell, R. Policy distillation. In *4th International Conference on Learning Representations, ICLR 2016*, 2016.

Scott, C., Blanchard, G., and Handy, G. Classification with asymmetric label noise: consistency and maximal denoising. In *Conference on Learning Theory (COLT)*, pp. 489–511, 2013.

Senge, R., Bösner, S., Dembczyński, K., Haasenritter, J., Hirsch, O., Donner-Banzhoff, N., and Hüllermeier, E. Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty. *Information Sciences*, 255:16–29, 2014. ISSN 0020-0255.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 2818–2826, 2016.

Tang, J., Shivanna, R., Zhao, Z., Lin, D., Singh, A., Chi, E. H., and Jain, S. Understanding and improving knowledge distillation. *CoRR*, abs/2002.03532, 2020.

Xu, Y., Xu, Y., Qian, Q., Li, H., and Jin, R. Towards understanding label smoothing. *CoRR*, abs/2006.11653, 2020.

Yalniz, I. Z., Jégou, H., Chen, K., Paluri, M., and Mahajan, D. Billion-scale semi-supervised learning for image classification. *CoRR*, abs/1905.00546, 2019. URL http://arxiv.org/abs/1905.00546.

Yang, Z., Yu, Y., You, C., Steinhardt, J., and Ma, Y. Rethinking bias-variance trade-off for generalization of neural networks. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10767–10777. PMLR, 13–18 Jul 2020.

Yim, J., Joo, D., Bae, J., and Kim, J. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7130–7138, July 2017.

Yuan, L., Tay, F. E. H., Li, G., Wang, T., and Feng, J. Revisiting knowledge distillation via label smoothing regularization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3902–3910. IEEE, 2020.

Zagoruyko, S. and Komodakis, N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *5th International Conference on Learning Representations, ICLR 2017*. OpenReview.net, 2017.

Zhang, Z. and Sabuncu, M. R. Self-distillation as instance-specific label smoothing. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, 2020.

Zhou, H., Song, L., Chen, J., Zhou, Y., Wang, G., Yuan, J., and Zhang, Q. Rethinking soft labels for knowledge distillation: A bias–variance tradeoff perspective. In *International Conference on Learning Representations*, 2021.

Supplementary material for "A statistical perspective on distillation"

A. Theory: discussion and additional results

A.1. Comparison to existing bounds

Our bound in Proposition 3 is not directly comparable to prior work; e.g., Phuong & Lampert (2019) bound the probability that the student and teacher disagree, not the generalisation error. Foster et al. (2019) assume the student is constrained to be close to a teacher, not trained with soft-labels. We remark that, unlike the latter, we assume the teacher is trained on an independent sample from the student; the more challenging case of sample reuse on teacher and student is an interesting topic of future study.

A.2. Proof of the claim in (8)

Note that by Jensen's inequality, and the definition of variance,

$$\begin{aligned}\left(\mathop{\mathbb{E}}_{x}\left[\|\mathbf{p}^{\mathrm{t}}(x)-\mathbf{p}^{*}(x)\|_2\right]\right)^2 &\leq \mathop{\mathbb{E}}_{x}\left[\|\mathbf{p}^{\mathrm{t}}(x)-\mathbf{p}^{*}(x)\|_2^2\right]\\ &= \|\mathbb{E}\left[\mathbf{p}^{\mathrm{t}}(x)\right]-\mathbf{p}^{*}(x)\|_2^2+\mathbb{V}\left[\mathbf{p}^{\mathrm{t}}(x)\right].\end{aligned}$$

Thus, we further have

$$R(\mathbf{f}) \leq \frac{1}{N}\cdot\mathbb{V}\left[\mathbf{p}^{\mathrm{t}}(x)^{\top}\ell(\mathbf{f}(x))\right]+c^2\cdot\left(\|\mathbb{E}\left[\mathbf{p}^{\mathrm{t}}(x)\right]-\mathbf{p}^{*}(x)\|_2^2+\mathbb{V}\left[\mathbf{p}^{\mathrm{t}}(x)\right]\right). \tag{15}$$

A.3. Additional results

We now explicate how to convert Proposition 3 into a generalisation bound for the student's performance, mirroring Proposition 2 for the case of a Bayes teacher.

