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 $\ell_2$ 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 $P$. Each feature vector $x$ belongs to a set $\mathcal{X}$, each label vector $y \in \{e_1,..,e_j\} \subset \{0,1\}^k$ 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 $P$.

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 $[k] \equiv \{1,2,...,k\}$. The most popular distillation loss functions\footnote{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.} $\ell(z; f(x), \hat{p}(x))$ include the squared error logit loss (Ba & Caruana, 2014)

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

(1)
and the annealed cross-entropy loss (Hinton et al., 2015)

\[ \ell_{\beta}(z; f(x), \hat{p}(x)) = -\sum_{j \in [k]} \frac{p_j(x)^\beta}{\sum_{i \in [k]} p_i(x)^\beta} \log \left( \frac{\exp(\beta f_j(x))}{\sum_{i \in [k]} \exp(\beta f_i(x))} \right) \]

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 (Bucilu 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.

Before presenting our main theorem we introduce some technical notation.

Technical definitions

For a vector valued function \( f \) that takes as input a random variable \( X \), we use the shorthand notation \( \| f \|_{L^q} \triangleq \| f(X) \|_{L^q} = E[\| f(X) \|_p^q]^{1/q} \). Let \( \nabla_\phi \) and \( \nabla_\pi \) 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., \( \nabla_{\phi, \pi} \ell(z; \phi, \pi)_{i,j} = \frac{\partial^2}{\partial \phi_i \partial \pi_j} \ell(z; \phi, \pi) \). Finally, let

\[ q_{f,p}(x) = E[\nabla_\phi \ell(Z; f(X), p(X)) | X = x] \quad \text{and} \quad \gamma_{f,p}(x) = E_{U \sim \text{Unif}(0,1)} [q_{f,p}(x) - q_{f_0,p}(x)]. \]

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 \( F \) we define the localized Rademacher complexity as:

\[ R(\delta; F) = E_{X_1, \ldots, X_n, \epsilon_1, \ldots, \epsilon_n} \sup_{f \in F} \| f \|_{L^1} = \frac{1}{n} \sum_{i=1}^n \epsilon_i f(X_i), \]

where \( \epsilon_i \) are i.i.d. random variables taking values equiprobably in \( \{-1,1\} \). The critical radius of a class \( F \), taking values in \( [-H, H] \), is the smallest solution \( \delta_n \) to the inequality \( R(\delta; F) \leq \frac{H^2}{d_\phi} \).

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

\[ \mathcal{G} \triangleq \{ z \to r(\ell(z; f(x), p(x)) - \ell(z; f_0(x), p(x))) : f \in F, p \in P, r \in [0,1] \} \]
Assume that, for all $z \in \mathbb{Z}$ and $\pi$, the loss $\ell(z; \phi, \pi)$ is $\sigma$-strongly convex with respect to $\phi$, uniformly bounded in $[-H, H]$ for all $z, \pi$, and let: $\mu(z) = \sup_{\theta}[\nabla_{\phi} \ell(z; \phi, \hat{p}(x))]$. Moreover, suppose that the function class $f$ satisfies the $\ell_2/\ell_4$ ratio condition: $\sup_{f \in F} \|f - f_0\|_{\ell_4, 2} \leq C$. Then $\hat{f}$ satisfies

$$\|\hat{f} - f_0\|_{\ell_2, 2} \leq \frac{1}{\sqrt{T}} \left( O\left(\delta_{n, C}^2 H^2 \|\mu\|_4^4 + \phi_{f \to \theta}(\hat{p} - p_0)\|_2^2, 2\right) \right) \quad \text{with probability} \quad 1 - \zeta.$$ 

Remark

The proof is in App. D. We note that the $\ell_2/\ell_4$ ratio requirement can be removed, albeit the final bound will depend on a uniform bound $\|\mu\|_{\infty} = \sup_{x}[\mu(x)]$, instead of $\|\mu\|_4$. Moreover, we note that the strong convexity requirement for $f$ 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: $\|\hat{f} - f_0\|_{\ell_2, 2} = O(\delta_{n, C} + \phi_{f \to \theta}(\hat{p} - p_0)\|_2^2, 2)$.

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 $\mu(x) \hat{p}(x)$, which now contains extra variation due to the variance of the labels $y$. Thus in practice it can be that $\|\mu\|_4 \ll \|\mu\|_4$.

**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 $P$ is a large model space, with a large critical radius, then $G$ 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., $\|p - p_0\|_{\ell_2, 2}$ 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 $c > 0$, belongs to the set

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

Suppose moreover that $X \in \mathbb{R}^d$ has Lebesgue density bounded away from 0 and $\epsilon$ and that $\epsilon < 1/\mathbb{E}[(1-p_0_j(X))1/(1-p_0_j(X))^2]$ for each $j$. Consider the teacher estimates $p_j(x) = \max(c, p_j(x))$ for $p_0$ the Nadaraya-Watson kernel smoothing estimator (Nadaraya, 1964; Watson, 1964)

$$\hat{p}(x) \triangleq \frac{\sum_{i=1}^n y_i K((x - x_i)/h)}{\sum_{i=1}^n K((x - x_i)/h)} \quad \text{if} \quad x = x_i$$

$$\text{otherwise}$$

with kernel $K(x) = \|x\|^2 a_n \|x\|^2 \leq 1$, $a \in (0, d/2)$, and $n = n^{-1/(4+d)}$. Notably, $\hat{p}$ exactly interpolates the observed labels (i.e., $\hat{p}(x_i) = y_i$) but, by Belkin et al. (2019, Thm. 1), also satisfies $\mathbb{E}[\|p_0 - \hat{p}\|_{\ell_2, 2}^2] = O(n^{-4/(4+d)})$. Consider now a student that learns a constant prediction rule via distilled ERM (3) with the squared error logit loss (1) and:

