KNOWLEDGE DISTILLATION AS SEMIPARAMETRIC INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

A popular approach to model compression is to train an inexpensive student model to mimic the class probabilities of a highly accurate but cumbersome teacher model. Surprisingly, this two-step knowledge distillation process often leads to higher accuracy than training the student directly on labeled data. To explain and enhance this phenomenon, we cast knowledge distillation as a semiparametric inference problem with the optimal student model as the target, the unknown Bayes class probabilities as nuisance, and the teacher probabilities as a plug-in nuisance estimate. By adapting modern semiparametric tools, we derive several new guarantees for the prediction error of standard distillation and develop several enhancements with improved guarantees. We validate our findings empirically on both tabular data and image data and observe consistent improvements from our knowledge distillation enhancements.

1 INTRODUCTION

Knowledge distillation (KD) (Craven & Shavlik, 1996; Breiman & Shang, 1996; Bucila et al., 2006; Li et al., 2014; Ba & Caruana, 2014; Hinton et al., 2015) is a widely used model compression technique that enables deployment of ML models on devices such as phones, watches, and virtual assistants (Stock et al., 2020). KD works by training a large teacher model and then a smaller student model that tries to match the teacher's predictions. It is easily applied to any classification procedure and any teacher and works remarkably well across a wide variety of domains including computer vision (Hinton et al., 2015), natural language processing (Sanh et al., 2019), and tabular data settings (Liu et al., 2018; Tan et al., 2018; Fakoor et al., 2020). Despite KD's ubiquity and importance, the reasons for its success remains elusive. The practice of KD is often guided by heuristics.

In practice, KD is often performed in diverse ways, for different tasks and different domains. It is often tailored to particular model types (e.g., CNN, RNNs, Transformer). This precludes a general theoretical understanding of KD. Hinton et al. (2015) remarks that KD works by transferring the "dark knowledge" from the teacher's predicted class probabilities. As we detail below, several works have tried to explain how this dark knowledge can improve the student's performance beyond training from scratch. Our work is motivated by the observation of Menon et al. (2020) that the teacher probabilities serve as a proxy for the *Bayes probabilities* (i.e., the true class probabilities) and that the closer the teacher and Bayes probabilities, the better the student's performance should be.

Building on this observation, we cast KD as a plug-in approach to *semiparametric inference*: distillation aims to fit a student model $\hat{f}$ in the presence of nuisance (the Bayes probabilities p_0) with the teacher's probabilities $\hat{p}$ as a plug-in estimate of p_0. This insight allows us to adapt modern tools from semiparametric inference to analyze the error of a distilled student in Sec. 3. Our analysis further suggests several enhancements for KD. For example, in Sec. 4 we show that avoiding sample reuse between the teacher and student through *cross-fitting* (see, e.g., Chernozhukov et al., 2018) can improve KD performance. Moreover, a common strategy in semiparametric inference is to improve plug-in estimation by reducing plug-in bias. The method of *orthogonal machine learning* (Chernozhukov et al., 2018; Foster & Syrgkanis, 2019) suggests correcting the teacher's probabilities by stepping in the $y-\hat{p}$ direction where y is an observed label vector. We argue in Sec. 4 that while this minimizes the teacher bias, the variance could explode. We instead propose a correction that balances the bias and variance, leading to a better overall generalization bound.

We empirically validate our theoretical analysis in Sec. 5. In a synthetic setting where we have access to the Bayes classifier, our modified loss yields estimators that are closer to the Bayes classifier. Moreover, our enhanced distillation improves dramatically over vanilla distillation in failure regimes of the

teacher, e.g., when the teacher is heavily biased or overfitted. On five real tabular datasets, cross-fitting and corrected distillation improve the student's performance by up to 4% AUC point compared to vanilla KD. Furthermore, on a common image classification dataset (CIFAR-10, Krizhevsky & Hinton, 2009), when the teacher model overfits, cross-fitting and our corrected loss improve the student accuracy by up to 1.5%. This demonstrates the wide applicability of our enhancements.

Related work. Since we do not aim review the vast literature on KD in its entirety here, we point the interested reader to Gou et al. (2020) for a good overview of the results. In the following we review theory research studying why distillation works. We further go over some empirical studies and applications of knowledge distillation in the extended literature review in App. A.

A number of papers have argued that the availability of soft class probabilities from the teacher rather than hard labels enables us to improve training of the student model. This was hypothesized in Hinton et al. (2015) with empirical justification. Phuong & Lampert (2019) consider the case where the teacher is a fixed linear classifier and the student is a linear classifier or a deep linear network. They show that the student can learn the teacher perfectly if it has access to training examples more than the ambient dimensions. Vapnik & Izmailov (2015) discusses the setting of learning with privileged information where one has additional information at training time which is not available at test time. Lopez-Paz et al. (2015) draws a connection between this and distillation, arguing that distillation is effective because the teacher learns a better representation allowing the student to learn at a faster rate. They hypothesize that a teacher's soft predictions indicating if an example is hard or easy enables this. Tang et al. (2020) argues using empirical evidence that label smoothing and reweighting of training examples using the teacher's predictions are key to benefits offered by knowledge distillation. Mobahi et al. (2020) analyzed the case of self-distillation where the student and teacher function classes are the same. They showed that this is equivalent to increasing regularisation strength for kernelized models. Bu et al. (2020) considers model compression in a rate-distortion framework, where the rate is the size of the student model and distortion is the difference in excess risk between the teacher and the student. Menon et al. (2020) consider the case of losses such that the population risk is linear in the Bayes class probabilities. They consider *distilled empirical risk* and *Bayes distilled empirical risk* which are the risk computed using the teacher class probabilities and Bayes class probabilities respectively rather than the observed label. They show that the variance of the Bayes distilled empirical risk is lower than the empirical risk. Then using analysis from Maurer & Pontil (2009); Bennett (1962), they derive the excess risk of the distilled empirical risk as a function of the ℓ_2 distance between the teacher's class probabilities and the Bayes class probabilities. We significantly depart from Menon et al. (2020) in multiple ways: i) our Thm. 1 allows for the common practice of data re-use, ii) our results cover the standard distillation losses (1) and (2) which are non-linear in p_0, iii) we use localized Rademacher analysis to achieve tight fast rates for strongly-convex losses, which are commonly used for distillation, and iv) we use techniques from semiparametric inference to improve upon vanilla distillation.

2 Knowledge Distillation Background

We consider a multiclass classification problem with k classes and n training datapoints $z_i = (x_i, y_i)$ sampled independently from some distribution $\mathbb{P}$. Each feature vector x belongs to a set $\mathcal{X}$, each label vector $y \in \{e_1, \ldots, e_j\} \subset \{0,1\}^k$ is a one-hot encoding of the class label, and the conditional probability of observing each label is the *Bayes class probability* function $p_0(x) = \mathbb{E}[Y \mid X = x]$. Our aim is to identify a scoring rule $f : \mathcal{X} \to \mathbb{R}^k$ that minimizes a prediction loss on average under the distribution $\mathbb{P}$.

Knowledge distillation. Knowledge distillation (KD) is a two-step training process where one first uses a labeled dataset to train a teacher model and then trains a student model to predict the teacher's predicted class probabilities. Typically the teacher model is larger and more cumbersome, while the student is smaller and more efficient. Knowledge distillation was first motivated by model compression (Bucila et al., 2006), to find compact yet high-performing models to be deployed (such as on mobile devices).

In training the student to match the teacher's prediction probability, there are several types of loss functions that are commonly used. Let $\hat{p}(x) \in \mathbb{R}^k$ be the teacher's vector of predicted class probabilities, $f(x) \in \mathbb{R}^k$ be the student model's output, and $[k] \triangleq \{1, 2, \ldots, k\}$. The most popular distillation loss functions[1] $\ell(z; f(x), \hat{p}(x))$ include the squared error logit loss (Ba & Caruana, 2014)

$$\ell_{\text{se}}(z; f(x), \hat{p}(x)) \triangleq \sum_{j \in [k]} \tfrac{1}{2} \left(f_j(x) - \log(\hat{p}_j(x)) \right)^2 \tag{1}$$

[1]These loss functions do not depend on the ground-truth label y, but we use the augmented notation $\ell(z; f(x), \hat{p}(x))$ to accommodate the enhanced distillation losses presented in Sec. 4.

and the annealed cross-entropy loss (Hinton et al., 2015)

$$\ell_\beta(z;f(x),\hat{p}(x)) = -\sum_{j\in[k]} \frac{\hat{p}_j(x)^\beta}{\sum_{l\in[k]}\hat{p}_l(x)^\beta} \log\left(\frac{\exp(\beta f_j(x))}{\sum_{l\in[k]}\exp(\beta f_l(x))}\right) \tag{2}$$

for an inverse temperature $\beta > 0$. These loss functions measure the divergence between the probabilities predicted by the teacher and the student.

A student model trained with knowledge distillation often performs better than the same model trained from scratch (Bucila et al., 2006; Hinton et al., 2015). In Secs. 3 and 4, we will adapt modern tools from semiparametric inference to understand and enhance this phenomenon.

3 Distillation as Semiparametric Inference

In semiparametric inference (Kosorok, 2007), one aims to estimate a target parameter or function f_0, but that estimation depends on an auxiliary *nuisance function* p_0 that is unknown and not of primary interest. We cast the knowledge distillation process as a semiparametric inference problem, by treating the unknown Bayes class probabilities p_0 as nuisance and the teacher's predicted probabilities as a plug-in estimate of that nuisance. This perspective allows us bound the generalization of the student in terms of the mean squared error (MSE) between the teacher and the Bayes probabilities. In the next section (Sec. 4) we use techniques from semiparametric inference to enhance the performance of the student. The interested reader could consult Tsiatis (2007) for more details on semiparametric inference.

Our analysis starts from taking the following perspective on distillation. For a given pointwise loss function $\ell(z;f(x),p_0(x))$, we view the goal of the student as minimizing an oracle population loss over a function class $\mathcal{F}$,

$$L_D(f,p_0) = \mathbb{E}[\ell(Z;f(X),p_0(X))] \quad \text{with} \quad f_0 \triangleq \operatorname{argmin}_{f\in\mathcal{F}} L_D(f,p_0).$$

The main hurdle is that this is objective depends on the unknown Bayes probabilities p_0. We view the teacher's model $\hat{p}$ as an approximate version of p_0 and bound the distillation error of the student as a function of the teacher's estimation error.

Typical semiparametric inference considers cases where f_0 is a finite dimensional parameter; however recent work of Foster & Syrgkanis (2019) extends this framework to infinite dimensional models f_0 and to develop statistical learning theory with a nuisance component framework. The distillation problem fits exactly into this setup.

Bounds on vanilla distillation with data re-use As a first step we derive a vanilla bound on the error of the distilled student model without any further modifications of the distillation process, i.e., we assume that the student is trained on the same data as the teacher and is trained by running empirical risk minimization (ERM) on the plug-in loss, plugging in the teacher's model instead of p_0, i.e.,

$$\hat{f} = \operatorname{argmin}_{f\in\mathcal{F}} L_n(f,\hat{p}) \quad \text{for} \quad L_n(f,\hat{p}) \triangleq \mathbb{E}_n[\ell(Z;f(X),\hat{p}(X))] \tag{3}$$

where $\mathbb{E}_n[X] = \frac{1}{n}\sum_{i=1}^n X_i$ denotes the empirical expectation of a random variable.