Proposition 4. *Pick any bounded loss ℓ. Fix a hypothesis class $\mathcal{F}$ of predictors $f\colon \mathcal{X}\to\mathbb{R}^L$, with induced class $\mathcal{H}\subset[0,1]^{\mathcal{X}}$ of functions $h(x)\doteq \mathbf{p}^{\mathrm{t}}(x)^{\top}\ell(\mathbf{f}(x))$. Suppose $\mathcal{H}$ has uniform covering number $\mathcal{N}_\infty$. Then, for any $\delta\in(0,1)$, with probability at least $1-\delta$ over $S\sim\mathbb{P}^N$,*

$$R(\mathbf{f}) \leq \tilde{R}(\mathbf{f};S)+\mathcal{O}\left(\sqrt{\tilde{\mathbb{V}}_N(\mathbf{f})\cdot\frac{\log\frac{\mathcal{M}_N}{\delta}}{N}}+\frac{\log\frac{\mathcal{M}_N}{\delta}}{N}\right)+\mathcal{O}\left(\mathbb{E}\|\mathbf{p}^{\mathrm{t}}(x)-\mathbf{p}^{*}(x)\|_2\right),$$

where $\mathcal{M}_N\doteq\mathcal{N}_\infty(\frac{1}{N},\mathcal{H},2N)$ and $\tilde{\mathbb{V}}_N(\mathbf{f})$ is the empirical variance of the loss values.

Proof of Proposition 4. Let $\tilde{R}(\mathbf{f})=\mathbb{E}\left[\tilde{R}(\mathbf{f};S)\right]$ and $\Delta\doteq\tilde{R}(\mathbf{f};S)-R(\mathbf{f})$. Following the proof of Proposition 2, we get that with probability $1-\delta$,

$$\tilde{R}(\mathbf{f}) \leq \tilde{R}(\mathbf{f};S)+\mathcal{O}\left(\sqrt{\tilde{\mathbb{V}}_N(\mathbf{f})\cdot\frac{\log\frac{\mathcal{M}_N}{\delta}}{N}}+\frac{\log\frac{\mathcal{M}_N}{\delta}}{N}\right), \tag{16}$$

where $\mathcal{M}_N\doteq\mathcal{N}_\infty(\frac{1}{N},\mathcal{H},2N)$ and $\tilde{\mathbb{V}}_N(\mathbf{f})$ is the empirical variance of the loss values. Furthermore, the following holds

$$\begin{aligned}|\tilde{R}(\mathbf{f})-R(\mathbf{f})| &= \left|\mathbb{E}\left[\tilde{R}(\mathbf{f};S)\right]-\mathbb{E}\left[\hat{R}_*(\mathbf{f};S)\right]\right|\\ &\leq \mathbb{E}\left[\|\mathbf{p}^{\mathrm{t}}(x)-\mathbf{p}^{*}(x)\|_2\cdot\|\ell(\mathbf{f}(x))\|_2\right].\end{aligned}$$

Thus, we have

$$R(\mathbf{f}) \leq \tilde{R}(\mathbf{f})+C\cdot\mathbb{E}\left[\|\mathbf{p}^{\mathrm{t}}(x)-\mathbf{p}^{*}(x)\|_2\right]. \tag{17}$$

for some constant $C>0$. The result follows by combining (16) and (17). □

B. Additional applications of the statistical framework

The statistical framework espoused above gives a simple yet generic way to understand and use distillation: for population objectives that make complex use of the Bayes class-probabilities, one may derive empirical versions that are based on the outputs of a teacher model. We present here some additional potential applications of the framework.