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

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

$$f_0(x) - E[f_0] = \mathbb{E}[\log(p_0_j(X)) - \log(max(Y_j, \epsilon))] \geq E[\frac{P_{0_j}(X) - \max(Y_j, \epsilon)}{P_{0_j}(X)} (P_{0_j}(X) - \frac{\log(Y_j, \epsilon)}{P_{0_j}(X)})^2]$$

$$= E[\frac{P_{0_j}(X)(1-P_{0_j}(X))^2}{P_{0_j}(X)^2} (1-P_{0_j}(X))(1-P_{0_j}(X) - \frac{\log(Y_j, \epsilon)}{P_{0_j}(X)})^2] - E[\frac{1-P_{0_j}(X)}{P_{0_j}(X)}] \geq E[\frac{P_{0_j}(X)(1-P_{0_j}(X))^2}{P_{0_j}(X)^2}]$$

by Taylor’s theorem with Lagrange remainder. This non-vanishing student error reflects the non-vanishing critical radius $\delta_1$ of the composite student-teacher function class $G$ defined in Thm. 1; since the student function class $F$ has low complexity, the complexity of $G$ is driven by the highly flexible interpolating teacher. We will demonstrate how to avoid this dependence on teacher complexity in Sec. 4.2.
We will show that orthogonal correction (5) can significantly improve student bias due to teacher
bias. We can view the plug-in distillation loss
\[ f_0 - \hat{f} - \log(p_0)^\top - \log(\hat{p}) \geq \text{diag}(\frac{1}{p_0}) (p_0 - \hat{p}) = \gamma_{f_0,p_0}^\top (p_0 - \hat{p}) = \gamma_{f_0,p_0}^\top \Lambda_{p_0}(\gamma_{f_0,p_0}^\top) \]
by the concavity of the logarithm. Since \( P(\hat{y}_j - p_{0,j}(x) \geq \theta_{p_{0,j}(x)}) \leq \frac{\zeta}{\theta} \) for
\( \theta \geq \sqrt{\frac{2(1-p_0(x))\log(\frac{2}{\zeta})}{np_0(x)}} + \frac{1}{\sqrt{2p_0(x)}} \log(\frac{2}{\zeta}) \) by Bernstein’s inequality (Bernstein, 1946), we have
\( P(\|\gamma_{f_0,p_0}^\top \Lambda_{p_0}(\gamma_{f_0,p_0}^\top)\|_2^2 \geq 1 - \zeta \) matching the teacher MSE dependence of the Thm. 1
upper bound whenever \( \lambda \geq \sqrt{\frac{2\log(\frac{2}{\zeta})}{n}} + \frac{1}{\sqrt{2p_0(x)}} \log(\frac{2}{\zeta}) \). Moreover, since \( \limsup_{n \to \infty} \sqrt{\frac{\eta(\gamma_{p_0})}{\eta(p_0(1-p_0))\sqrt{\log(n)}}} = 1 \)
with probability 1 by the law of the iterated logarithm, \( \|f_0 - \hat{f}\|_2^2 = \Omega((\min(1,\lambda^2)) \) with probability
1 whenever \( \lambda \geq \sqrt{\frac{2\log(\log(n))}{n}} \).

4 Enhancing Knowledge Distillation

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

4.1 Combating Teacher Underfitting with \( \gamma \)-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
\[ \ell(z; f(x), \hat{p}(x)) + \langle \nabla_\hat{p} \ell(z; f(x), \hat{p}(x)) \rangle \].
However, its computation also requires knowledge of \( p_0 \). Nevertheless, since \( p_0(x) = \mathbb{E}[Y | X = x] \), we can always construct an unbiased estimate of the ideal first order term by replacing \( p_0(x) \) with \( y \):
\[ \ell_{\text{ortho}}(z; f(x), \hat{p}(x)) = \ell(z; f(x), \hat{p}(x)) + \langle y - \hat{p}(x), \mathbb{E}[\nabla_\hat{p} \ell(z; f(x), \hat{p}(x)) | x] \rangle. \] (5)
For standard distillation base losses like (1) and (2), the orthogonal loss (5) has an especially simple form, as \( \nabla_\hat{p} \ell(z; f(x), \hat{p}(x)) \) 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
\[ \ell(z; f(x), p(x)) \triangleq \Psi(f(x)) - \Psi(g(p(x))) - \langle \nabla_\hat{p} \Psi(g(p(x))), f(x) - g(p(x)) \rangle \]
has \( \ell_{\text{ortho}}(z; f(x), p(x)) = \ell(z; f(x), p(x)) + \langle y - p(x), \nabla_\hat{p} \Psi(g(p(x))) f(x) \rangle \) 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 \( \nabla_\hat{p} \Psi(g(p(x))) = \nabla_\hat{p} \Psi(g(p(x))) = \text{diag}(\frac{1}{p(x)}) \). 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 \( \gamma \)-corrected losses, parameterized by a matrix valued function \( \gamma: \mathcal{X} \to \mathbb{R}^k \times \mathbb{R}^k \):
\[ \ell_{\gamma}(z; f(x), p(x)) \triangleq \ell(z; f(x), p(x)) + \langle y - p(x), \gamma(x) f(x) \rangle \]
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
\[ L_D(f, p, \gamma) \triangleq \mathbb{E}[\ell_{\gamma}(Z; f(X), p(X))] \] and \( L_n(f, p, \gamma) \triangleq \mathbb{E}[\ell_{\gamma}(Z; f(X), p(X))] \).
Observe that at \( p_0 \) the correction term is mean-zero and hence \( L_D(f, p_0, \gamma) \) is independent of \( \gamma \)
\[ L_D(f, p_0) \triangleq \mathbb{E}[\ell(Z; f(X), p_0(X))] = L_D(f, p_0, \gamma) \] for all \( \gamma \).