Technical definitions Before presenting our main theorem we introduce some technical notation. For a vector valued function f that takes as input a random variable X, we use the shorthand notation $\|f\|_{p,q} \triangleq \|\|f(X)\|_p\|_{L^q} = \mathbb{E}\left[\|f(X)\|_p^q\right]^{1/q}$. Let ∇_ϕ and ∇_π denote the partial derivatives of $\ell(z;\phi,\pi)$, with respect to its second and third input correspondingly and $\nabla_{\phi\pi}$ the Jacobian of cross partial derivatives, i.e., $[\nabla_{\phi\pi}\ell(z;\phi,\pi)]_{i,j} = \frac{\partial^2}{\partial\phi_j\partial\pi_i}\ell(z;\phi,\pi)$. Finally, let

$$q_{f,p}(x) = \mathbb{E}[\nabla_{\phi\pi}\ell(Z;f(X),p(X)) \,|\, X = x] \quad \text{and} \quad \gamma_{f,p}(x) = \mathbb{E}_{U\sim\text{Unif}([0,1])}[q_{f,Up+(1-U)p_0}(x)].$$

Critical radius Finally, we need to define the notion of the critical radius (see e.g., Wainwright (2019)) of a function class, which typically provides tight learning rates for statistical learning theory tasks. For any function class $\mathcal{F}$ we define the localized Rademacher complexity as:

$$\mathcal{R}(\delta;\mathcal{F}) = \mathbb{E}_{X_{1:n},\epsilon_{1:n}}[\sup_{f\in\mathcal{F}:\|f\|_2\le\delta} \tfrac{1}{n}\textstyle\sum_{i=1}^n \epsilon_i f(X_i)]$$

where ϵ_i are i.i.d. random variables taking values equiprobably in $\{-1,1\}$. The *critical radius* of a class $\mathcal{F}$, taking values in $[-H,H]$, is the smallest solution δ_n to the inequality $\mathcal{R}(\delta;\mathcal{F}) \le \frac{\delta^2}{H}$.

Theorem 1 (Vanilla distillation analysis)**.** *Assume that $f_0 \in \mathcal{F}$ and $\mathcal{F}$ is a convex set and that the teacher estimates a model $\hat{p}\in\mathcal{P}$ on the same sample as the student. Let $\delta_{n,\zeta} = \delta_n + c_0\sqrt{\frac{\log(c_1/\zeta)}{n}}$ for some universal constants c_0,c_1 and δ_n is an upper bound on the critical radius the function class:*

$$\mathcal{G} \triangleq \{z \to r(\ell(z;f(x),p(x)) - \ell(z;f_0(x),p(x))) : f\in\mathcal{F}, p\in\mathcal{P}, r\in[0,1]\}$$

Assume that, for all $z \in \mathcal{Z}$ and π, the loss $\ell(z;\phi,\pi)$ is σ-strongly convex with respect to ϕ, uniformly bounded in $[-H,H]$ for all z,ϕ,π and let: $\mu(z) = \sup_\phi \|\nabla_\phi \ell(z;\phi,\hat{p}(x))\|_2$. Moreover, suppose that the function class f satisfies the ℓ_2/ℓ_4 ratio condition: $\sup_{f\in\mathcal{F}} \frac{\|f-f_0\|_{2,4}}{\|f-f_0\|_{2,2}} \leq C$. Then $\hat{f}$ (3) satisfies

$$\|\hat{f}-f_0\|_{2,2}^2 \leq \frac{1}{\sigma^2} O(\delta_{n,\zeta}^2 C^2 H^2 \|\mu\|_4^2 + \|\gamma_{f_0,\hat{p}}^\top (\hat{p}-p_0)\|_{2,2}^2) \quad \textit{with probability} \quad 1-\zeta.$$

Remark The proof is in App. D. We note that the ℓ_2/ℓ_4 ratio requirement can be removed, albeit the final bound will depend on a uniform bound $\|\mu\|_\infty = \sup_z |\mu(z)|$, instead of $\|\mu\|_4$. Moreover, we note that the strong convexity requirement for ℓ is satisfied by all standard distillation objectives, as it is strong convexity with respect to the output of f and not the parameters of f. Even this requirement could be removed, but this would yield slow rate bounds of the form: $\|\hat{f}-f_0\|_{2,2}^2 = O(\delta_{n,\zeta} + \|\gamma_{f_0,\hat{p}}^\top (\hat{p}-p_0)\|_{2,2}^2)$.

Thm. 1 shows that the standard approach to distillation can provably work to provide an accurate student and potentially improve upon training from scratch. For instance, if the teacher is very accurate, i.e., $\hat{p} \approx p_0$, then the term that remains contains the fourth moment of the loss gradient $\|\mu\|_4$. If we were to replace the teacher $\hat{p}(x)$ with the raw labels y, then we would have gotten a dependence on the fourth moment of the loss gradient $\tilde{\mu}(z) \triangleq \sup_{f\in\mathcal{F}} \|\nabla_\phi \ell(z;f(x),y)\|_2$, which now contains extra variation due to the variance of the labels y. Thus in practice it can be that $\|\mu\|_4 \ll \|\tilde{\mu}\|_4$.

Drawbacks of vanilla distillation There are two major drawbacks of Thm. 1, corresponding to two distinct sources of KD student error: error due to teacher overfitting and error due to teacher underfitting. First, the function class $\mathcal{G}$ contains the variation of the student's model over the whole $\mathcal{P}$, the function class of the teacher. If $\mathcal{P}$ is a large model space, with a large critical radius, then $\mathcal{G}$ will inherit this complexity and the final bound will depend on the sample complexity of the teacher. Second, if the teacher is not very accurate, i.e., $\|\hat{p}-p_0\|_{2,2}$ is large, then the student's model will be significantly impacted. In the next section, we use techniques from semiparametric inference to tackle both of these problems.

We show two examples to illustrate how the problems of teacher overfitting/underfitting could manifest themselves in knowledge distillation. These examples serve to lower bound student performance, in the worst case, by the teacher's critical radius and the teacher's class probability MSE, matching the upper bounds given in Thm. 1. However, we note that in other better-case scenarios vanilla distillation can perform better than the upper-bounding Thm. 1 would imply.

Example 1 (Impact of teacher overfitting on vanilla distillation)**.** *Suppose that the Bayes probability function p_0 has Lipschitz gradient and, for known $\epsilon > 0$, belongs to the set*

$$\mathcal{P} = \{p : p_j(x) \in [\epsilon,1], \forall x \in \mathcal{X}, j \in [k]\}. \tag{4}$$

Suppose moreover that $X \in \mathbb{R}^d$ has Lebesgue density bounded away from 0 and ∞ and that $\epsilon < \frac{1}{4} \frac{\mathbb{E}[p_{0,j}(X)(1-p_{0,j}(X))^2]}{\mathbb{E}[(1-p_{0,j}(X))/p_{0,j}(X)]}$ for each j. Consider the teacher estimates $\hat{p}_j(x) = \max(\epsilon, \tilde{p}_j(x))$ for $\tilde{p}$ the Nadaraya-Watson kernel smoothing estimator (Nadaraya, 1964; Watson, 1964)

$$\tilde{p}(x) \triangleq \begin{cases} y_i & \textit{if} \quad x = x_i \\ \sum_{i=1}^n y_i K((x-x_i)/h) / \sum_{i=1}^n K((x-x_i)/h) & \textit{otherwise} \end{cases}$$

with kernel $K(x) = \|x\|_2^{-a} \mathbb{I}[\|x\|_2 \leq 1]$, $a \in (0, d/2)$, and $h = n^{-1/(4+d)}$. Notably, $\tilde{p}$ exactly interpolates the observed labels (i.e., $\tilde{p}(x_i) = y_i$) but, by Belkin et al. (2019, Thm. 1), also satisfies $\mathbb{E}[\|p_0 - \hat{p}\|_{2,2}^2] = O(n^{-4/(4+d)})$. Consider now a student that learns a constant *prediction rule via distilled ERM* (3) *with the squared error logit loss* (1) *and*

$$\mathcal{F} = \{f : f(x) = f(x') \in [\log(\epsilon), 0]^k \textit{ for all } x,x' \in \mathcal{X}\}.$$

Since the student only has access to the teacher's training set probabilities, its estimate $\hat{f}(x) = \frac{1}{n}\sum_{i=1}^n \log(\max(y_i,\epsilon))$ is inconsistent for the optimal constant rule $f_0(x) = \mathbb{E}[\log(p_0(X))]$ as

$$f_{0,j}(x) - \mathbb{E}\hat{f}_j(x) = \mathbb{E}[\log(p_{0,j}(X)) - \log(\max(Y_j,\epsilon))] \geq \mathbb{E}[\tfrac{p_{0,j}(X) - \max(Y_j,\epsilon)}{p_{0,j}(X)} + \tfrac{(\max(Y_j,\epsilon) - p_{0,j}(X))^2}{2}]$$
$$= \mathbb{E}[\tfrac{p_{0,j}(X)(1-p_{0,j}(X))^2 + (1-p_{0,j}(X))(p_{0,j}(X)-\epsilon)^2}{2}] - \epsilon\mathbb{E}[\tfrac{1-p_{0,j}(X)}{p_{0,j}(X)}] \geq \mathbb{E}[\tfrac{p_{0,j}(X)(1-p_{0,j}(X))^2}{4}]$$

by Taylor's theorem with Lagrange remainder. This non-vanishing student error reflects the non-vanishing critical radius δ_n of the composite student-teacher function class $\mathcal{G}$ defined in Thm. 1; since the student function class $\mathcal{F}$ has low complexity, the complexity of $\mathcal{G}$ is driven by the highly flexible interpolating teacher. We will demonstrate how to avoid this dependence on teacher complexity in Sec. 4.2.

Example 2 (Impact of teacher underfitting on vanilla distillation)**.** *Suppose that p_0 belongs to the set (4) and does not vary with x. Consider the constant teacher estimate $\hat{p}(x)=\frac{\bar{y}}{1+\lambda}$ obtained via ridge regression with regularization strength λ and $\bar{y}\triangleq\frac{1}{n}\sum_{i=1}^{n}y_i$. A constant student prediction rule trained via distilled ERM (3) with the squared error logit loss (1) yields $\hat{f}(x)=\log(\frac{\bar{y}}{1+\lambda})$ for all $x\in\mathcal{X}$ and satisfies*

$$f_0-\hat{f}=\log(p_0)-\log(\hat{p})\geq\mathrm{diag}(\tfrac{1}{p_0})(p_0-\hat{p})=\gamma_{f_0,p_0}^{\top}(p_0-\hat{p})=\gamma_{f_0,p_0}^{\top}\tfrac{\lambda p_0-(\bar{y}-p_0)}{(1+\lambda)}$$

by the concavity of the logarithm. Since $P(\bar{y}_j-p_{0,j}(x)\geq\theta p_{0,j}(x))\leq\frac{\zeta}{k}$ for $\theta\geq\sqrt{\frac{2(1-p_0(x))}{np_0(x)}\log(\frac{k}{\zeta})}+\frac{4}{3}\frac{1}{np_0(x)}\log(\frac{k}{\zeta})$ by Bernstein's inequality (Bernstein, 1946), we have $P(\|f_0-\hat{f}\|_{2,2}^2\geq\|\gamma_{f_0,p_0}^{\top}(p_0-\hat{p})\|_{2,2}^2)\geq 1-\zeta$ matching the teacher MSE dependence of the Thm. 1 upper bound whenever $\lambda\geq\sqrt{\frac{2}{n\epsilon}\log(\frac{k}{\zeta})}+\frac{4}{3}\frac{1}{n\epsilon}\log(\frac{k}{\zeta})$. Moreover, since $\limsup_{n\to\infty}\frac{\sqrt{n}(\bar{y}-p_0)}{\sqrt{2p_0(1-p_0)\log\log(n)}}=1$ with probability 1 by the law of the iterated logarithm, $\|f_0-\hat{f}\|_{2,2}^2=\Omega(\min(1,\lambda^2))$ with probability 1 whenever $\lambda\geq\sqrt{\frac{2\log\log(n)}{n\epsilon}}$.

4 Enhancing Knowledge Distillation

To address the two distinct inefficiencies of vanilla distillation revealed in Sec. 3, we will adapt and generalize two distinct techniques from semiparametric inference: orthogonal correction and cross-fitting.

4.1 Combating teacher underfitting with γ-correction

We can view the plug-in distillation loss $\ell(z;f(x),\hat{p}(x))$ as a zeroth order Taylor approximation to the ideal loss $\ell(z;f(x),p_0(x))$ around $\hat{p}$. An ideal first-order approximation would take the form

$$\ell(z;f(x),\hat{p}(x))+\langle p_0(x)-\hat{p}(x),\nabla_{\pi}\ell(z;f(x),\hat{p}(x))\rangle.$$

However, its computation also requires knowledge of p_0. Nevertheless, since $p_0(x)=\mathbb{E}[Y\mid X=x]$, we can always construct an unbiased estimate of the ideal first order term by replacing $p_0(x)$ with y:

$$\ell_{ortho}(z;f(x),\hat{p}(x))=\ell(z;f(x),\hat{p}(x))+\langle y-\hat{p}(x),\mathbb{E}[\nabla_{\pi}\ell(z;f(x),\hat{p}(x))\mid x]\rangle. \tag{5}$$

For standard distillation base losses like (1) and (2), the *orthogonal loss* (5) has an especially simple form, as $\nabla_{\pi}\ell(z;f(x),\hat{p}(x))$ is linear in f. Indeed, this is true more generally for the following class of Bregman divergence losses.