B.1. Robustness to label noise

Our statistical perspective gives a way to interpret the viability of distillation under label noise. Given samples from a distribution $\bar{\mathbb{P}}$ that is subject to class-conditional label noise (Scott et al., 2013; Natarajan et al., 2013) — i.e., $\bar{\mathbb{P}}(y \mid x) = \mathbf{T}_{y,:}\, \mathbb{P}(\cdot \mid x)$ for noise transition matrix $\mathbf{T}$ — a common family of loss-correction techniques involve learning with the loss $\mathbf{T}^{-1}\ell$. This can be interpreted as constructing a plug-in estimate of $\mathbb{P}(y \mid x)$ via $\mathbb{P}(\cdot \mid x) = \mathbf{T}^{-1}\bar{\mathbb{P}}(y \mid x)$.

Given a teacher model that is trained on *noisy* data — and thus produces estimates of the noisy $\bar{\mathbb{P}}(y \mid x)$ — we may thus compute a tighter estimate to $\mathbf{T}^{-1}\bar{\mathbb{P}}(y \mid x)$, and use this to weigh the loss. In fact, such a procedure was recently explored in Lukasik et al. (2020), but with a purely empirical motivation. Our statistical framework gives a means of justifying this procedure.

B.2. Ranking with a push-loss

Motivated by the bipartite ranking problem in § 5.1, consider now a multiclass classification problem over $\mathcal{X} \times [L]$. We may consider a contextual version of the bipartite ranking loss,

$$R(f) = \mathop{\mathbb{E}}_{x} \mathop{\mathbb{E}}_{y \sim P_+(x)} \mathop{\mathbb{E}}_{y' \sim P_-(x)} [\![f_y(x) < f_{y'}(x)]\!],$$

where $P_+, P_- \in \Delta_L$ are distributions over "positive" and "negative" labels respectively. For the positives, the natural choice is $P_+ = \mathbb{P}(y \mid x)$. For the negatives, one possible choice is $P_- \propto C - \mathbb{P}(y' \mid x)$ for $C = \max_{y''} \mathbb{P}(y'' \mid x)$, so that the labels with the lowest probability under $\mathbb{P}(\cdot)$ are most likely to be negative. We may rewrite the risk as

$$R(f) = \mathop{\mathbb{E}}_{x} \Big[\sum_{y,y'} \mathbb{P}(y \mid x) \cdot (C - \mathbb{P}(y' \mid x)) \cdot [\![f_y(x) < f_{y'}(x)]\!] \Big].$$

As before, we may replace $\mathbb{P}(\cdot \mid x)$ with the estimates from a teacher model.

One may generalise the above to use a *push loss* (Rudin, 2009) as follows: for increasing $g \colon \mathbb{R} \to \mathbb{R}$, define

$$R_{\text{push}}(f) = \mathop{\mathbb{E}}_{x} \mathop{\mathbb{E}}_{y \sim P_+(x)} g \left(\mathop{\mathbb{E}}_{y' \sim P_-(x)} [\![f_y(x) < f_{y'}(x)]\!] \right),$$

so that one penalises false negatives more strongly. As an example, when $g(z) = z^p$, as $p \to +\infty$ we have a contextual analogue of the p-norm push loss of Rudin (2009):

$$R_{\text{push}}(f) = \mathop{\mathbb{E}}_{x} \mathop{\mathbb{E}}_{y \sim P_+(x)} \max_{y' \in \text{supp}(P_-(x))} [\![f_y(x) < f_{y'}(x)]\!],$$

where the inner quantity may be understood as the rank of the highest scoring negative sample. As before, we may rewrite the risk as

$$R_{\text{push}}(f) = \mathop{\mathbb{E}}_{x} \left[\sum_{y} \mathbb{P}(y \mid x) \cdot g \left(\sum_{y'} (C - \mathbb{P}(y' \mid x)) \cdot [\![f_y(x) < f_{y'}(x)]\!] \right) \right].$$

For example, when $g(z) = \log(1 + z)$, replacing the indicator function with an exponential surrogate yields

$$\bar{R}_{\text{push}}(f) = \mathop{\mathbb{E}}_{x} \left[\sum_{y} \mathbb{P}(y \mid x) \cdot \log \left(1 + \sum_{y'} (C - \mathbb{P}(y' \mid x)) \cdot e^{f_{y'}(x) - f_y(x)} \right) \right],$$

which is similar to the negative-aware distillation objective (14).