The \( \gamma \)-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 \( \gamma \) is set appropriately, then one can show that the \( \gamma \)-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
We now provide a more sophisticated version of sample splitting to make use of all datapoints in a more robust manner. Moreover, suppose that the function class is a convex set. Let $\tau$ be a convex set. Let $f$ be a convex function. Now, suppose that $f(\hat{\beta},\gamma)$ is strongly convex with respect to $\beta$ and $F$ is a convex set. Then:

$$
\frac{2}{\tau} \| f - f_0 \|_2^2 \leq \epsilon(\hat{\beta},\gamma) + \frac{1}{\tau} \| (\gamma f_0,\hat{\beta} - \gamma)^T (\hat{p} - p_0) \|_2^2
$$

If, further, $\sup_{z,\phi,\pi,\epsilon} \| \nabla_\phi,\pi \epsilon(z,\phi,\pi) \|_{\text{op}} \leq M$, then:

$$
\| (\gamma f_0,\hat{\beta} - \gamma)^T (\hat{p} - p_0) \|_2^2 \leq 2(\| (q f_0,\hat{\beta} - \gamma)^T (\hat{p} - p_0) \|_2^2 + M^2 k \| \hat{p} - p_0 \|_4^2)
$$

**Connection to Neyman orthogonality** Note that if we set $\gamma = q f_0,\hat{\beta}$, 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 $\gamma = q f_0,\hat{\beta}$ renders the loss function Neyman orthogonal (Foster & Syrgkanis, 2019). Moreover, $q f_0,\hat{\beta}$ is an observable quantity for any Bregman divergence loss (Definition 1) as $q f_0,\hat{\beta}$ is independent of $f_0$. However, we note that this setting of the $\gamma$ can lead to larger variance, i.e., the achievable excess risk can be much larger than the excess risk without the $\gamma$-correction. For instance, in the case of the squared error logit loss (1) $q f_0,\hat{\beta}(x) = \frac{1}{1 + e^{z x}}$, 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 $\gamma$ in practice to balance bias and variance.

**Example instantiation of student’s estimation algorithm** If we use *plug-in empirical risk minimization*, i.e., $\hat{f} = \arg\min_{f \in F} L_n(f,\hat{\beta},\gamma)$, to estimate $f_0$ with $\hat{\beta}$ 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. $1 - \delta$:

$$
\epsilon(\hat{f},\hat{\beta},\gamma) \leq O\left(\sqrt{\frac{\sup_{f \in F} \text{Var}(\ell(z,f(X),\hat{\beta}(X))) \log(\tau(n)/\delta)}{n} + \frac{H \log(\tau(n)/\delta)}{n}}\right)
$$

where $\tau(n) = N_{\infty}(1/n,F,2n)$ and $N_{\infty}(\epsilon,F,m)$ is the $\epsilon$-covering number of function class $F$ in the worst-case over all realizations of $m$ samples and at approximation level $\epsilon$. 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. 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 $\hat{\beta}(t)$ and $\hat{\gamma}(t)$ using all the out-of-fold samples.
3. Estimate $\hat{f}$ by minimizing the empirical loss:

$$
\hat{f} = \arg\min_{f \in F} \frac{1}{B} \sum_{i=1}^{B} \sum_{\ell \in [B]} \ell_{\gamma}(Z_i,f(X_i),\hat{\beta}(t)(X_i))
$$

In other words, the nuisance estimates $\hat{\gamma}(t),\hat{\beta}(t)$ 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 $f_0 \in F$ and $F$ is a convex set. Let $\delta_{n,\epsilon} = \delta_n + c_0 \sqrt{\log(c_1/\epsilon)}$ for universal constants $c_0,c_1$ and $\delta_n$ an upper bound on the critical radii of the classes:

$$
G(\hat{\beta}(t),\hat{\gamma}(t)) = \{ z \rightarrow \ell_{\gamma}(z,f(X),\hat{\beta}(t)(X)) - \ell_{\gamma}(z,f_0(X),\hat{\beta}(t)(X)) : f \in F, \forall \epsilon \in [0,1] \}
$$

for each $t \in [B]$. Assume that, for all $z \in Z$, $\phi$, and $\pi$, the loss $\ell(z,\phi,\pi)$ is uniformly bounded in $[-H,H]$ and $\sigma$-strongly convex with respect to $\phi$, and let $\mu(z) = \sup_{f \in F, \epsilon \in [0,1]} [\nabla_\phi,\pi \ell(z;f(X),\hat{\beta}(t)(X))].$

Moreover, suppose that the function class $f$ satisfies the $\ell_2/\ell_4$ ratio condition: $\sup_{f \in F, \epsilon \in [0,1]} [\| f - f_0 \|_2 / \| f - f_0 \|_4] \leq C.$
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 – \( \zeta \):

\[
\frac{1}{\sigma} \| \hat{f} - f_0 \|_{2,2}^2 \leq \frac{1}{\sigma} O \left( \frac{\delta^2_{\gamma, f_0, B}}{\sqrt{n}} \right) + \frac{1}{\sigma} \sum_{t=1}^{B} \sqrt{E \left[ \left( \gamma \left( \hat{f}(X) - \hat{p}(X) \right) - \gamma(\hat{p}(X) - p_0(X)) \right) \right]}
\]