Definition 1 (Bregman divergence losses)**.** *Any* Bregman divergence *loss function of the form*

$$\ell(z;f(x),p(x))\triangleq\Psi(f(x))-\Psi(g(p(x)))-\langle\nabla_g\Psi(g(p(x))),f(x)-g(p(x))\rangle \quad \textit{has}$$

$$\ell_{ortho}(z;f(x),p(x))=\ell(z;f(x),p(x))+(y-p(x))^{\top}\nabla_p g(p(x))^{\top}\nabla_{gg}^2\Psi(g(p(x)))f(x)+\mathrm{const} \tag{6}$$

with the second term bilinear in $f(x)$ and $y-p(x)$. For the squared error logit loss (1), $\Psi(s)=\frac{1}{2}\|s\|_2^2$, $g(p)=\log(p)$, and the correction matrix $\nabla_p g(p(x))^{\top}\nabla_{gg}^2\Psi(g(p(x)))=\mathrm{diag}(\frac{1}{p(x)})$. Similarly, the annealed cross-entropy loss (2) falls into the class of Bregman divergence losses.

We will show that orthogonal correction (5) can significantly improve student bias due to teacher underfitting; however, for our standard distillation losses (1) and (2), the same orthogonal correction term often introduces unreasonably large variance due to division by small probabilities appearing in the correction matrix (see Definition 1). To grant ourselves more flexibility in balancing bias and variance, we propose and analyze a family of *γ-corrected losses*, parameterized by a matrix valued function $\gamma:\mathcal{X}\to\mathbb{R}^k\times\mathbb{R}^k$:

$$\ell_{\gamma}(z;f(x),p(x))\triangleq\ell(z;f(x),p(x))+(y-p(x))^{\top}\gamma(x)f(x)$$

to mimic the bilinear structure of Bregman orthogonal losses (6). Note that we can always recover the vanilla distillation loss by taking $\gamma\equiv 0$. We denote the associated population and empirical risks by

$$L_D(f,p,\gamma)\triangleq\mathbb{E}[\ell_{\gamma}(Z;f(X),p(X))] \quad \text{and} \quad L_n(f,p,\gamma)\triangleq\mathbb{E}_n[\ell_{\gamma}(Z;f(X),p(X))].$$

Observe that at p_0 the correction term is mean-zero and hence $L_D(f,p_0,\gamma)$ is independent of γ

$$L_D(f,p_0)\triangleq\mathbb{E}[\ell(Z;f(X),p_0(X))]=L_D(f,p_0,\gamma) \quad \text{for all} \quad \gamma.$$

The γ-corrected loss has strong connections to the literature on Neyman orthogonality (Chernozhukov et al., 2018; Chernozhukov et al., 2016; Nekipelov et al., 2018; Chernozhukov et al., 2018; Foster & Syrgkanis, 2019). In particular, if the function γ is set appropriately, then one can show that the γ-corrected loss function satisfies the condition of a Neyman orthogonal loss defined by Foster & Syrgkanis (2019). We begin our analysis by showing a general lemma for any estimator $\hat{f}$, which

adapts the main theorem of Foster & Syrgkanis (2019) to account for approximate orthogonality; the proof can be found in App. C.

Lemma 2 (Algorithm-agnostic analysis)**.** *Consider any estimation algorithm that produces an estimate $\hat{f}$ with small plug-in excess risk, i.e.,*

$$L_D(\hat{f},\hat{p},\gamma)-L_D(f_0,\hat{p},\gamma)\leq\epsilon(\hat{f},\hat{p},\gamma)$$

and that the loss L_D is σ-strongly convex with respect to f and $\mathcal{F}$ is a convex set. Then:

$$\tfrac{\sigma}{4}\|\hat{f}-f_0\|_{2,2}^2\leq\epsilon(\hat{f},\hat{p},\gamma)+\tfrac{1}{\sigma}\|(\gamma_{f_0,\hat{p}}-\gamma)^\top(\hat{p}-p_0)\|_{2,2}^2$$

If, further, $\sup_{z,\phi,\pi,i\in[d]}\|\nabla_{\phi_i\pi\pi}\ell(z;\phi,\pi)\|_{\text{op}}\leq M$*, then:*

$$\|(\gamma_{f_0,\hat{p}}-\gamma)^\top(\hat{p}-p_0)\|_{2,2}^2\leq 2\big(\|(q_{f_0,\hat{p}}-\gamma)^\top(\hat{p}-p_0)\|_{2,2}^2+M^2k\|\hat{p}-p_0\|_{2,4}^4\big)$$

Connection to Neyman orthogonality Note that if we set $\gamma=q_{f_0,\hat{p}}$, then we obtain that the MSE error of the student depends on the MSE error of the teacher only in a second-order manner! Setting $\gamma=q_{f_0,\hat{p}}$ renders the loss function Neyman orthogonal (Foster & Syrgkanis, 2019). Moreover, $q_{f_0,\hat{p}}$ is an observable quantity for any Bregman divergence loss (Definition 1) as $q_{f_0,\hat{p}}$ is independent of f_0. However, we note that this setting of the γ can lead to larger variance, i.e., the achievable excess risk can be much larger than the excess risk without the γ-correction. For instance, in the case of the squared error logit loss (1) $q_{f_0,\hat{p}}(x)=\frac{1}{\hat{p}(x)}$, which can be excessively large when $\hat{p}$ is close to 0, leading to a large increase in the variance of our loss. Thus, in a departure from the standard approach in semiparametric inference, we will be choosing γ in practice to balance bias and variance.

Example instantiation of student's estimation algorithm If we use *plug-in empirical risk minimization*, i.e., $\hat{f}=\text{argmin}_{f\in\mathcal{F}}L_n(f,\hat{p},\gamma)$, to estimate f_0 with $\hat{p}$ estimated on an independent sample, then the results of Maurer & Pontil (2009) directly imply that as long as the loss function $\ell(z;\phi,\pi)$ is uniformly bounded in $[-H,H]$, then w.p. $1-\delta$:

$$\epsilon(\hat{f},\hat{p},\gamma)\leq O\Big(\sqrt{\frac{\sup_{f\in\mathcal{F}}\text{Var}(\ell_\gamma(Z;f(X),\hat{p}(X)))\log(\tau(n)/\delta)}{n}}+\frac{H\log(\tau(n)/\delta)}{n}\Big)$$

where $\tau(n)=\mathcal{N}_\infty(1/n,\mathcal{F},2n)$ and $\mathcal{N}_\infty(\epsilon,\mathcal{F},m)$ is the ℓ_∞ empirical covering number of function class $\mathcal{F}$ in the worst-case over all realizations of m samples and at approximation level ϵ. This result has two drawbacks: it is a slow rate result that scales as $1/\sqrt{n}$ for parametric of bounded VC-dimension classes and it requires the student to be fit on a completely separate sample than the teacher. In the next theorem, we address both of these drawbacks: i) we invoke localized Rademacher complexity analysis to provide a fast rate result which would be of the order of $1/n$ for VC or parametric function classes, ii) we use a more sophisticated sample-splitting technique called cross-fitting, which allows the student to be trained on all the available data.

4.2 Combating teacher overfitting with cross-fitting

We now provide a more sophisticated version of sample splitting to make use of all datapoints in our final estimation, while at the same time not taking a hit by the sample complexity of the teacher's function space. This approach is referred to as *cross-fitting* in the semiparametric inference literature (see e.g., Chernozhukov et al. (2018)):

1. Partition the dataset into B equally sized partitions $P_1,\ldots,P_B$
2. For each fold $t\in[B]$ estimate $\hat{p}^{(t)}$ and $\hat{\gamma}^{(t)}$ using all the out-of-fold samples
3. Estimate $\hat{f}$ by minimizing the empirical loss:

$$\hat{f}=\text{argmin}_{f\in\mathcal{F}}\tfrac{1}{n}\textstyle\sum_{t=1}^B\sum_{i\in P_t}\ell_{\hat{\gamma}^{(t)}}(Z_i;f(X_i),\hat{p}^{(t)}(X_i))$$

In other words, the nuisance estimates $\hat{\gamma}^{(t)},\hat{p}^{(t)}$ that are evaluated on the samples in fold t, are estimated based only on samples outside of P_t.

Theorem 3 (Cross-fitted ERM analysis)**.** *Assume that $f_0\in\mathcal{F}$ and $\mathcal{F}$ is a convex set. Let $\delta_{n,\zeta}=\delta_n+c_0\sqrt{\frac{\log(c_1/\zeta)}{n}}$ for universal constants c_0,c_1 and δ_n an upper bound on the critical radii of the classes:*

$$\mathcal{G}(\hat{p}^{(t)},\hat{\gamma}^{(t)})=\{z\to r\big(\ell_{\hat{\gamma}^{(t)}}(z;f(x),\hat{p}^{(t)}(x))-\ell_{\hat{\gamma}^{(t)}}(z;f_0(x),\hat{p}^{(t)}(x))\big):f\in\mathcal{F},r\in[0,1]\}$$

for each $t\in[B]$. Assume that, for all $z\in\mathcal{Z}$, ϕ, and π, the loss $\ell(z;\phi,\pi)$ is uniformly bounded in $[-H,H]$ and σ-strongly convex with respect to ϕ, and let $\mu(z)=\sup_{f\in\mathcal{F},t\in[B]}\big\|\nabla_\phi\ell(z;f(X),\hat{p}^{(t)}(x))\big\|_2$. Moreover, suppose that the function class f satisfies the ℓ_2/ℓ_4 ratio condition: $\sup_{f\in\mathcal{F}}\frac{\|f-f_0\|_{2,4}}{\|f-f_0\|_{2,2}}\leq C$.

If $\hat{f}$ is the output of the cross-fitted empirical risk minimization algorithm on a fresh set of n samples, then, w.p. $1-\zeta$:

$$\frac{\sigma}{8}\|\hat{f}-f_0\|_{2,2}^2 \leq \frac{1}{\sigma}O\Big(\delta_{n/B,\zeta/B}^2 C^2 H^2\Big(\|\mu\|_4^2+\frac{1}{B}\sum_{t=1}^B\sqrt{\mathbb{E}\big[\|(Y-\hat{p}^{(t)}(X))^\top\hat{\gamma}^{(t)}(X)\|_2^4\big]}\Big)\Big) + \frac{1}{\sigma}O\big(\frac{1}{B}\sum_{t=1}^B\|(\gamma_{f_0,\hat{p}^{(t)}}-\hat{\gamma}^{(t)})^\top(\hat{p}^{(t)}-p_0)\|_{2,2}^2\big).$$

The proof is found in App. E. Observe that unlike Thm. 1, the function classes $\mathcal{G}(\hat{p}^{(t)},\hat{\gamma}^{(t)})$ in the latter theorem do not vary the teacher's model over $\mathcal{P}$ but rather only vary $f\in\mathcal{F}$ and evaluate p at the out-of-fold estimate $p^{(t)}$. Since in practice the teacher's model can be quite complex, getting rid of this dependence on the sample complexity of the teacher's function space can bring immense improvement. Thus the critical radius of $\mathcal{G}(\hat{p}^{(t)},\hat{\gamma}^{(t)})$, can be significantly smaller than $\mathcal{G}$. For instance, if $\mathcal{P}$ and $\mathcal{F}$ are VC-subgraph classes with VC dimensions $d_{\mathcal{P}}\gg d_{\mathcal{F}}$, correspondingly, then the critical radius of $\mathcal{G}(\hat{p}^{(t)},\hat{\gamma}^{(t)})$, will be of the order $\sqrt{d_{\mathcal{F}}\log(n)/n}$, while that of $\mathcal{G}$ of the order of $\sqrt{d_{\mathcal{P}}\log(n)/n}$.

