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 $B = 3$.

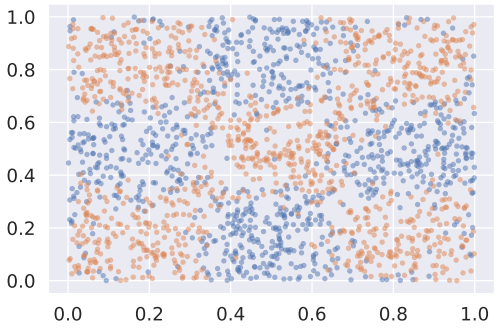


Figure 6. Checkerboard data used for decision tree.

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.

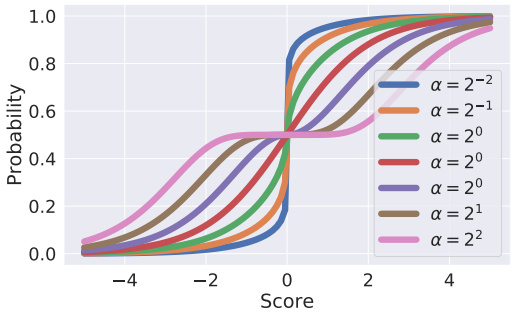


Figure 7. As tuning parameter α is increased, the teacher probabilities $\bar{p}(x) = \Psi_\alpha(\mathbf{p}^*(x))$ increasingly deviate from the Bayes probabilities $\mathbf{p}^*(x)$.

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).

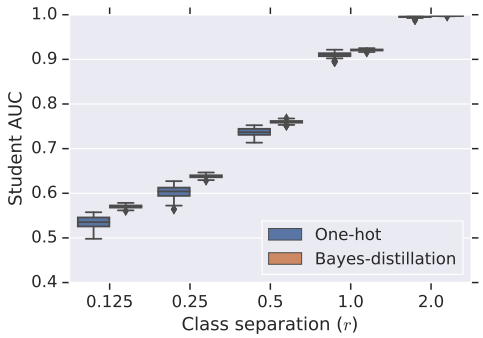


Figure 8. Distillation versus one-hot encoding on a synthetic dataset comprising Gaussian class-conditionals with means $r \cdot (+1, +1)$ and $r \cdot (-1, 1)$. We vary r so as to change the separation between the classes. Both methods see worse performance as r is smaller, but the gains of distillation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

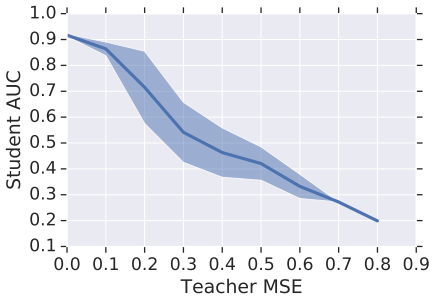
C.4. Bias-variance tradeoff: alternate distortion

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

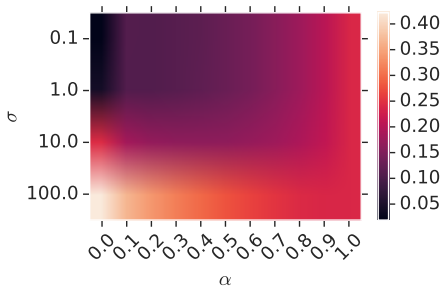
$$\mathbf{p}^{\mathrm{t}}(x) = (1 - \alpha) \cdot \Psi((\theta^*)^\top x + \sigma^2 \cdot \epsilon) + \frac{\alpha}{2}, \tag{18}$$

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.

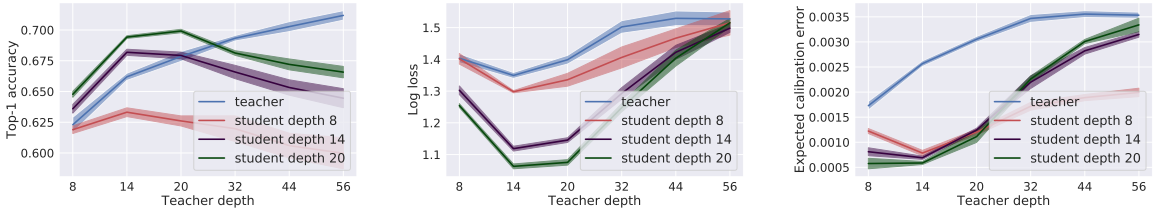


(b) Relationship between teacher's bias and variance, and corresponding MSE against true class-probability. The teacher can achieve a given MSE through multiple possible bias-variance combinations.