B.3. Robustness to covariate shift

The covariate shift problem involves a test distribution whose marginal distribution over instances differs from that observed during training. One means of guarding against such problem is to adopt a distributionally robust optimisation objective, such as

$$R_{\mathrm{dro}}(f) = \sup_{\mu' \in B(\mu, \epsilon)} \mathop{\mathbb{E}}_{x \sim \mu'} \mathop{\mathbb{E}}_{y|x} \ell(y, f(x)),$$

where μ is the observed training distribution over instances, and $B(\cdot, \epsilon)$ denotes a suitable ball of size ϵ. As observed in Duchi et al. (2020), when B is a *CVaR-ball*,

$$R_{\mathrm{dro}}(f) = \inf_{\lambda} \left[\frac{1}{\epsilon} \cdot \mathop{\mathbb{E}}_{x \sim \mu} \left(\mathop{\mathbb{E}}_{y|x} \ell(y, f(x)) - \lambda \right)_+ + \lambda \right].$$

Intuitively, we only retain those samples whose expected losses exceed some threshold λ^*, which in turn is some distribution-dependent quantity.

Typically, given a sample $S = \{(x_n, y_n)\}_{n=1}^N \sim \mathbb{P}^N$, estimating $\mathop{\mathbb{E}}_{y|x} \left[\ell(y, f(x)) \right]$ reliably is infeasible, since we often have only one observation for a given x. This motivated a procedure in Duchi et al. (2020) that constructs a different bound to $R_{\mathrm{dro}}(f)$. However, in a distillation setting, we may estimate $\mathop{\mathbb{E}}_{y|x} \left[\ell(y, f(x)) \right]$ using the scores of a teacher model. This gives a significantly simpler means of approximately minimising R_{dro}, albeit at the expense of increased bias.

C. Additional experiments

We present additional experiments to complement those in the main body. We illustrate the following:

(i) we visualise the checkerboard data used to illustrate the bias-variance tradeoff for decision trees (§C.1)

(ii) we visualise the distortion function Ψ_α used to show that teacher accuracy can be wholly at odds with student generalisation (§C.2)

(iii) distilling with a Bayes teacher becomes increasing useful as the underlying problem becomes noisier (§C.3)

(iv) the bias-variance tradeoff can be illustrated by explicitly distortion the Bayes class-probability function (§C.4)

(v) the bias-variance tradeoff can be illustrated on ResNets with varying depth (§C.5)

(vi) the distilled bipartite ranking objective can benefit over standard one-hot training (§C.6)

(vii) we show that on synthetic Gaussian data as well as the AMAZONCAT-13K data, temperature scaling of the teacher probabilities can improve their calibration and student performance.

C.1. Checkerboard data

Figure 6 shows the checkerboard data used in §4. Here, our samples are drawn from a marginal that is uniform on $[0, 1]^2$. We choose the class-probability function to be

$$\mathbb{P}(y = +1 \mid x) = \sum_{i=0}^{(B+1)/2} \sum_{j=0}^{(B+1)/2} \sigma(40 \cdot s_{2i,2j}(x)) + \sum_{i=1}^{(B-1)/2} \sum_{j=1}^{(B-1)/2} \sigma(40 \cdot s_{2i,2j}(x)) \sigma(40 \cdot s_{2i+1,2j+1}(x))$$

$$s_{i,j}(x) = \frac{1}{2 \cdot B} - \|x - \mu_{i,j}\|_\infty$$

for B^2 equally spaced squares with centroids $\mu_{i,j}$, and $B = 3$.