Moreover, we can see in the bound of Thm. 3 the interplay between bias and variance introduced by γ. In particular, the part of the bound that depends on $\hat{\gamma}^{(t)}$ can be further simplified as:

$$\sqrt{\mathbb{E}[\delta_{n,\zeta}^4 C^4\big\|(Y-\hat{p}(X))^\top\hat{\gamma}^{(t)}(X)\big\|_2^4+\|(\gamma_{\hat{p},0}(X)-\hat{\gamma}^{(t)}(X))^\top(\hat{p}(X)-p_0(X))\|_2^4]} \tag{7}$$

where the first part corresponds to the bias increase due to γ, while the second part the bias decrease.

4.3 Biased stochastic gradient descent analysis

When the set of candidate prediction rules f_θ is parameterized by a vector $\theta\in\mathbb{R}^d$, we may alternatively fit θ via stochastic gradient descent (SGD) (Robbins & Monro, 1951; Bottou & Bousquet, 2008) on the γ-corrected objective $L_D(f_\theta,\hat{p},\hat{\gamma})$. With a minibatch size of 1 and a starting point θ_0, the parameter updates take the form

$$\theta_{t+1}=\theta_t-\eta_t\nabla_\theta f_\theta(X_t)^\top\nabla_\phi\ell_\gamma(W_t;f_\theta(X_t),p(X_t)) \quad \text{for} \quad t+1\in[n]. \tag{8}$$

Ideally, these updates would converge to a minimizer of the ideal risk $\mathcal{L}(\theta;p_0)=L_D(f_\theta,p_0)$. Our next result shows that, if the teacher $\hat{p}$ is independent of $(W_t)_{t\in[n]}$, then the SGD updates (8) have excess ideal risk governed by a bias term $\zeta(\hat{\gamma})$ and a variance term $\sigma(\hat{\gamma})^2/n$. Here, $\sigma_0^2(\theta)$ represents the baseline stochastic gradient variance that would be incurred if SGD were run directly on the ideal risk $\mathcal{L}(\theta;p_0)$ rather than our surrogate risk. Our proof in App. F builds upon the biased SGD bounds of Ajalloeian & Stich (2020).

Theorem 4 (Biased SGD analysis)**.** *Suppose that the loss $\mathcal{L}(\theta;p_0)$ is λ-strongly smooth in θ. Define the bias and root-variance parameters*

$$\zeta(\hat{\gamma})\triangleq\sup_{\theta\in\mathbb{R}^d}\big\|\nabla_\theta f_\theta^\top(\gamma_{f_\theta,\hat{p}}-\hat{\gamma})^\top(\hat{p}-p_0)\big\|_{2,2}$$

$$\sigma(\hat{\gamma})\triangleq\sup_{\theta\in\mathbb{R}^d}\sigma_0(\theta)+\sqrt{\mathbb{E}[\|\nabla_\theta f_\theta(X)^\top\gamma(X)^\top(Y-p_0(X))\|_2^2]+\big\|\nabla_\theta f_\theta^\top(\gamma_{f_\theta,\hat{p}}-\hat{\gamma})^\top(\hat{p}-p_0)\big\|_{2,2}^2}$$

for $\sigma_0^2(\theta)\triangleq\sum_{i\in[d]}\mathrm{Var}[\nabla_{\theta_i}\ell(W;\ f_\theta(X),\ p_0(X))]$ the unbiased SGD variance. If $F_0=\mathcal{L}(\theta_0;p_0)-\min_{\theta\in\mathbb{R}^d}\mathcal{L}(\theta;p_0)$, then the iterates $\{\theta_t\}_{t=1}^n$ of the γ-corrected SGD algorithm satisfy

$$\min_{t\in[n]}\mathbb{E}\big[\|\nabla_\theta\mathcal{L}(\theta_t;p_0)\|_2^2\big]=O\Big(\frac{\sigma(\hat{\gamma})\sqrt{\lambda F_0}}{\sqrt{n}}+\zeta^2(\hat{\gamma})\Big).$$

If, in addition, $\mathcal{L}(\theta;p_0)$ is μ-strongly convex in θ, then the iterates satisfy

$$\mathbb{E}[\mathcal{L}(\theta_n;p_0)-\min_{\theta\in\mathbb{R}^d}\mathcal{L}(\theta;p_0)]=\frac{1}{\mu}O\big(\frac{\lambda}{\mu}\frac{\sigma(\hat{\gamma})^2}{n}+\zeta^2(\hat{\gamma})\big)+O\big(F_0e^{-\frac{\mu}{2\lambda}n}\big).$$

Similar to Thm. 3, the bound in Thm. 4, portrays the interplay of bias and variance as $\hat{\gamma}$ ranges from 0 to $q_{f_\theta,\hat{p}}$ (which we remind that for any Bregman loss is independent of f_θ). In particular, the part of the bound for strongly convex losses that depends on $\hat{\gamma}$ can be further simplified to:

$$\mathbb{E}\Big[\Big(\frac{\lambda\|\hat{\gamma}(X)\|_2^2\|Y-p_0(X)\|_2^2}{\mu n}+\|(\gamma_{f_\theta,\hat{p}}(X)-\hat{\gamma}(X))^\top(\hat{p}(X)-p_0(X))\|_2^2\Big)\|\nabla_\theta f_{\hat{\theta}}(X)\|_2^2\Big] \tag{9}$$

This has a very intuitive form: the first term is the impact of $\hat{\gamma}(X)$ on the variance, which is also related to the square of the noise of y, divided by the standard error scaling. The second controls how $\hat{\gamma}$ improves the bias introduced by the error in the teacher's $\hat{p}$. Both the variance and the bias are re-weighted based on the gradient of θ, which expresses that some regions of X are more important in identifying the parameters and therefore, when we minimize the average bias-variane trade-off, we should be re-weighting the samples based on how informative they are about the parameters θ.

5 EXPERIMENTS

We empirically validate that the two components of our solution, cross-fitting and γ-correction, can address the two failure modes of KD (teacher overfitting/underfitting), in a variety of settings: (i) a synthetic setting that validates the theory in Secs. 3 and 4, (ii) tabular data showing that cross-fitting and γ-correction yield improvement, and (iii) image data to show the wide applicability of our methods. Throughout, we use the squared error logit loss (1). Additional details and results are in App. G.

Selecting the correction matrix $\hat{\gamma}$ Motivated by the analyses in Sec. 4, for each training point (x,y), we will select our correction matrix $\hat{\gamma}(x)$ to balance bias and variance by minimizing a pointwise upper bound on the γ-correction error (9) (ideally with a closed-form solution to avoid excessive computational overhead).[2] To eliminate dependence on the unobserved p_0, we observe that the bias term $\|(\gamma_{f_\theta,\hat{p}}(x) - \hat{\gamma}(x))^\top (\hat{p}(x) - p_0(x))\|_2^2 = O(\|q_{f_\theta,\hat{p}}(x) - \hat{\gamma}(x)\|_{\text{op}}^2)$ up to additive terms independent of $\hat{\gamma}$. We introduce a tunable hyperparameter $\alpha > 0$ to trade off between this bias bound and the variance term in (9) and select $\hat{\gamma}(x) = \text{diag}(v(x))$ to minimize

$$\|\hat{\gamma}(x)\|_{\text{op}}^2 \mathbb{E}[\|y - \hat{p}(x)\|_2^2 \,|\, x] + \alpha \|q_{f_\theta,\hat{p}}(x) - v(x)\|_{\text{op}}^2 = \|v(x)\|_2^2 \mathbb{E}[\|y - \hat{p}(x)\|_2^2 \,|\, x] + \alpha \|\tfrac{1}{\hat{p}(x)} - v(x)\|_2^2.$$

Since the conditional expectation involves the unknown quantity p_0, we approximate $\mathbb{E}[\|y - \hat{p}(x)\|_2^2 \,|\, x]$ by its sample $\|y - \hat{p}(x)\|_2^2$.[3] This objective is quadratic in $v(x)$ and thus has a closed-form solution. Moreover, once we solve for $\hat{\gamma}(x)$, the student's γ-corrected objective can be shown to be equivalent to a square loss with labels $\log(p(x)) + \hat{\gamma}(x)^\top (y - p(x))$.

Validation of theoretical analysis. When the Bayes probability $p_0(x)$ is known, we validate that the corrected loss function yields a student model whose accuracy is closer to the performance of the Bayes classifier, compared to a student model trained from scratch.

The synthetic data is generated from a mixture of two Gaussians with equal weights, where the Bayes classifier is known. The student model is a linear logistic regression model, which can capture the Bayes classifier. To simulate the teacher probabilities, we take the true probability $p_0(x)$ and move it closer to either 0 or 1, whichever is closer to $p_0(x)$. We compare 4 training objectives for the student model: (i) trained from scratch (ii) trained with vanilla KD (iii) trained with γ-corrected KD loss (iv) trained to match the Bayes probability $p_0(x)$. Figure 1 shows that the corrected KD loss brings the student's accuracy close to the case where it is trained on the true probabilities.

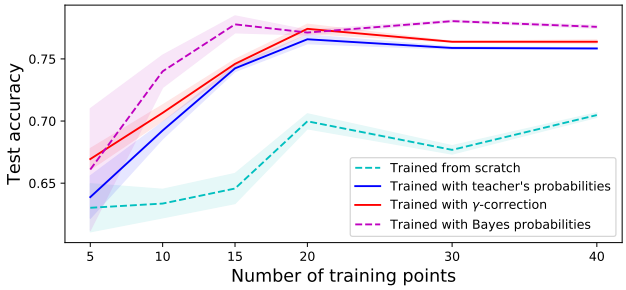


Figure 1: On synthetic data, γ-correction reduces student error due to teacher underfitting and more closely matches training with Bayes class probabilities p_0.

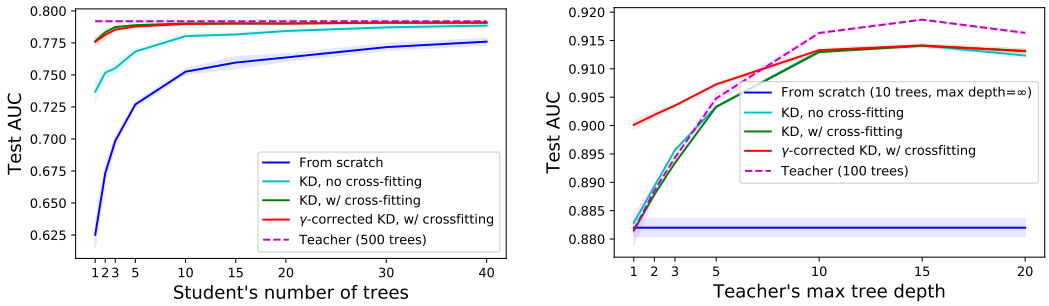
Tabular data. We validate cross-fitting and corrected KD loss on five real-world tabular datasets datasets: Adult (Dheeru & Karra Taniskidou, 2017), FICO (FIC), Higgs (Dheeru & Karra Taniskidou, 2017), MAGIC (Dheeru & Karra Taniskidou, 2017), and StumbleUpon (Eve; Liu et al., 2017). Cross-fitting can improve the student performance by up to 4% AUC points when the teacher model likely overfits, while γ-corrected KD loss is most helpful when the teacher has high bias.

We validate that cross-fitting can improve KD when the teacher overfits, and γ-correction is helpful when the teacher has high bias. Both teacher and student models are random forest. In the first case (Fig. 2a), we use a large teacher (500 trees) and vary the student's capacity, while in the second case (Fig. 2b) we vary the teacher's bias by limiting the teacher's max tree depth. Figure 2 shows the result of the first case on the FICO dataset, and the results of the second case on the Adult dataset. The results on the rest of the datasets are similar (App. G.2). When the teacher overfits, cross-fitting

[2] Balancing the bias and variance terms (7) of Thm. 3 yields a similar objective.

[3] An alternative that performs slightly worse is to approximate $\mathbb{E}[\|y - \hat{p}(x)\|^2 \,|\, x]$ by $\|\hat{p}(x)(1 - \hat{p}(x))\|_2^2$.

shows significant improvement over using the same teacher on the training points. The γ correction successfully reduces the teacher's bias and thus improves the student's performance. The effect is most pronounced when the teacher has large bias (low tree depth).