The proof is found in App. E. Observe that unlike Thm. 1, the function classes \( G(\hat{p}(t), \hat{\gamma}(t)) \) in the latter theorem do not vary the teacher’s model over \( P \) but rather only vary \( f \in F \) and evaluate \( p \) at the out-of-fold estimate \( \hat{p}(t) \). 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 \( G(\hat{p}(t), \hat{\gamma}(t)) \), can be significantly smaller than \( G \). For instance, if \( P \) and \( F \) are VC-subgraphs classes with VC dimensions \( d_P \gg d_F \), correspondingly, then the critical radius of \( G(\hat{p}(t), \hat{\gamma}(t)) \), will be of the order \( \sqrt{d_F \log(n)/\mu} \), that of \( G \) of the order of \( \sqrt{d_P \log(n)/n} \). Moreover, we can see in the bound of Thm. 3 the interplay between bias and variance introduced by \( \gamma \). In particular, the part of the bound that depends on \( \hat{\gamma}(t) \) can be further simplified as:

\[
\sqrt{\mathbb{E} \left[ \frac{\delta^4_{\gamma, \gamma}}{n} \right] \left( \| \hat{f}(X) - \hat{p}(X) \|_2 + \| \gamma(X) - \gamma(\hat{p}(X) - p_0(X)) \|_2^2 \right)}
\]

where the first part corresponds to the bias increase due to \( \gamma \), while the second part the bias decrease.

### 4.3 Biased stochastic gradient descent analysis

When the set of candidate prediction rules \( f_0 \) is parameterized by a vector \( \theta \in \mathbb{R}^d \), we may alternatively fit \( \theta \) via stochastic gradient descent (SGD) (Robbins & Monro, 1951; Bottou & Bousquet, 2008) on the \( \gamma \)-corrected objective \( L_D(f_0, \hat{\gamma}, \hat{\gamma}) \). With a minibatch size of 1 and a starting point \( \theta_0 \), the parameter updates take the form

\[
\theta_{t+1} = \theta_t - \eta \nabla_{\theta} L_D(f_0, \hat{\gamma}, \hat{\gamma})(W_t; f_0(X_t), p_0(X_t)) \quad \text{for} \quad t 

Ideally, these updates would converge to a minimizer of the ideal risk \( L_D(f_0, \hat{\gamma}, \hat{\gamma}) \). Our next result shows that, if the teacher \( \hat{p} \) is independent of \( \langle W_t \rangle_{t \in [n]} \), 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, \( \sigma(\hat{\gamma})^2 \) represents the baseline stochastic gradient variance that would be incurred if SGD were run directly on the ideal risk \( L_D(f_0, \hat{\gamma}, \hat{\gamma}) \) rather than our surrogate risk. Our proof in App. F builds upon the biased SGD bounds of Ajalleian & Stich (2020).

**Theorem 4 (Biased SGD analysis).** Suppose that the loss \( L_D(f_0, \hat{\gamma}, \hat{\gamma}) \) is \( \lambda \)-strongly smooth in \( \theta \). Define the bias and root-variance parameters

\[
\zeta(\hat{\gamma}) \triangleq \sup_{\theta \in \mathbb{R}^d} \| \nabla_{\theta} f_0^\top (\gamma(X) - \hat{\gamma}(\hat{p}(X) - p_0(X))) \|_2 \]

\[
\sigma^2(\hat{\gamma}) \triangleq \sup_{\theta \in \mathbb{R}^d} \| \nabla_{\theta} f_0^\top (\gamma(X) - \hat{\gamma}(\hat{p}(X) - p_0(X))) \|_2^2 \]

for \( \eta(\theta) = \eta \sum_{i \in [d]} \text{Var}[\nabla_{\theta_i} L(\hat{p}; f_0(X), p_0(X))] \) the unbiased SGD variance. If \( F_0 = L_D(\theta_0, p_0) - \min_{\theta \in \mathbb{R}^d} L_D(\theta, p_0) \), then the iterates \( \{\theta_t\}_{t=1}^n \) of the \( \gamma \)-corrected SGD algorithm satisfy

\[
\min_{\theta \in \mathbb{R}^d} E \left[ \| \nabla_{\theta} L_D(\theta, p_0) \|_2^2 \right] = O \left( \frac{\sigma^2(\hat{\gamma})}{\sqrt{n}} + \zeta^2(\hat{\gamma}) \right)
\]

If, in addition, \( L_D(f_0, \hat{\gamma}, \hat{\gamma}) \) is \( \mu \)-strongly convex in \( \theta \), then the iterates satisfy

\[
E \left[ L_D(\theta_n, p_0) - \min_{\theta \in \mathbb{R}^d} L_D(\theta, p_0) \right] = O \left( \frac{\sigma^2(\hat{\gamma})}{\sqrt{n}} + \zeta^2(\hat{\gamma}) \right) + O \left( F_0 e^{-\frac{\mu}{\lambda} n} \right).
\]

Similar to Thm. 3, the bound in Thm. 4, portrays the interplay of bias and variance as \( \hat{\gamma} \) ranges from 0 to \( \eta(\theta) \) (which we remind that for any Bregman loss is independent of \( f_0 \)). In particular, the part of the bound for strongly convex losses that depends on \( \hat{\gamma} \) can be further simplified to:

\[
E \left[ \left( \frac{\lambda}{\mu} \| \hat{\gamma}(X) - \hat{p}(X) \|_2 \right)^2 + \| \gamma(X) - \hat{\gamma}(X) \|_2^2 \right] + \| \nabla_{\theta} f_0(X) \|_2^2
\]