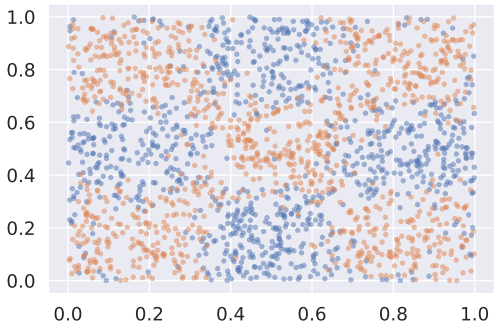


Figure 6. Checkerboard data used for decision tree.

C.2. Teacher probability distortion function

Figure 7 plots the result of applying the distortion function Ψ_α to the teacher probabilities. When $\alpha = 1$, we obtain the standard sigmoid function. When $\alpha \gg 1$, the probabilities become nearly uninformative, as they are strongly concentrated around 0.5; this makes the student's learning problem significantly noisier, and thus more challenging. When $\alpha \ll 1$, the probabilities becomes overly concentrated near the extremes; this becomes tantamount to training on the original labels itself.

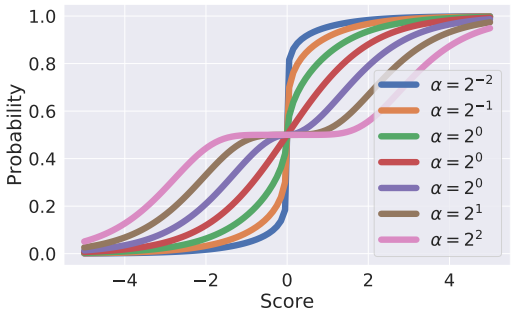


Figure 7. As tuning parameter α is increased, the teacher probabilities $\bar{p}(x) = \Psi_\alpha(\mathbf{p}^*(x))$ increasingly deviate from the Bayes probabilities $\mathbf{p}^*(x)$.

C.3. Bayes distillation is valuable for non-separable problems

Figure 8 continues the exploration of the Gaussian setting in §3.1 for $N = 100$ samples. We now vary the distance r between the means of each of the Gaussians. When r is small, the two distributions grow closer together, making the classification problem more challenging. At the same time, smaller r makes the one-hot labels have higher variance compared to the Bayes class-probabilities. Consequently, the gains of distillation over the one-hot encoding are greater for this setting, in line with our guarantee on the lower-variance Bayes-distilled risk aiding generalisation (Proposition 2).

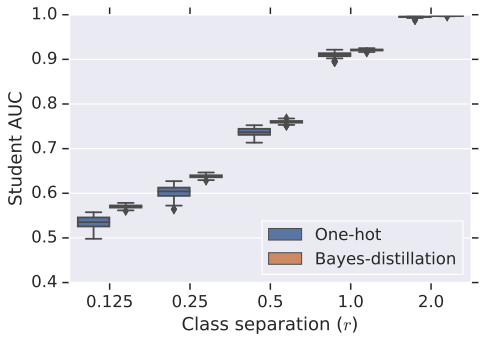


Figure 8. Distillation versus one-hot encoding on a synthetic dataset comprising Gaussian class-conditionals with means $r \cdot (+1, +1)$ and $r \cdot (-1, 1)$. We vary r so as to change the separation between the classes. Both methods see worse performance as r is smaller, but the gains of distillation over the one-hot encoding are greater for this setting.

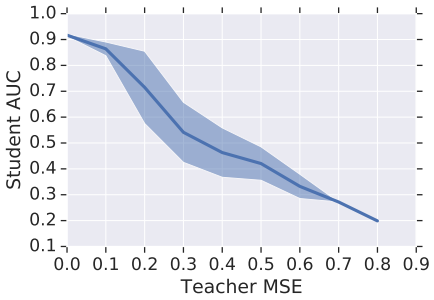
C.4. Bias-variance tradeoff: alternate distortion

We present an alternate verification of the bias-variance tradeoff, wherein we distort the Bayes probabilities in a different manner. Continuing the same synthetic Gaussian data as in §3.2, we now consider a family of teachers of the form