(a) FICO dataset, when teacher overfits

(b) Adult dataset, when teacher has high bias

Figure 2: Cross-fitting leads to better KD when the teacher overfits, while γ-correction improves student's performance when the teacher has high bias.

Image data. On the image classification dataset CIFAR-10 (Krizhevsky & Hinton, 2009), we show that cross-fitting can improve the student's performance in the case where the teacher overfits. We use ResNet-8 as the student and ResNet-14/20/32/44/56 (He et al., 2016) as the teacher. It has been observed that larger or deeper teachers might not yield better students, as the teacher might overfit to the training set (Cho & Hariharan, 2019; Müller et al., 2019). To simulate overfitting, we turn off data augmentation (random horizontal flipping and cropping). We compare students trained with vanilla KD objective with and without cross-fitting and γ-correction in Fig. 3. Cross-fitting can reduce the effect of teacher's overfitting, especially for deep teacher models. This effect is most evident in the test loss, where the student that is trained with vanilla KD has much higher loss than the student trained with KD and cross-fitting. The γ-correction consistently provides a small performance boost.

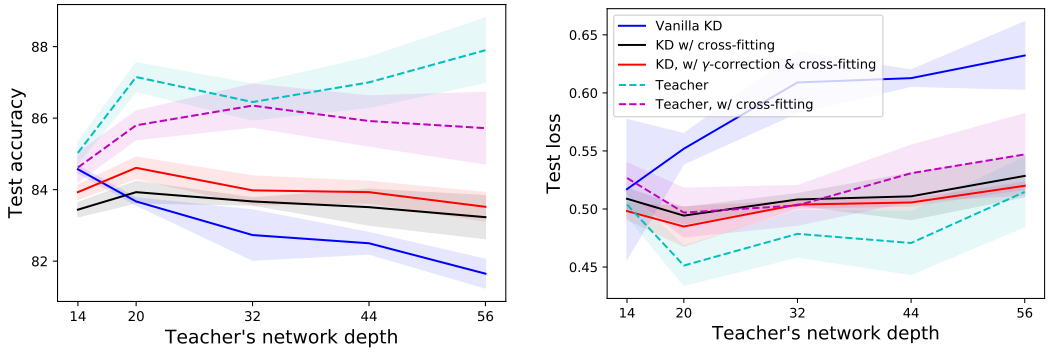


Figure 3: On CIFAR-10, cross-fitting reduces the effect of overfitting teacher, and γ-correction yields an additional small performance boost.

App. G.3 shows the effect of varying the hyperparameter α. Large α (closer to orthogonal loss) leads to high variance and lower test accuracy. Intermediate α improves on the vanilla KD objective ($\alpha = 0$).

6 Conclusion

We provide a new analysis of knowledge distillation under the lens of semiparametric inference. By viewing the KD process as learning a student model with the true class probabilities as a nuisance function and the teacher's probabilities as a plug-in estimator, we obtain generalization bounds on the student model. This perspective also suggests two techniques to improve KD, namely cross-fitting and γ-correction, which we show leads to better KD performance, both theoretically and empirically. This also opens the door to combining KD with other techniques and corrections in the semiparametric literature, such as those based on targeted maximum likelihood framework (Van Der Laan & Rubin, 2006). We are excited about future work on unraveling many surprising phenomena in KD, such as the success of self-distillation (Furlanello et al., 2018) and noisy student training (Xie et al., 2020).

REFERENCES

Stumbleupon evergreen dataset. https://www.kaggle.com/c/stumbleupon.

FICO: Explanable machine learning challenge. https://community.fico.com/s/explainable-machine-learning-challenge.

Ahmad Ajalloeian and Sebastian U Stich. Analysis of sgd with biased gradient estimators. *arXiv preprint arXiv:2008.00051*, 2020.

Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? In *Advances in neural information processing systems*, pp. 2654–2662, 2014.

Mikhail Belkin, Alexander Rakhlin, and Alexandre B Tsybakov. Does data interpolation contradict statistical optimality? In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1611–1619. PMLR, 2019.

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

S. Bernstein. The theory of probabilities. *Gastehizdat Publishing House*, 1946.

Léon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. In *Advances in neural information processing systems*, pp. 161–168, 2008.

Leo Breiman and Nong Shang. Born again trees. *University of California, Berkeley, Berkeley, CA, Technical Report*, 1:2, 1996.

Yuheng Bu, Weihao Gao, Shaofeng Zou, and Venugopal V Veeravalli. Information-theoretic understanding of population risk improvement with model compression. In *AAAI*, pp. 3300–3307, 2020.

Cristian Bucila, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, August 20-23, 2006*, pp. 535–541, 2006. doi: 10.1145/1150402.1150464.

Yevgen Chebotar and Austin Waters. Distilling knowledge from ensembles of neural networks for speech recognition. In *Interspeech*, pp. 3439–3443, 2016.

Wei-Chun Chen, Chia-Che Chang, and Che-Rung Lee. Knowledge distillation with feature maps for image classification. In *Asian Conference on Computer Vision*, pp. 200–215. Springer, 2018.

Xu Cheng, Zhefan Rao, Yilan Chen, and Quanshi Zhang. Explaining knowledge distillation by quantifying the knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12925–12935, 2020.

Victor Chernozhukov, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K. Newey, and James M. Robins. Locally Robust Semiparametric Estimation. *arXiv e-prints*, art. arXiv:1608.00033, July 2016.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018. doi: 10.1111/ectj.12097. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ectj.12097.

Victor Chernozhukov, Whitney Newey, and Rahul Singh. De-Biased Machine Learning of Global and Local Parameters Using Regularized Riesz Representers. *arXiv e-prints*, art. arXiv:1802.08667, February 2018.

Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4794–4802, 2019.

Mark Craven and Jude W Shavlik. Extracting tree-structured representations of trained networks. In *Advances in neural information processing systems*, pp. 24–30, 1996.

Dua Dheeru and Efi Karra Taniskidou. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Rasool Fakoor, Jonas Mueller, Nick Erickson, Pratik Chaudhari, and Alexander J Smola. Fast, accurate, and simple models for tabular data via augmented distillation. *arXiv preprint arXiv:2006.14284*, 2020.

Dylan J Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *arXiv preprint arXiv:1901.09036*, 2019.

Markus Freitag, Yaser Al-Onaizan, and Baskaran Sankaran. Ensemble distillation for neural machine translation. *arXiv preprint arXiv:1702.01802*, 2017.

Tommaso Furlanello, Zachary C Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. *arXiv preprint arXiv:1805.04770*, 2018.

Yotam Gil, Yoav Chai, Or Gorodissky, and Jonathan Berant. White-to-black: Efficient distillation of black-box adversarial attacks. *arXiv preprint arXiv:1904.02405*, 2019.

Micah Goldblum, Liam Fowl, Soheil Feizi, and Tom Goldstein. Adversarially robust distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 3996–4003, 2020.

Jianping Gou, Baosheng Yu, Stephen John Maybank, and Dacheng Tao. Knowledge distillation: A survey. *arXiv preprint arXiv:2006.05525*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Minghao Hu, Yuxing Peng, Furu Wei, Zhen Huang, Dongsheng Li, Nan Yang, and Ming Zhou. Attention-guided answer distillation for machine reading comprehension. *arXiv preprint arXiv:1808.07644*, 2018.

Michael R Kosorok. *Introduction to empirical processes and semiparametric inference*. Springer Science & Business Media, 2007.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Jinyu Li, Rui Zhao, Jui-Ting Huang, and Yifan Gong. Learning small-size dnn with output-distribution-based criteria. In *Fifteenth annual conference of the international speech communication association*, 2014.

Quanquan Li, Shengying Jin, and Junjie Yan. Mimicking very efficient network for object detection. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pp. 6356–6364, 2017.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

Ruishan Liu, Nicolo Fusi, and Lester Mackey. Teacher-student compression with generative adversarial networks. *arXiv preprint arXiv:1812.02271*, 2018.

Yu Liu, Hantian Zhang, Luyuan Zeng, Wentao Wu, and Ce Zhang. MLBench: How good are machine learning clouds for binary classification tasks on structured data. *ArXiv e-prints*, 2017.

Raphael Gontijo Lopes, Stefano Fenu, and Thad Starner. Data-free knowledge distillation for deep neural networks. *arXiv preprint arXiv:1710.07535*, 2017.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. *arXiv preprint arXiv:1511.03643*, 2015.

Liang Lu, Michelle Guo, and Steve Renals. Knowledge distillation for small-footprint highway networks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4820–4824. IEEE, 2017.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

Aditya Krishna Menon, Ankit Singh Rawat, Sashank J Reddi, Seungyeon Kim, and Sanjiv Kumar. Why distillation helps: a statistical perspective. *arXiv preprint arXiv:2005.10419*, 2020.

Hossein Mobahi, Mehrdad Farajtabar, and Peter L Bartlett. Self-distillation amplifies regularization in hilbert space. *arXiv preprint arXiv:2002.05715*, 2020.

Lili Mou, Ran Jia, Yan Xu, Ge Li, Lu Zhang, and Zhi Jin. Distilling word embeddings: An encoding approach. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pp. 1977–1980, 2016.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems*, pp. 4694–4703, 2019.

Elizbar A Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1): 141–142, 1964.

Ndapandula Nakashole and Raphael Flauger. Knowledge distillation for bilingual dictionary induction. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pp. 2497–2506, 2017.

Denis Nekipelov, Vira Semenova, and Vasilis Syrgkanis. Regularized Orthogonal Machine Learning for Nonlinear Semiparametric Models. *arXiv e-prints*, art. arXiv:1806.04823, June 2018.

Aaron Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George Driessche, Edward Lockhart, Luis Cobo, Florian Stimberg, et al. Parallel wavenet: Fast high-fidelity speech synthesis. In *International conference on machine learning*, pp. 3918–3926. PMLR, 2018.

Nicolas Papernot, Martín Abadi, Ulfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. *arXiv preprint arXiv:1610.05755*, 2016a.

Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE Symposium on Security and Privacy (SP)*, pp. 582–597. IEEE, 2016b.

Mary Phuong and Christoph Lampert. Towards understanding knowledge distillation. In *International Conference on Machine Learning*, pp. 5142–5151, 2019.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *Ann. Math. Statist.*, 22(3):400–407, 09 1951. doi: 10.1214/aoms/1177729586. URL https://doi.org/10.1214/aoms/1177729586.

Andrew Slavin Ross and Finale Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. *arXiv preprint arXiv:1711.09404*, 2017.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Peng Shen, Xugang Lu, Sheng Li, and Hisashi Kawai. Feature representation of short utterances based on knowledge distillation for spoken language identification. In *Interspeech*, pp. 1813–1817, 2018.

Pierre Stock, Armand Joulin, Rémi Gribonval, Benjamin Graham, and Hervé Jégou. And the bit goes down: Revisiting the quantization of neural networks. 2020.

Sarah Tan, Rich Caruana, Giles Hooker, Paul Koch, and Albert Gordo. Learning global additive explanations for neural nets using model distillation. *arXiv preprint arXiv:1801.08640*, 2018.

Jiaxi Tang, Rakesh Shivanna, Zhe Zhao, Dong Lin, Anima Singh, Ed H Chi, and Sagar Jain. Understanding and improving knowledge distillation. *arXiv preprint arXiv:2002.03532*, 2020.

Anastasios Tsiatis. *Semiparametric theory and missing data*. Springer Science & Business Media, 2007.

Mark J Van Der Laan and Daniel Rubin. Targeted maximum likelihood learning. *The international journal of biostatistics*, 2(1), 2006.

Vladimir Vapnik and Rauf Izmailov. Learning using privileged information: similarity control and knowledge transfer. *J. Mach. Learn. Res.*, 16(1):2023–2049, 2015.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. doi: 10.1017/9781108627771.

Chong Wang, Xipeng Lan, and Yangang Zhang. Model distillation with knowledge transfer from face classification to alignment and verification. *arXiv preprint arXiv:1709.02929*, 2017.