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 \( \theta \), which expresses that some regions of \( X \) are more important in identifying the parameters and therefore, when we minimize the average bias-variance trade-off, we should be re-weighting the samples based on how informative they are about the parameters \( \theta \).
5 Experiments

We empirically validate that the two components of our solution, cross-fitting and $\gamma$-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 $\gamma$-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 $\gamma$-correction error (9) (ideally with a closed-form solution to avoid excessive computational overhead). To eliminate dependence on the unobserved $p_0$, we observe that the bias term $\| (\gamma_{fs,p}(x) - \hat{\gamma}(x)) \| (\hat{p}(x) - p_0(x)) \|_2^2 = O(\| q_{fs,p}(x) - \hat{\gamma}(x) \|_D^2)$ up to additive terms independent of $\hat{\gamma}$. We introduce a tunable hyperparameter $\alpha \geq 0$ to trade off between this bias bound and the variance term in (9) and select $\hat{\gamma}(x) = \text{diag}(v(x))$ to minimize

$$\| \hat{\gamma}(x) \|_D^2 \mathbb{E}[\| y - \hat{p}(x) \|_2^2 | x] + \alpha \| q_{fs,p}(x) - v(x) \|_D^2 = \mathbb{E}[\| y - \hat{p}(x) \|_2^2 | x] + \alpha \| \frac{1}{p_0(x)} - v(x) \|_2^2.$$  

Since the conditional expectation involves the unknown quantity $p_0$, we approximate $\mathbb{E}[\| y - \hat{p}(x) \|_2^2 | x]$ by its sample $\| y - \hat{p}(x) \|_2^2$. 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 $\gamma$-corrected objective can be shown to be equivalent to a square loss with labels $\log(p(x)) + \gamma(x) (y - p(x))$.

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, compared to a student model trained from scratch.

We compare 4 training objectives for the student model: (i) trained from scratch (ii) trained with vanilla KD (iii) trained with $\gamma$-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, $\gamma$-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: 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 $\gamma$-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 $\gamma$-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 $\mathbb{E}[\| y - \hat{p}(x) \|_2^2 | x]$ by $\| \hat{p}(x)(1 - \hat{p}(x)) \|_2^2$.  

---

Under review as a conference paper at ICLR 2021
shows significant improvement over using the same teacher on the training points. The $\gamma$ 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).

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 $\gamma$-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 $\gamma$-correction consistently provides a small performance boost.

App. G.3 shows the effect of varying the hyperparameter $\alpha$. Large $\alpha$ (closer to orthogonal loss) leads to high variance and lower test accuracy. Intermediate $\alpha$ 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 $\gamma$-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


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>
<thead>
<tr>
<th>Notation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\ell(Z; f(X), p_0(X))$</td>
<td>Loss function on a random data point</td>
</tr>
<tr>
<td>Population risk $L_D(f, p)$</td>
<td>$\mathbb{E}[\ell(Z; f(X), p(X))]$</td>
</tr>
<tr>
<td>Empirical risk $L_n(f, p)$</td>
<td>$\mathbb{E}_n[\ell(Z; f(X), p(X))]$</td>
</tr>
<tr>
<td>Population optimal student model $f_0$</td>
<td>$\argmin_{f \in \mathcal{F}} L_D(f, p_0)$</td>
</tr>
<tr>
<td>Empirical optimal student model $\hat{f}$</td>
<td>$\argmin_{f \in \mathcal{F}} L_n(f, \hat{p})$</td>
</tr>
<tr>
<td>$|f|_{p,q}$</td>
<td>$||f(X)|<em>p|</em>{L^q} = \mathbb{E}[|f(X)|_p^{1/q}]$</td>
</tr>
<tr>
<td>$\nabla \phi$</td>
<td>Partial derivative of $\ell(z, \phi, \pi)$ wrt to the second input</td>
</tr>
<tr>
<td>$\nabla \pi$</td>
<td>Partial derivative of $\ell(z, \phi, \pi)$ wrt to the third input</td>
</tr>
<tr>
<td>$\nabla_{\phi \pi}$</td>
<td>$[\nabla_{\phi \pi} \ell(z, \phi, \pi)]_{i,j} = \frac{\partial^2}{\partial \phi_i \partial \pi_j} \ell(z, \phi, \pi)$</td>
</tr>
<tr>
<td>$q_{f,p}(x)$</td>
<td>$\mathbb{E} [\nabla_{\phi \pi} \ell(Z; f(X), p(X)) \mid X = x]$</td>
</tr>
<tr>
<td>$\gamma_{f,p}(x)$</td>
<td>$\mathbb{E} [\ell_{\gamma}(Z; f(X), p(X)) \mid X = x]$</td>
</tr>
<tr>
<td>$\mathcal{R}(\delta; \mathcal{F})$</td>
<td>Localized Rademacher complexity of function class $\mathcal{F}$</td>
</tr>
<tr>
<td>$\delta_n$</td>
<td>Critical radius</td>
</tr>
<tr>
<td>$\gamma$-corrected loss $\ell_{\gamma}(z; f(x), p(x))$</td>
<td>$\ell(z; f(x), p(x)) + (y - p(x))^\top \gamma(x) f(x)$</td>
</tr>
<tr>
<td>Population $\gamma$-risk $L_D(f, p, \gamma)$</td>
<td>$\mathbb{E}[\ell_{\gamma}(Z; f(X), p(X))]$</td>
</tr>
<tr>
<td>Empirical $\gamma$-risk $L_n(f, p, \gamma)$</td>
<td>$\mathbb{E}<em>n[\ell</em>{\gamma}(Z; f(X), p(X))]$</td>
</tr>
</tbody>
</table>