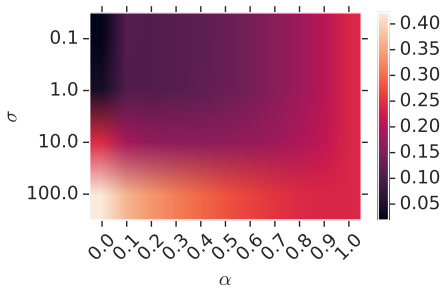
$$\mathbf{p}^{\mathrm{t}}(x) = (1 - \alpha) \cdot \Psi((\theta^*)^\top x + \sigma^2 \cdot \epsilon) + \frac{\alpha}{2}, \tag{18}$$

where $\Psi(z) \doteq (1 + e^{-z})^{-1}$ is the sigmoid, $\alpha \in [0, 1]$, $\sigma > 0$, and and $\epsilon \sim \mathcal{N}(0, 1)$ comprises independent Gaussian noise. Increasing α induces a *bias* in the teacher's estimate of $\mathbf{p}^*(x)$, while increasing σ induces a *variance* in the teacher over fresh draws. Combined, these control the teacher's mean squared error (MSE) $\mathbb{E}\left[\|\mathbf{p}^*(x) - \mathbf{p}^{\mathrm{t}}(x)\|_2^2\right]$, which by Proposition 3 bounds the gap between the population and distilled empirical risk.

For each such teacher, we compute its MSE, as well as the test set AUC of the corresponding distilled student. Figure 9(a) shows the relationship between the the teacher's MSE and the student's AUC. In line with the theory, more accurate estimates of $\mathbf{p}^*$ result in better students. Figure 9(b) also shows how the teacher's MSE depends on the choice of σ and α, demonstrating that multiple such pairs can achieve a similar MSE. As before, we see that a teacher may trade-off bias for variance in order to achieve a low MSE.



(a) Relationship between teacher's MSE against true class-probability and student's test set AUC. In keeping with the theory, teachers which better approximate $\mathbf{p}^*$ yield better students.

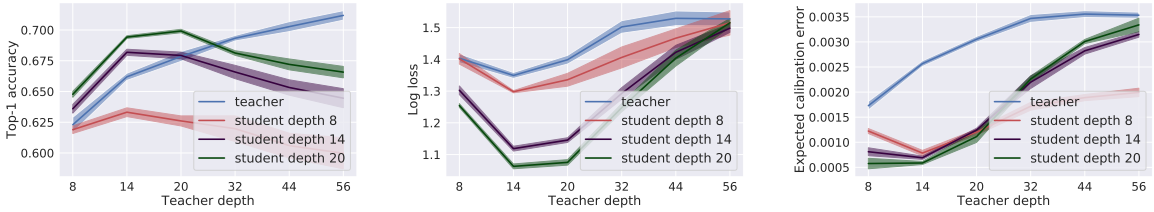


(b) Relationship between teacher's bias and variance, and corresponding MSE against true class-probability. The teacher can achieve a given MSE through multiple possible bias-variance combinations.

Figure 9. Bias-variance tradeoff on Gaussian data.

C.5. Trading off bias for variance: ResNet

Recall that in Figure 1, we train teacher ResNets of varying depths on CIFAR-100, and distill these to a student ResNet of fixed depth 8. We see that teachers with better probabilities (in an MSE sense) generally yield better students. Further, even though the teacher model gets increasingly more accurate as its depth increases, improved accuracy does *not* correspond to improved MSE. Prior work has observed that mismatch between the sizes of the student and teacher can also affect distillation (Cho & Hariharan, 2019; Mirzadeh et al., 2020). To mitigate such confounders, in Figure 10, we extend Figure 1 to include students with depth 14 and 20, and find the general trends for depth 8 hold.



(a) Top-1 accuracy. (b) Log-loss. (c) Expected calibration error.

Figure 10. Illustration of bias-variance tradeoff on CIFAR-100: teachers with better probability estimates generally yield better students. Results extend Figure 1 to include students of varying depth.