Ji Wang, Weidong Bao, Lichao Sun, Xiaomin Zhu, Bokai Cao, and S Yu Philip. Private model compression via knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1190–1197, 2019.

Shinji Watanabe, Takaaki Hori, Jonathan Le Roux, and John R Hershey. Student-teacher network learning with enhanced features. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5275–5279. IEEE, 2017.

Geoffrey S Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 359–372, 1964.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10687–10698, 2020.

A ADDITIONAL LITERATURE REVIEW

There has been a lot of work on knowledge distillation in the last half of a decade. Gou et al. (2020) gives a survey of the various directions that have been studied in this regard. In this appendix, we sketch some directions. In addition to references discussing theoretic aspects discussed in Sec. 1, here we mention some work empirically studying why distillation works. Cho & Hariharan (2019) show that larger teacher models do not necessarily improve the performance of student models as parsimonious student models are not able to mimic the teacher model. They suggest early stopping in training large teacher neural networks as means of regularizing. Cheng et al. (2020) show that when applied to image data, distillation allows the student neural net to learn multiple visual concepts simultaneously, while in knowledge distillation neural networks learn concepts sequentially.

Knowledge distillation has been used in various other settings such as adversarial attacks (Papernot et al., 2016b; Ross & Doshi-Velez, 2017; Gil et al., 2019; Goldblum et al., 2020), data security (Papernot et al., 2016a; Lopes et al., 2017; Wang et al., 2019), image processing (Li & Hoiem, 2017; Wang et al., 2017; Chen et al., 2018; Li et al., 2017), natural language processing (Nakashole & Flauger, 2017; Mou et al., 2016; Hu et al., 2018; Freitag et al., 2017), speech processing (Chebotar & Waters, 2016; Lu et al., 2017; Watanabe et al., 2017; Oord et al., 2018; Shen et al., 2018).

B GLOSSARY

Table 1: Glossary of notation

Notation	Definition
$\ell(Z;f(X),p_0(X)$	Loss function on a random data point
Population risk $L_D(f,p)$	$\mathbb{E}[\ell(Z;f(X),p(X))]$
Empirical risk $L_n(f,p)$	$\mathbb{E}_n[\ell(Z;f(X),p(X))]$
Population optimal student model f_0	$\operatorname{argmin}_{f\in\mathcal{F}} L_D(f,p_0)$
Empirical optimal student model $\hat{f}$	$\operatorname{argmin}_{f\in\mathcal{F}} L_n(f,\hat{p})$
$\|f\|_{p,q}$	$\|\|f(X)\|_p\|_{L^q} = \mathbb{E}\left[\|f(X)\|_p^q\right]^{1/q})$
∇_ϕ	Partial derivative of $\ell(z,\phi,\pi)$ wrt to the second input
∇_π	Partial derivative of $\ell(z,\phi,\pi)$ wrt to the third input
$\nabla_{\phi\pi}$	$[\nabla_{\phi\pi}\ell(z;\phi,\pi)]_{i,j} = \frac{\partial^2}{\partial\phi_j\partial\pi_i}\ell(z;\phi,\pi)$
$q_{f,p}(x)$	$\mathbb{E}[\nabla_{\phi\pi}\ell(Z;f(X),p(X))\,\|\,X=x]$
$\gamma_{f,p}(x)$	$\mathbb{E}_{U\sim\text{Unif}([0,1])}[q_{f,Up+(1-U)p_0}(x)]$
$\mathcal{R}(\delta;\mathcal{F})$	Localized Rademacher complexity of function class $\mathcal{F}$
δ_n	Critical radius
γ-corrected loss $\ell_\gamma(z;f(x),p(x))$	$\ell(z;f(x),p(x))+(y-p(x))^\top\gamma(x)f(x)$
Population γ-risk $L_D(f,p,\gamma)$	$\mathbb{E}[\ell_\gamma(Z;f(X),p(X))]$
Empirical γ-risk $L_n(f,p,\gamma)$	$\mathbb{E}_n[\ell_\gamma(Z;f(X),p(X))]$

C PROOF OF LEMMA 2: ALGORITHM-AGNOSTIC ANALYSIS

First we define for any functional $L(f)$ the Frechet derivative as:

$$D_f L(f)[\nu] = \frac{\partial}{\partial t} L(f+t\nu)\,|_{t=0}$$

When L is an operator of the form: $\mathbb{E}[g(f(X))]$, then: $D_f L(f)[\nu] = \mathbb{E}[\nabla g(f(X))^\top \nu(X)]$.

By σ-strong convexity of L_D,[4] we have that

$$L_D(\hat{f},\hat{p},\gamma) \geq L_D(f_0,\hat{p},\gamma) + D_f L_D(f_0,\hat{p},\gamma)[\hat{f}-f_0] + \frac{\sigma}{2}\|\hat{f}-f_0\|_{2,2}^2.$$

[4]Notably this strong convexity assumption can be relaxed to $\mathbb{E}\left[\nabla_\phi \ell(W;f_0(X),p_0(X))(\hat{f}(X)-f_0(X)\right] \geq 0$.

Furthermore, our excess risk assumption and the optimality of f_0 give us

$$\frac{\sigma}{2}\|\hat{f}-f_0\|_{2,2}^2 \leq \underbrace{L_D(\hat{f},\hat{p},\gamma)-L_D(f_0,\hat{p},\gamma)}_{\text{excess risk of } \hat{f}} - D_f L_D(f_0,\hat{p},\gamma)[\hat{f}-f_0]$$

$$\overset{(a)}{\leq} \epsilon(\hat{f},\hat{p},\gamma) - \underbrace{D_f L_D(f_0,p_0,\gamma)[\hat{f}-f_0]}_{\geq 0 \text{ by optimality of } f_0} + D_f(L_D(f_0,p_0,\gamma)-L_D(f_0,\hat{p},\gamma))[\hat{f}-f_0].$$

By Taylor's theorem with integral remainder,

$$\begin{aligned}&\mathbb{E}[\langle \nabla_\phi \ell(W;f_0(x),p_0(x)) - \nabla_\phi \ell(W;f_0(x),\hat{p}(x)), \hat{f}(x)-f_0(x)\rangle \,|\, X=x] \\ &= (p_0(x)-\hat{p}(x))^\top \gamma_{f_0,\hat{p}}(x)(\hat{f}(x)-f_0(x))\end{aligned} \tag{10}$$

whenever $\nabla_{\phi\pi}\ell$ is well-defined. We can now invoke the expansion (10) and Cauchy-Schwarz to obtain the bound

$$\begin{aligned}&D_f(L_D(f_0,p_0,\gamma)-L_D(f_0,\hat{p},\gamma))[\hat{f}-f_0] \\ &= \mathbb{E}[\langle \nabla_\phi \ell(W;f_0(X),p_0(X)) - \nabla_\phi \ell(W;f_0(X),p(X)), \hat{f}(X)-f_0(X)\rangle] \\ &\quad - \mathbb{E}[(p_0(X)-\hat{p}(X))^\top \gamma(X)(\hat{f}(X)-f_0(X))] \\ &= \mathbb{E}[(p_0(X)-\hat{p}(X))^\top (\gamma_{f_0,\hat{p}}(X)-\gamma(X))(\hat{f}(X)-f_0(X))] \\ &\leq \mathbb{E}[\|(p_0(X)-\hat{p}(X))^\top (\gamma_{f_0,\hat{p}}(X)-\gamma(X))\|_2 \|\hat{f}(X)-f_0(X)\|_2] \\ &\leq \|(p_0-\hat{p})^\top(\gamma_{f_0,\hat{p}}-\gamma)\|_{2,2}\|\hat{f}-f_0\|_{2,2}\end{aligned}$$

Thus combining all the above inequalities:

$$\frac{\sigma}{2}\|\hat{f}-f_0\|_{2,2}^2 \leq \epsilon(\hat{f},\hat{p},\gamma) + \|(\hat{p}-p_0)^\top(\gamma_{f_0,\hat{p}}-\gamma)\|_{2,2}\|\hat{f}-f_0\|_{2,2}$$

By an AM-GM inequality, for all $a,b\geq 0$: $a\cdot b \leq \frac{1}{2}(\frac{2}{\sigma}a^2+\frac{\sigma}{2}b^2)$. Applying this to the product of norms on the RHS and re-arranging yields

$$\frac{\sigma}{4}\|\hat{f}-f_0\|_{2,2}^2 \leq \epsilon(\hat{f},\hat{p},\gamma) + \frac{1}{\sigma}\|(\hat{p}-p_0)^\top(\gamma_{f_0,\hat{p}}-\gamma)\|_{2,2}^2.$$

To get the final ineguality, observe that:

$$\|(\hat{p}-p_0)^\top(\gamma_{f_0,\hat{p}}-\gamma)\|_{2,2}^2 \leq 2\|(\hat{p}-p_0)^\top(q_{f_0,\hat{p}}-\gamma)\|_{2,2}^2 + 2\|(\hat{p}-p_0)^\top(\gamma_{f_0,\hat{p}}-q_{f_0,\hat{p}})\|_{2,2}^2$$

Moreover, by the boundedness of the third derivative, we have:

$$\begin{aligned}\|(\hat{p}-p_0)^\top(\gamma_{f_0,\hat{p}}-q_{f_0,\hat{p}})\|_{2,2}^2 &\leq \mathbb{E}[\|\hat{p}(X)-p_0(X)\|_2^2\|\gamma_{f_0,\hat{p}}(X)-q_{f_0,\hat{p}}(X)\|_2^2] \\ &\leq \mathbb{E}[\|\hat{p}(X)-p_0(X)\|_2^2 M^2 k\|\hat{p}(X)-p_0(X)\|_2^2] \\ &\leq M^2 k\|\hat{p}-p_0\|_{2,4}^4\end{aligned}$$

Combining all the above yields the final bound.

D Proof of Theorem 1

Proof First by localized Rademacher analysis (see Lemma 11 of Foster & Syrgkanis (2019)) if δ_n upper bounds the critical radius of the function class $\mathcal{G}$ and if we denote with $\ell_{\hat{p},f}(w)=\ell(w;f(x),\hat{p}(x))$, then w.p. $1-\zeta$:[5]

$$\left|L_n(\hat{f},\hat{p})-L_n(f_0,\hat{p})-(L_D(\hat{f},\hat{p})-L_D(f_0,\hat{p}))\right| \leq O\Big(\delta_{n,\zeta}\|\ell_{\hat{p},\hat{f}}-\ell_{\hat{p},f_0}\|_{2,2}+\delta_{n,\zeta}^2\Big)$$

Moreover, if we let $\mu(w)=\sup_\phi\|\nabla_\phi \ell(w;\phi,\hat{p}(x))\|_2$, then we have by Cauchy-Scwarz inequality:

$$\|\ell_{\hat{p},f}-\ell_{\hat{p},f_0}\|_{2,2} \leq \|\mu\|_4\|f-f_0\|_{2,4}$$

If we further assume that the function class $\mathcal{F}$ satisfies an ℓ_2/ℓ_4 condition that:

$$\sup_{f\in\mathcal{F}}\frac{\|f-f_0\|_{2,4}}{\|f-f_0\|_{2,2}} \leq C$$

then:

$$\epsilon(\hat{f},\hat{p},\gamma) \leq O\big(\delta_{n,\zeta}C\|\mu\|_4\|f-f_0\|_{2,2}+\delta_{n,\zeta}^2\big)$$

Plugging in the bound above to Lemma 2 (which holds irrespective of data re-use or sample splitting or cross-fitting) and applying again the AM-GM inequality yields:

$$\frac{\sigma}{8}\|\hat{f}-f_0\|_{2,2}^2 \leq \frac{1}{\sigma}O\big(\delta_{n,\zeta}^2 C^2\|\mu\|_4^2+\|\gamma_{\hat{p},0}^\top(\hat{p}-p_0)\|_{2,2}^2\big)$$

[5] We apply the lemma with $\mathcal{L}_g=g$ and $g\in\mathcal{G}$ and $g^*=0$. Then we instantiate the concentration inequality with $g=\ell_{\hat{p},\hat{f}}-\ell_{\hat{p},f_0}\in\mathcal{G}$.