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 $\sigma$-strong convexity of $L_D$, 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
\[
\frac{\sigma}{2} \| \hat{f} - f_0 \|_{2,2}^2 \leq \underbrace{L_D(f, \hat{p}, \gamma) - L_D(f_0, \hat{p}, \gamma) - D_f L_D(f_0, \hat{p}, \gamma)}_{\text{excess risk of } f} - D_f L_D(f_0, \hat{p}, \gamma)] [\hat{f} - f_0] 
\]
\[\leq \epsilon(\hat{f}, \hat{p}, \gamma) - D_f L_D(f_0, \hat{p}, \gamma)] [\hat{f} - f_0] + D_f (L_D(f_0, \hat{p}, \gamma) - L_D(f_0, \hat{p}, \gamma)] [\hat{f} - f_0].
\]
By Taylor’s theorem with integral remainder,
\[
\mathbb{E}[(\nabla_{\phi} \ell(W; f_0(x), p_0(x)) - \nabla_{\phi} \ell(W; f_0(x), \hat{p}(x)), \hat{f}(x) - f_0(x) | X = x] 
\]
\[
= \langle p_0(x) - \hat{p}(x), \gamma_{f_0, \hat{p}}(x) \rangle \hat{f}(x) - f_0(x) \rangle 
\]
whenver \(\nabla_{\phi} \ell\) is well-defined. We can now invoke the expansion (10) and Cauchy-Schwarz to obtain the bound
\[
D_f (L_D(f_0, \hat{p}, \gamma) - L_D(f_0, \hat{p}, \gamma)] [\hat{f} - f_0] 
\]
\[= \mathbb{E}[(\nabla_{\phi} \ell(W; f_0(X), p_0(X)) - \nabla_{\phi} \ell(W; f_0(X), \hat{p}(X)), \hat{f}(X) - f_0(X) ] 
\]
\[= \mathbb{E}[(p_0(X) - \hat{p}(X)) \gamma_{f_0, \hat{p}}(X) \hat{f}(X) - f_0(X) ] 
\]
\[\leq \mathbb{E}[(p_0(X) - \hat{p}(X)) \gamma_{f_0, \hat{p}}(X) \hat{f}(X) - f_0(X)] 
\]
\[\leq \mathbb{E}[(p_0(X) - \hat{p}(X)) \gamma_{f_0, \hat{p}}(X) \hat{f}(X) - f_0(X)] 
\]
\[\leq \|p_0 - \hat{p}\| \gamma_{f_0, \hat{p}}(X) \|[\hat{f}(X) - f_0(X)] \| \|\hat{f} - f_0\|_{2,2} 
\]
Thus combining all the above inequalities:
\[
\frac{\sigma}{2} \| \hat{f} - f_0 \|_{2,2}^2 \leq \epsilon(\hat{f}, \hat{p}, \gamma) + \|p_0 - \hat{p}\| \gamma_{f_0, \hat{p}}(X) \|\hat{f} - f_0\|_{2,2} 
\]
By an AM-GM inequality, for all \(a, b \geq 0\): \(a \cdot b \leq \frac{1}{2}(a^2 + b^2)\). Applying this to the product of norms on the RHS and re-arranging yields
\[
\frac{\sigma}{4} \| \hat{f} - f_0 \|_{2,2}^2 \leq \epsilon(\hat{f}, \hat{p}, \gamma) + \|p_0 - \hat{p}\| \gamma_{f_0, \hat{p}}(X) \|\hat{f} - f_0\|_{2,2} 
\]
To get the final inequality, observe that:
\[
\|p_0 - \hat{p}\| \gamma_{f_0, \hat{p}}(X) \|\hat{f} - f_0\|_{2,2} \leq \mathbb{E}[\|p_0(X) - \hat{p}(X)\| \gamma_{f_0, \hat{p}}(X) \|\hat{f} - f_0\|_{2,2}] 
\]
Moreover, by the boundedness of the third derivative, we have:
\[
\|p_0 - \hat{p}\| \gamma_{f_0, \hat{p}}(X) \|\hat{f} - f_0\|_{2,2} \leq \mathbb{E}[\|p_0(X) - \hat{p}(X)\| \gamma_{f_0, \hat{p}}(X) \|\hat{f} - f_0\|_{2,2}] 
\]
\[\leq \mathbb{E}[\|p_0(X) - \hat{p}(X)\| \gamma_{f_0, \hat{p}}(X) \|\hat{f} - f_0\|_{2,2}] 
\]
\[\leq M^2 \|p_0 - \hat{p}\| \gamma_{f_0, \hat{p}}(X) \|\hat{f} - f_0\|_{2,2} 
\]
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 \(\delta_n\) 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. \(1 - \zeta\):
\[
|L_n(\hat{f}(\hat{p}) - L_n(f_0, \hat{p}) - (L_D(\hat{f}(\hat{p}) - L_D(f_0, \hat{p})))| \leq O(\delta_n, \zeta \|\ell_{\hat{p}, f} - \ell_{\hat{p}, f_0}\|_{2,2} + \delta_n^2, \zeta) 
\]
Moreover, if we let \(\mu(w) = \sup_{\|\|\|} \|\nabla_{\phi} \ell(w; \phi, \hat{p}(x))\|_{2,2}\) then we have by Cauchy-Schwartz inequality:
\[
\|\ell_{\hat{p}, f} - \ell_{\hat{p}, f_0}\|_{2,2} \leq \|\mu\| \|f - f_0\|_{2,2} 
\]
If we further assume that the function class \(\mathcal{F}\) satisfies an \(\ell_2/\ell_4\) condition that:
\[
\sup_{f \in \mathcal{F}} \|f - f_0\|_{2,2} \leq C \|f - f_0\|_{2,2} 
\]
then:
\[
\epsilon(\hat{f}, \hat{p}, \gamma) \leq O(\delta_n, C \|\mu\| \|f - f_0\|_{2,2} + \delta_n^2, \zeta) 
\]
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:
\[
\frac{\sigma}{8} \| \hat{f} - f_0 \|_{2,2}^2 \leq \frac{1}{\sigma} O(\delta_n^2, C \|\mu\| \|f - f_0\|_{2,2} + \|\gamma_{\hat{p}, 0}(\hat{p} - p_0)\|_{2,2}^2) 
\]
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 $\hat{\gamma}(t)$ the nuisance functions used on the samples in the $k$-th fold. For any $t \in [K]$ and conditional on $\hat{\gamma}(t)$, suppose that $\delta_n$ upper bounds the critical radius of the function class $\mathcal{G}(\hat{\gamma}(t))$, then by Lemma 11 of Foster & Syrgkanis (2019), if we denote with $\ell_{t,f}(w) = \ell_{\gamma(t)}(w; f, \hat{\gamma}(t)(x))$, w.p. $1-\zeta$:

$$L_{n,t}(\hat{f}, \hat{\gamma}(t)) - L_{n,t}(f_0, \hat{\gamma}(t)) - (L_D(\hat{f}, \hat{\gamma}(t)) - L_D(f_0, \hat{\gamma}(t))) \leq O(\delta_n/B, \zeta)$$

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

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

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

$$\frac{1}{B} \sum_{t=1}^B L_D(\hat{f}, \hat{\gamma}(t)) - L_D(f_0, \hat{\gamma}(t)) \leq O(\delta_n/B, \zeta)$$

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

$$\|\ell_t - \ell_{t,f_0}\|_{2,2} \leq \|\mu\|_4 f - f_0\|_{2,4} + \sqrt{E\left[\left((Y - \hat{p}(t)(X))^\top \hat{\gamma}(t)(X) (f(X) - f_0(X))\right)^2\right]}$$

$$\leq \|\mu\|_4 f - f_0\|_{2,4} + E\left[\left((Y - \hat{p}(t)(X))^\top \hat{\gamma}(t)(X)\right)^2\right] f - f_0\|_{2,4}$$

If we further assume that the function class $\mathcal{F}$ satisfies an $\ell_2/\ell_4$ condition that:

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

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

$$\frac{1}{B} \sum_{t=1}^B \ell(\hat{f}, \hat{p}(t), \hat{\gamma}(t)) \leq O(\delta_n/B, \zeta/B) \frac{1}{B} \sum_{t=1}^B C\left(\|\mu\|_4 + E\left[\left((Y - \hat{p}(t))(X))^\top \hat{\gamma}(t)(X)\right)^2\right]^{1/4}\right) f - f_0\|_{2,2} + \delta_n/B$$

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

$$\sigma^2 \frac{1}{4} \|\hat{f} - f_0\|_{2,2}^2 \leq \frac{1}{8} \sum_{t=1}^B \left(\|\gamma_{f_0\hat{p}(t)} - \hat{\gamma}(t)\|_{2,2}^2 + \|\gamma_{f_0\hat{p}(t)} - \hat{\gamma}(t)\|_{2,2}^2\right)$$

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

$$\sigma^2 \frac{1}{8} \|\hat{f} - f_0\|_{2,2}^2 \leq \frac{1}{8} \sum_{t=1}^B \left(\|\gamma_{f_0\hat{p}(t)} - \hat{\gamma}(t)\|_{2,2}^2 + \frac{1}{B} \sum_{t=1}^B E\left[\left((Y - \hat{p}(t))\right)^\top \hat{\gamma}(t)(X)\right]^{2}\right)$$

$$+ \frac{1}{8} \sum_{t=1}^B \left(\|\gamma_{f_0\hat{p}(t)} - \hat{\gamma}(t)\|_{2,2}^2\right).$$

F PROOF OF THM. 4: BIAISED 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 $\|v\|_\text{op} \triangleq \|v\|_2$ and $\|T\|_\text{op} \triangleq \sup_{\|v\|_2 = 1} \|T[v]\|_\text{op}$. Recall the definition

$$\nabla(W; \theta, p, \gamma) = \nabla_{\theta} f_0(X)^\top \nabla_{\phi} \ell_t(W; f_0(X), p(X))$$

$$= \nabla_{\theta} f_0(X)^\top (\nabla_{\phi} \ell(W; f_0(X), p(X)) + \gamma(X)^\top (Y - p(X))).$$

We apply the lemma with $L_{\gamma} = g$ and $g \in \mathcal{G}(\hat{\gamma}(t), \hat{\gamma}(t))$ and $g^* = 0$. Then we instantiate the concentration inequality with $g = \ell_{t,f} - \ell_{t,f_0} \in \mathcal{G}(\hat{\gamma}(t), \hat{\gamma}(t)).$
Observe that since $E[Y \mid X = x] = p_0(x)$, we can write for any $\gamma$:
\[
\mathcal{L}(\theta; p_0) = E[\ell(W; f_0(X), p_0(X)) + (Y - p_0(X))^\top \gamma(X) f_0(X)] = E[\ell_\gamma(W; f_0(X), p_0(X))]
\]
Thus we also have that:
\[
\forall \theta, \gamma : \nabla_\theta \mathcal{L}(\theta; p_0) = E[\nabla(W; \theta, p_0, \gamma)]
\]
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:
\[
\nabla(W; \theta, p, \gamma) = \nabla_\theta \mathcal{L}(\theta; p_0) + E[\nabla(W; \theta, p_0, \gamma)] - E[\nabla(W; \theta, p_0, \gamma)] - E[\nabla(W; \theta, p_0, \gamma)]
\]
The following two lemmas bound the gradient bias and noise terms.