C.6. Distillation for bipartite ranking

Recall the following distillation objective for bipartite ranking problems (§5.1): given a training sample $S = \{(x_n, y_n)\}_{n=1}^N$ where $y_n \in \{\pm 1\}$, we construct

$$\widetilde{\mathrm{PD}}(f) \propto \sum_{i \in S, j \in S - \{i\}} p^{\mathrm{t}}(x_i) \cdot (1 - p^{\mathrm{t}}(x_j)) \cdot [\![f(x_i) < f(x_j)]\!]$$

for teacher model p^{t}. This may be contrast to the standard bipartite ranking objective, which effectively corresponds to a "one-hot" teacher $p^{\mathrm{t}}(x_n) = (y_n + 1)/2$.

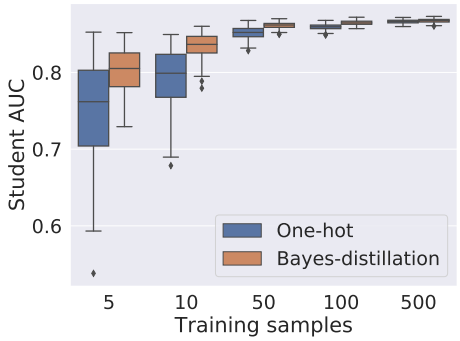
As in the classification setting, we show that learning with the distilled objective can significantly boost student performance. We consider the same synthetic Gaussian problem as §3.2, and compare training with the "one-hot" versus "Bayes teacher", with the latter employing probabilities given by the true $\mathbf{p}^*(x) = (\mathbb{P}(y = -1 \mid x), \mathbb{P}(y = +1 \mid x))$. To facilitate gradient-based optimisation, we replace the indicator function with convex surrogate $\phi(z) = \log(1 + e^{-z})$, yielding

$$\widetilde{\mathrm{PD}}(f) \propto \sum_{i \in S, j \in S - \{i\}} p^{\mathrm{t}}(x_i) \cdot (1 - p^{\mathrm{t}}(x_j)) \cdot \phi(f(x_i) - f(x_j)).$$

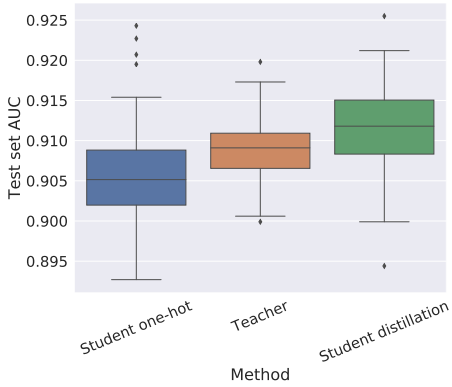
Figure 11(a) compares the student area under the ROC curve (AUC) on the test sample. Distilling with the Bayes teacher is seen to significantly boost performance in the low-sample regime.

To further assess the efficacy of the formulation in a real-world setting, we consider the Fashion MNIST dataset. While the data is inherently multi-class, we construct a binarised version suitable for bipartite ranking by focussing on samples with the labels T-Shirt and Shirt only. We train a teacher LeNet-5 model, which is distilled into a student model that shares the LeNet-5 architecture, but has all filter sizes reduced by half; such a setup has been considered in Lopes et al. (2017); Nayak et al. (2019). When applying distillation, we do not use the raw teacher predictions $\mathbf{p}^{\mathrm{t}}(x)$, but rather the common trick of mixing them with the training labels via $(1-\alpha)\cdot\mathbf{e}_y + \alpha\cdot\mathbf{p}^{\mathrm{t}}(x)$; following Nayak et al. (2019), we use $\alpha = 0.7$. (This can be understood as mitigating the bias of the target labels.)

Figure 11(b) compares the test set AUC for the teacher, student trained with one-hot labels, and student trained with distillation; the results are presented for 100 independent trials. We see that distillation notably improves performance over one-hot training, and in fact can sometimes exceed the performance of the teacher.