□

E PROOF OF THM. 3: CROSS-FITTED ERM ANALYSIS

Let $L_{n,t}$ denote the empirical loss over the samples in the t-th fold and $\hat{p}^{(t)},\hat{\gamma}^{(t)}$ the nuisance functions used on the samples in the k-th fold. For any $t\in[K]$ and conditional on $\hat{p}^{(t)},\hat{\gamma}^{(t)}$, suppose that δ_n upper bounds the critical radius of the function class $\mathcal{G}(\hat{p}^{(t)},\hat{\gamma}^{(t)})$, then by Lemma 11 of Foster & Syrgkanis (2019),[6] if we denote with $\ell_{t,f}(w)=\ell_{\hat{\gamma}^{(t)}}(w;f(x),\hat{p}^{(t)}(x))$, w.p. $1-\zeta$:

$$\left|L_{n,t}(\hat{f},\hat{p}^{(t)},\hat{\gamma}^{(t)})-L_{n,t}(f_0,\hat{p}^{(t)},\hat{\gamma}^{(t)})-(L_D(\hat{f},\hat{p}^{(t)},\hat{\gamma}^{(t)})-L_D(f_0,\hat{p}^{(t)},\hat{\gamma}^{(t)}))\right|\leq O\left(\delta_{n/B,\zeta}\|\ell_{t,f}-\ell_{t,f_0}\|_{2,2}+\delta^2_{n/B,\zeta}\right)$$

Moreover, we have that by the definition of cross-fitted ERM:

$$\frac{1}{B}\sum_{t=1}^{B}L_{n,t}(\hat{f},\hat{p}^{(t)},\hat{\gamma}^{(t)})-L_{n,t}(f_0,\hat{p}^{(t)},\hat{\gamma}^{(t)})\leq 0$$

Thus we have that w.p. $1-\zeta B$:

$$\frac{1}{B}\sum_{t=1}^{B}L_D(\hat{f},\hat{p}^{(t)},\hat{\gamma}^{(t)})-L_D(f_0,\hat{p}^{(t)},\hat{\gamma}^{(t)})\leq O\left(\delta_{n/B,\zeta}\frac{1}{B}\sum_{t=1}^{B}\|\ell_{t,f}-\ell_{t,f_0}\|_{2,2}+\delta^2_{n/B,\zeta}\right)$$

Moreover, if we let $\mu(w)=\sup_{\phi,t}\left\|\nabla_\phi\ell(w;\phi,\hat{p}^{(t)}(x))\right\|_2$, then we have by Cauchy-Schwarz inequality:

$$\begin{aligned}\|\ell_{t,f}-\ell_{t,f_0}\|_{2,2}&\leq\|\mu\|_4\|f-f_0\|_{2,4}+\sqrt{\mathbb{E}\left[\left((Y-\hat{p}^{(t)}(X))^\top\hat{\gamma}^{(t)}(X)(f(X)-f_0(X))\right)^2\right]}\\&\leq\|\mu\|_4\|f-f_0\|_{2,4}+\mathbb{E}\left[\left\|(Y-\hat{p}^{(t)}(X))^\top\hat{\gamma}^{(t)}(X)\right\|_2^2\|f(X)-f_0(X))\|^2\right]\\&\leq\left(\|\mu\|_4+\mathbb{E}\left[\left\|(Y-\hat{p}^{(t)}(X))^\top\hat{\gamma}^{(t)}(X)\right\|_2^4\right]^{1/4}\right)\|f-f_0\|_{2,4}\end{aligned}$$

If we further assume that the function class $\mathcal{F}$ satisfies an ℓ_2/ℓ_4 condition that:

$$\sup_{f\in\mathcal{F}}\frac{\|f-f_0\|_{2,4}}{\|f-f_0\|_{2,2}}\leq C$$

then w.p. $1-\zeta$:

$$\frac{1}{B}\sum_{t=1}^{B}\epsilon(\hat{f},\hat{p}^{(t)},\hat{\gamma}^{(t)})\leq O\left(\delta_{n/B,\zeta/B}\frac{1}{B}\sum_{t=1}^{B}C\left(\|\mu_p\|_4+\mathbb{E}\left[\left\|(Y-\hat{p}(X))^\top\gamma(X)\right\|_2^4\right]^{1/4}\right)\|f-f_0\|_{2,2}+\delta^2_{n/B,\zeta/B}\right).$$

Applying Lemma 2 for any $\hat{p}^{(t)},\hat{\gamma}^{(t)}$ and averaging the final inequality we get:

$$\frac{\sigma}{4}\|\hat{f}-f_0\|_{2,2}^2\leq\frac{1}{B}\sum_{t=1}^{B}\left(\epsilon(\hat{f},\hat{p}^{(t)},\hat{\gamma}^{(t)})+\frac{1}{\sigma}\|(\gamma_{f_0,\hat{p}^{(t)}}-\hat{\gamma}^{(t)})^\top(\hat{p}^{(t)}-p_0)\|_{2,2}^2\right).$$

Plugging in the bound above to Lemma 2 and applying the AM-GM inequality and Jensen's inequality, yields:

$$\begin{aligned}\frac{\sigma}{8}\|\hat{f}-f_0\|_{2,2}^2\leq&\frac{1}{\sigma}O\left(\delta^2_{n/B,\zeta/B}C^2\left(\|\mu\|_4^2+\frac{1}{B}\sum_{t=1}^{B}\sqrt{\mathbb{E}\left[\left\|(Y-\hat{p}^{(t)}(X))^\top\hat{\gamma}^{(t)}(X)\right\|_2^4\right]}\right)\right)\\&+\frac{1}{\sigma}O\left(\frac{1}{B}\sum_{t=1}^{B}\|(\gamma_{f_0,\hat{p}^{(t)}}-\hat{\gamma}^{(t)})^\top(\hat{p}^{(t)}-p_0)\|_{2,2}^2\right).\end{aligned}$$

F PROOF OF THM. 4: BIASED SGD ANALYSIS

Below, for any integer s, we define the operator norm of any vector $v\in\mathbb{R}^s$ and any tensor T operating on $\mathbb{R}^s$ as

$$\|v\|_{\text{op}}\triangleq\|v\|_2\quad\text{and}\quad\|T\|_{\text{op}}\triangleq\sup_{v:\|v\|_2=1}\|T[v]\|_{\text{op}}.$$

Recall the definition

$$\begin{aligned}\nabla(W;\theta,p,\gamma)&=\nabla_\theta f_\theta(X)^\top\nabla_\phi\ell_\gamma(W;f_\theta(X),p(X))\\&=\nabla_\theta f_\theta(X)^\top(\nabla_\phi\ell(W;f_\theta(X),p(X))+\gamma(X)^\top(Y-p(X))).\end{aligned}$$

[6] We apply the lemma with $\mathcal{L}_g=g$ and $g\in\mathcal{G}(\hat{p}^{(t)},\hat{\gamma}^{(t)})$ and $g^*=0$. Then we instantiate the concentration inequality with $g=\ell_{t,\hat{f}}-\ell_{t,f_0}\in\mathcal{G}(\hat{p}^{(t)},\hat{\gamma}^{(t)})$.

Observe that since $\mathbb{E}[Y \mid X=x]=p_0(x)$, we can write for any γ:

$$\mathcal{L}(\theta;p_0)=\mathbb{E}[\ell(W;f_\theta(X),p_0(X))+(Y-p_0(X))^\top\gamma(X)f_\theta(X)]=\mathbb{E}[\ell_\gamma(W;f_\theta(X),p_0(X))]$$

Thus we also have that:

$$\forall\theta,\gamma:\nabla_\theta\mathcal{L}(\theta;p_0)=\mathbb{E}[\nabla(W;\theta,p_0,\gamma)]$$

Given this observation, we can decompose the gradient that is used in our SGD algorithm into a bias and variance component, when viewed from the perspective of a biased SGD algorithm for the population oracle loss:

$$\begin{aligned}\nabla(W;\theta,p,\gamma)=&\ \nabla_\theta\mathcal{L}(\theta;p_0)\\ &+\underbrace{\mathbb{E}[\nabla(W;\theta,p,\gamma)]-\mathbb{E}[\nabla(W;\theta,p_0,\gamma)]}_{\mathbf{b}(\theta,p,\gamma)}+\underbrace{\nabla(W;\theta,p,\gamma)-\mathbb{E}[\nabla(W;\theta,p,\gamma)]}_{\mathbf{n}(W;\theta,p,\gamma)}\end{aligned}$$

The following two lemmas bound the gradient bias and noise terms.

Lemma 5 (Gradient bias)**.** *If* $\sup_{x,\phi,\pi}\|\mathbb{E}[\nabla_{\pi\pi\phi}\ell(W;\phi,\pi)\mid X=x]\|_{\mathrm{op}}\le M$*, then for any parameter vector* θ *and functions* p *and* γ*, we have:*

$$\begin{aligned}\mathbf{b}(\theta,p,\gamma)&=\mathbb{E}[\nabla_\theta f_\theta(X)^\top(\gamma_{f_\theta,p}(X)-\gamma(X))^\top(p(X)-p_0(X))],\\ \|\mathbf{b}(\theta,p,\gamma)\|_2&\le\left\|\nabla_\theta f_\theta^\top(\gamma_{f_\theta,p}-\gamma)^\top(p-p_0)\right\|_{2,2},\quad \textit{and}\\ \|\mathbf{b}(\theta,p,\gamma)\|_2&\le\left\|\nabla_\theta f_\theta^\top(q_{f_\theta,p}-\gamma)^\top(p-p_0)\right\|_{2,2}+\frac{M}{2}\|\nabla_\theta f_\theta\|_{F,2}\|p-p_0\|_{2,4}^2.\end{aligned}$$

Proof By Taylor's theorem with integral remainder and Lagrange remainder respectively the SGD bias for each parameter i takes the form

$$\begin{aligned}\mathbf{b}_i(\theta,p,\gamma)=&\ \mathbb{E}[\nabla_i(W;\theta,p,\gamma)]-\mathbb{E}[\nabla_i(W;\theta,p_0,\gamma)]\\ =&\ \mathbb{E}[\langle\nabla_\phi\ell_\gamma(W;f_\theta(X),p(X))-\nabla_\phi\ell_\gamma(W;f_\theta(X),p_0(X)),\nabla_{\theta_i}f_\theta(X)\rangle]\\ =&\ \mathbb{E}[(p(X)-p_0(X))^\top(\gamma_{f_\theta,p}(X)-\gamma(X))\nabla_{\theta_i}f_\theta(X)]\\ =&\ \mathbb{E}[(p(X)-p_0(X))^\top(q_{f_\theta,p}(X)-\gamma(X))\nabla_{\theta_i}f_\theta(X)]\\ &+\frac{1}{2}\mathbb{E}[\nabla_{\pi\pi\phi}\ell(W;f_\theta(X),\bar{p}(X))[\nabla_{\theta_i}f_\theta(X),p(X)-p_0(X),p(X)-p_0(X)]].\end{aligned}$$

Furthermore, our operator norm assumption and Cauchy-Schwarz imply

$$\begin{aligned}|\mathbf{b}_i(\theta,p,\gamma)|&\le|\mathbb{E}[(p(X)-p_0(X))^\top(q_{f_\theta,p}(X)-\gamma(X))\nabla_{\theta_i}f_\theta(X)]|+\frac{M}{2}\mathbb{E}\Big[\|\nabla_{\theta_i}f_\theta(X)\|_2\|p(X)-p_0(X)\|_2^2\Big]\\ &\le|\mathbb{E}[(p(X)-p_0(X))^\top(q_{f_\theta,p}(X)-\gamma(X))\nabla_{\theta_i}f_\theta(X)]|+\frac{M}{2}\|\nabla_{\theta_i}f_\theta\|_{2,2}\|p-p_0\|_{2,4}^2.\end{aligned}$$

Thus, by the triangle inequality and Jensen's inequality we find that

$$\begin{aligned}\|\mathbf{b}(\theta,p,\gamma)\|_2&\le\left\|\nabla_\theta f_\theta^\top(\gamma_{f_\theta,p}-\gamma)^\top(p-p_0)\right\|_{2,2}\quad\text{and}\\ \|\mathbf{b}(\theta,p,\gamma)\|_2&\le\left\|\nabla_\theta f_\theta^\top(q_{f_\theta,p}-\gamma)^\top(p-p_0)\right\|_{2,2}+\frac{M}{2}\|\nabla_\theta f_\theta\|_{F,2}\|p-p_0\|_{2,4}^2.\end{aligned}$$