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 $p^{\mathrm{t}}(x_n) = (y_n + 1)/2$.

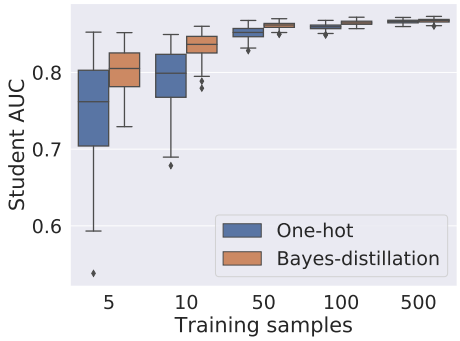
As in the classification setting, we show that learning with the distilled objective can significantly boost student performance. We consider the same synthetic Gaussian problem as §3.2, and compare training with the "one-hot" versus "Bayes teacher", with the latter employing probabilities given by the 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

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

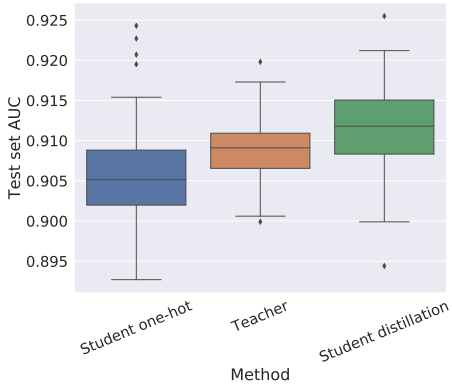
Figure 11(a) compares the student area under the ROC curve (AUC) on the test sample. Distilling with the Bayes teacher is seen to significantly boost performance in the low-sample regime.

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).

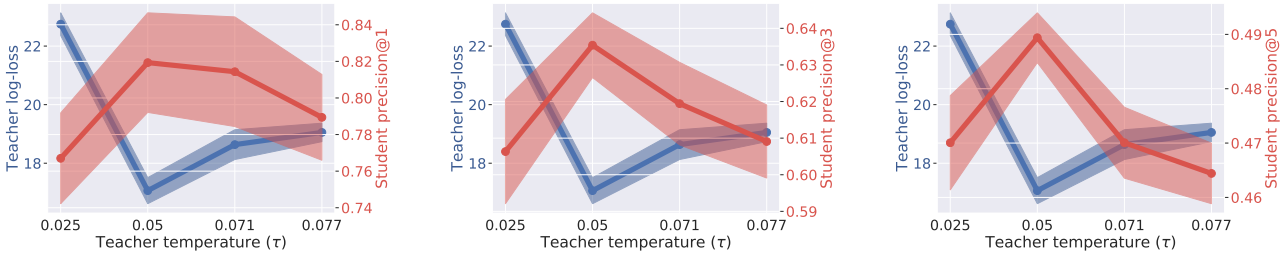


Figure 12. Temperature scaling versus accuracy: AMAZONCAT-13K data.

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.

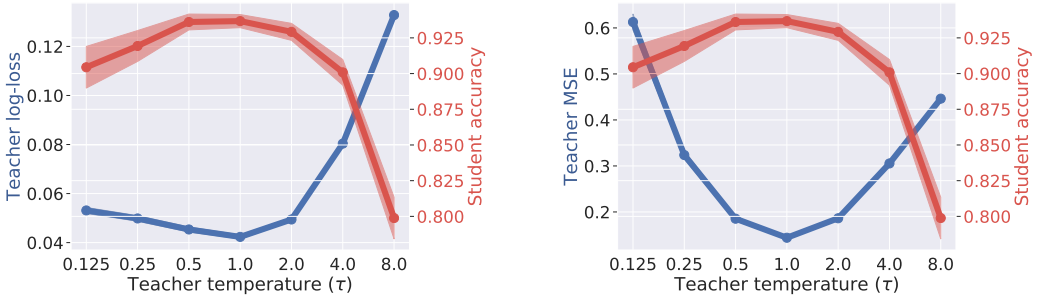


Figure 13. Temperature scaling versus accuracy: Gaussian data.