(a) Synthetic dataset comprising Gaussian class-conditionals. Here, we employ the "Bayes teacher", which uses the true $\mathbf{p}^*$ to train the student, which is a linear model.



(b) Fashion MNIST dataset, binarised to classify T-Shirt versus Shirt. Here, we use a LeNet-5 teacher, which is distilled to a LeNet-5 student with all filter sizes reduced by half.

Figure 11. Bipartite ranking version of distillation versus one-hot encoding. Our distillation objective significantly improves over one-hot training in terms of the student area under the ROC curve (AUC).

C.7. Temperature scaling and teacher calibration

We study the effect of temperature scaling on the student's performance, as well as the teacher's probability quality. In Figure 12, we study this on the AMAZONCAT-13K data. From left-to-right, we increased the temperature making the model generate less confident labels to the students. We see that the student's performance has a very high anti-correlation with the teacher's log-loss (a proxy for the distance between the Bayes label probability and teacher's prediction).

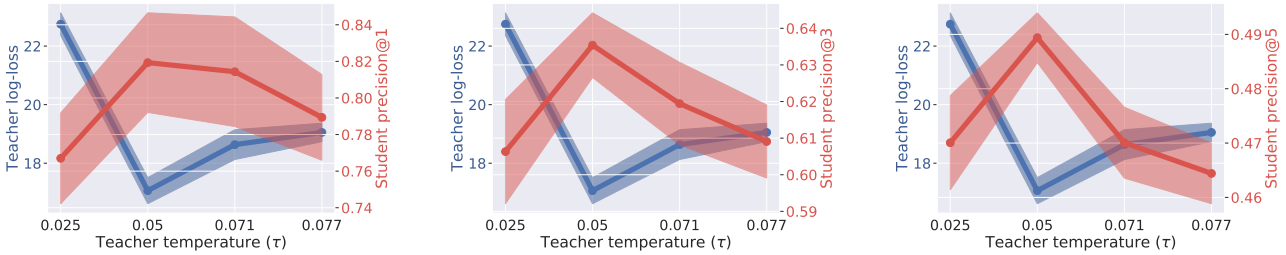


Figure 12. Temperature scaling versus accuracy: AMAZONCAT-13K data.

As further verification, we show that similar trends hold for the synthetic Gaussian data of §3.1. Here, we take the Bayes $\mathbf{p}^* = \sigma((\theta^*)^{\mathrm{T}}x)$ and apply temperature scaling inside the sigmoid. Evidently, we expect that applying no scaling should

give optimal student performance, as these offer the Bayes probabilities. Figure 13 confirms this, and also shows that as the temperature is varied, the calibration of the resulting teacher in terms of both log-loss and MSE is significantly harmed. This is a further corroboration of the quality of teacher probabilities playing an important role in distillation performance.

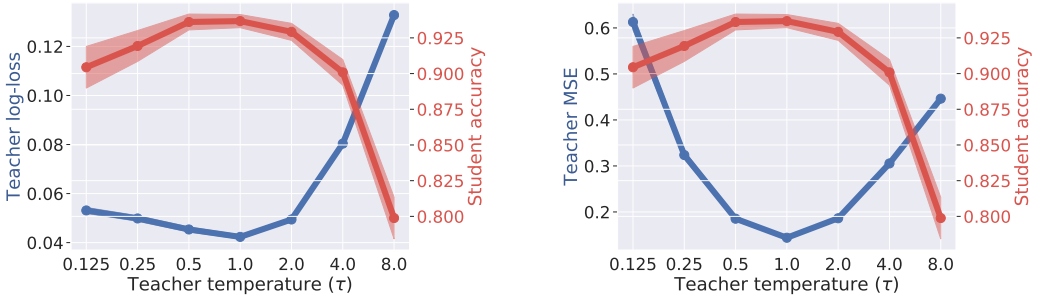


Figure 13. Temperature scaling versus accuracy: Gaussian data.