□

Lemma 6 (Gradient Variance)**.** *Define For any parameter* θ *and functions* p *and* γ*,*

$$\sqrt{\mathbb{E}[\|\mathbf{n}(W;\theta,p,\gamma)\|_2^2]}\le\sigma_0(\theta)+\sqrt{\mathbb{E}[\|\nabla_\theta f_\theta(X)^\top\gamma(X)^\top(Y-p_0(X))\|_2^2]+\left\|\nabla_\theta f_\theta^\top(\gamma_{f_\theta,p}-\gamma)^\top(p-p_0)\right\|_{2,2}^2}.$$

Proof For each $i\in[d]$, define the shorthand

$$\begin{aligned}\Delta_i&=\nabla_{\theta_i}\ell(W;f_\theta(X),p(X))+\nabla_{\theta_i}f_\theta(X)^\top\gamma(X)^\top(Y-p(X))-\nabla_{\theta_i}\ell(W;f_\theta(X),p_0(X))\quad\text{and}\\ Z_i&=\mathbb{E}[\Delta_i\mid X]\\ &=\nabla_{\theta_i}f_\theta(X)^\top(\gamma(X)-\gamma_{f_\theta,p}(X))^\top(p_0(X)-p(X))\\ &=\nabla_{\theta_i}f_\theta(X)^\top(\gamma(X)-q_{f_\theta,p}(X))^\top(p_0(X)-p(X))\\ &+\frac{1}{2}\mathbb{E}[\nabla_{\pi\pi\phi}\ell(W;f_\theta(X),\bar{p}(X))[\nabla_{\theta_i}f_\theta(X),p_0(X)-p(X),p(X)-p_0(X)]]\end{aligned}$$

for some convex combination $\bar{p}(X)$ of $p(X)$ and $p_0(X)$.

We begin by bounding the target expectation using Cauchy-Schwarz

$$
\begin{aligned}
&\mathbb{E}[\|\mathbf{n}(W;\theta,p,\gamma)\|_2^2] \\
&= \sum_{i\in[d]} \mathrm{Var}[\nabla_{\theta_i}\ell(W;f_\theta(X),p(X)) + \nabla_{\theta_i} f_\theta(X)^\top \gamma(X)^\top (Y-p(X))] \\
&= \sum_{i\in[d]} \mathrm{Var}[\nabla_{\theta_i}\ell(W;f_\theta(X),p_0(X)) + \Delta_i] \\
&= \sigma_0(\theta,p_0)^2 + \sum_{i\in[d]} \mathrm{Var}[\Delta_i] + 2\mathrm{Cov}(\nabla_{\theta_i}\ell(W;f_\theta(X),p_0(X)),\Delta_i) \\
&\leq \sigma_0(\theta,p_0)^2 + \sum_{i\in[d]} \mathrm{Var}[\Delta_i] + 2\sqrt{\mathrm{Var}[\nabla_{\theta_i}\ell(W;f_\theta(X),p_0(X))]\mathrm{Var}[\Delta_i]} \\
&\leq \sigma_0(\theta,p_0)^2 + (\sum_{i\in[d]} \mathrm{Var}[\Delta_i]) + 2\sqrt{\sum_{i\in[d]} \mathrm{Var}[\nabla_{\theta_i}\ell(W;f_\theta(X),p_0(X))] \sum_{i\in[d]} \mathrm{Var}[\Delta_i]} \\
&= (\sigma_0(\theta,p_0) + \sqrt{\sum_{i\in[d]} \mathrm{Var}[\Delta_i]})^2.
\end{aligned}
$$

We next employ the law of total variance to rewrite the variance terms:

$$
\begin{aligned}
\sum_{i\in[d]} \mathrm{Var}[\Delta_i] &= \sum_{i\in[d]} \mathrm{Var}[Z_i + \nabla_{\theta_i} f_\theta(X)^\top \gamma(X)^\top (Y-p_0(X))] \\
&= \mathbb{E}[\left\|\nabla_\theta f_\theta(X)^\top \gamma(X)^\top (Y-p_0(X))\right\|_2^2] + \sum_{i\in[d]} \mathrm{Var}[Z_i].
\end{aligned}
$$

Finally, we control $\mathrm{Var}[Z_i]$ using Cauchy-Schwarz

$$
\sqrt{\sum_{i\in[d]} \mathrm{Var}[Z_i]} \leq \left\|\nabla_\theta f_\theta^\top (\gamma_{f_\theta,p} - \gamma)^\top (p-p_0)\right\|_{2,2}.
$$

□

The two claims of Thm. 4 now follow from Theorems 2 and 3 of Ajalloeian & Stich (2020) respectively, with the parameters σ^2 and ζ instantiated with quantities $\sigma^2(\gamma)$ and $\zeta(\gamma)$ of Lemmas 5 and 6.

G Experiment Details and Additional Results

G.1 Synthetic experiment

The synthetic data is generated from a distribution $\mathbb{P}$ that is a mixture of two Gaussians with equal weights. with means at $\pm(r,r,\ldots,r) \in \mathbb{R}^{10}$, where $r \in \mathbb{R}$ controls the separation between the two Gaussian distributions. For each x, the label y is index of the Gaussian from which x was drawn. Given this generative process, the Bayes classifier has the form $p_0(x) = p(y=1\,|\,x) = \frac{1}{1+e^{\theta^T x}}$ for some vector $\theta \in \mathbb{R}^{10}$.

We set $r = 0.25$, and train with gradient descent for 40 epochs, with learning rate 0.3. The λ hyperparameter of the γ-correction was chosen by cross-validation. We repeat the experiments 5 times to measure the mean and standard deviation.

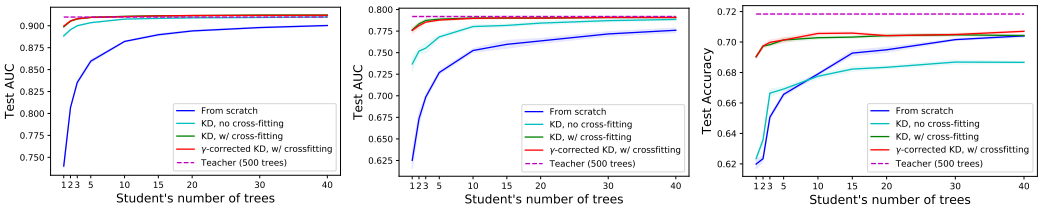
G.2 Tabular data

We use cross-fitting with 10 folds. The λ hyperparameter of the γ-correction was chosen by cross-validation. We repeat the experiments 5 times to measure the mean and standard deviation.

For the overfitting experiment, we use a random forest with 500 trees as the teacher, and a random forest with 1-40 trees as the student. The student is trained to match the class log probability $\log(\hat{p}(x))$ of the teacher. When $\hat{p}(x) = 0$, we clip its value to 10^{-3}.

We also evaluate the impact of γ correction compared to without correction (i.e., training on just the teacher's probabilities), especially when the teacher has high bias. We vary the teacher's bias by limited the tree depth in the teacher (from 1 to 20). Lower depth corresponds to more bias. The teacher has 100 trees and the student has 10 trees. For all of the dataset, γ correction successfully reduces the teacher's bias and thus improves the student's performance. The effect is most pronounced when the teacher has large bias (low tree depth).

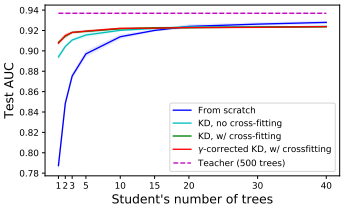
We show the full results for all 5 of the datasets in Figs. 4 and 5.



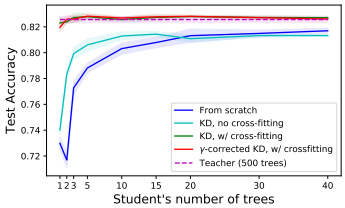
(a) Adult dataset

(b) FICO dataset

(c) Higgs dataset

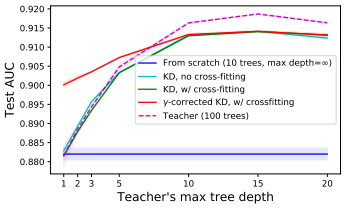


(d) MAGIC dataset

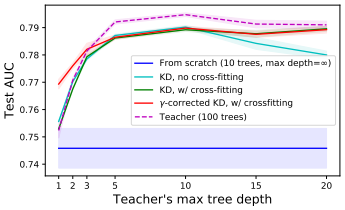


(e) StumbleUpon dataset

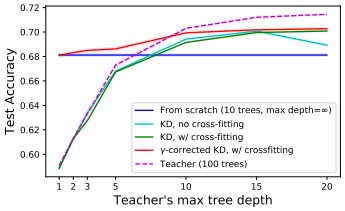
Figure 4: Comparing knowledge distillation, with and without cross-fitting.



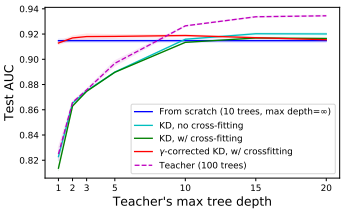
(a) Adult dataset



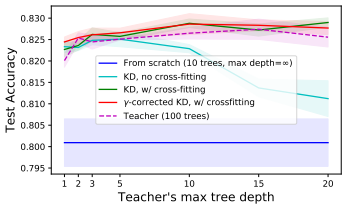
(b) FICO dataset



(c) Higgs dataset



(d) MAGIC dataset



(e) StumbleUpon dataset

Figure 5: Comparing knowledge distillation, with and without γ correction.

G.3 Experiment on image data (CIFAR-10)

We use SGD with initial learning rate 0.1 and momentum 0.9, batch size 128, to train for 200 epochs. We use the standard learning rate decay schedule, where the learning rate is divided by 5 at epoch 60, 120, and 160.

Ablation on the effect of hyperparameter α The hyperparameter α controls the tradeoff between bias and variance. When α is very small, the objective is close to the vanilla KD objective. When α is large, the objective is closer to the Neyman-orthogonal loss. In Figure 6, we show the effect of varying α, with ResNet-8 as the student and ResNet-20 as the teacher, on the CIFAR-10 dataset. Large values of α lead to high variance and thus lower test accuracy. Intermediate values of α improves on the vanilla KD objective, which corresponds to $\alpha = 0$. The test accuracy drops sharply beyond some threshold of α as the variance becomes too high (due to the terms $q_{\hat{p}}(x) = \mathrm{diag}\left(\frac{1}{\hat{p}_1(x)}, \ldots, \frac{1}{\hat{p}_K(x)}\right)$), causing training to become unstable.

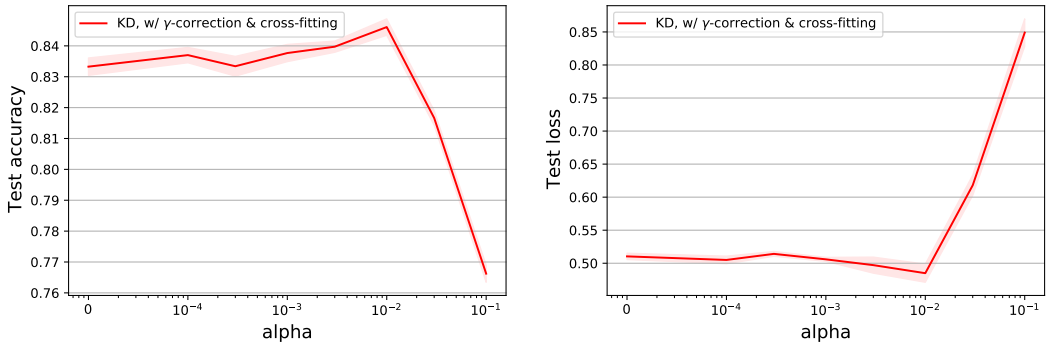


Figure 6: On CIFAR-10, large values of hyperparameter α (corresponding to orthogonal loss) is unstable, while intermediate values improve on vanilla KD.