**Lemma 5 (Gradient bias).** If $\sup_{p, \phi, \pi} \|E[\nabla_\pi \phi \ell(W; \phi, \pi) \mid X = x]\|_{op} \leq M$, then for any parameter vector $\theta$ and functions $p$ and $\gamma$, we have:
\[
\text{b}_i(\theta, p, \gamma) = E[\nabla_\theta f_\theta(X)^\top (\gamma_{f_\theta, p}(X) - \gamma(X))^\top (p(X) - p_0(X))],
\]
\[
\|\text{b}_i(\theta, p, \gamma)\|_2 \leq \left\|\nabla_\theta f_\theta^\top (\gamma_{f_\theta, p} - \gamma)^\top (p - p_0)\right\|_{2,2}, \quad \text{and}
\]
\[
\|\text{b}_i(\theta, p, \gamma)\|_2 \leq \left\|\nabla_\theta f_\theta^\top (q_{f_\theta, p} - \gamma)^\top (p - p_0)\right\|_{2,2} + \frac{M}{2} \|\nabla_\theta f_\theta\|_{F,2} \|p - p_0\|_{2,4}^2.
\]

**Proof** By Taylor’s theorem with integral remainder and Lagrange remainder respectively the SGD bias for each parameter $i$ takes the form
\[
\text{b}_i(\theta, p, \gamma) = E[\nabla_i(W; \theta, p, \gamma)] - E[\nabla_i(W; \theta, p_0, \gamma)] = E[(p(X) - p_0(X))^\top (\gamma_{f_\theta, p}(X) - \gamma(X))^\top (p(X) - p_0(X))]
\]
\[
E[\nabla_i(W; \theta, p, \gamma)] - E[\nabla_i(W; \theta, p_0, \gamma)] + E[\nabla_i(W; \theta, p_0, \gamma)] + E[\nabla_i(W; \theta, p_0, \gamma)]
\]
Furthermore, our operator norm assumption and Cauchy-Schwarz imply
\[
|\text{b}_i(\theta, p, \gamma)| \leq |E[(p(X) - p_0(X))^\top (\gamma_{f_\theta, p}(X) - \gamma(X))^\top (p(X) - p_0(X))]| + \frac{M}{2} E\left[\|\nabla_\theta f_\theta\|_{2,2} \|p - p_0\|_{2,4}^2\right]
\]
\[
\leq |E[(p(X) - p_0(X))^\top (\gamma_{f_\theta, p}(X) - \gamma(X))^\top (p(X) - p_0(X))]| + \frac{M}{2} E\left[\|\nabla_\theta f_\theta\|_{2,2} \|p - p_0\|_{2,4}^2\right].
\]
Thus, by the triangle inequality and Jensen’s inequality we find that
\[
\|\text{b}_i(\theta, p, \gamma)\|_2 \leq \left\|\nabla_\theta f_\theta^\top (\gamma_{f_\theta, p} - \gamma)^\top (p - p_0)\right\|_{2,2} \quad \text{and}
\]
\[
\|\text{b}_i(\theta, p, \gamma)\|_2 \leq \left\|\nabla_\theta f_\theta^\top (q_{f_\theta, p} - \gamma)^\top (p - p_0)\right\|_{2,2} + \frac{M}{2} \|\nabla_\theta f_\theta\|_{F,2} \|p - p_0\|_{2,4}^2.
\]

**Lemma 6 (Gradient Variance).** Define For any parameter $\theta$ and functions $p$ and $\gamma$,
\[
\sqrt{E[\|\nabla(W; \theta, p, \gamma)\|_{2,2}^2]} \leq \sigma_0(\theta) + \sqrt{E[\|\nabla_\theta f_\theta(X)^\top \gamma(X)^\top (Y - p_0(X))\|_{2,2}^2] + \|\nabla_\theta f_\theta^\top (\gamma_{f_\theta, p} - \gamma)^\top (p - p_0)\|_{2,2}^2}
\]

**Proof** For each $i \in [d]$, define the shorthand
\[
\Delta_i = \nabla_\theta \ell(W; f_\theta(X), p(X)) + \nabla_\theta f_\theta(X)^\top \gamma(X)^\top (Y - p(X)) - \nabla_\theta \ell(W; f_\theta(X), p_0(X)) \quad \text{and}
\]
\[
Z_i = E[\Delta_i | X]
\]
\[
= \nabla_\theta f_\theta(X)^\top (\gamma(X) - \gamma_{f_\theta, p}(X))^\top (p_0(X) - p(X))
\]
\[
= \nabla_\theta f_\theta(X)^\top (\gamma(X) - q_{f_\theta, p}(X))^\top (p_0(X) - p(X))
\]
\[
+ \frac{1}{2} E[\nabla_\pi \phi \ell(W; f_\theta(X), \tilde{p}(X))[\nabla_\theta f_\theta(X), p_0(X) - p(X), p(X) - p_0(X)]
\]
for some convex combination $\tilde{p}(X)$ of $p(X)$ and $p_0(X)$.
We next employ the law of total variance to rewrite the variance terms:
$$\mathbb{E}[\|n(W;\theta,p,\gamma)\|^2] = \sum_{i \in [d]} \text{Var}[\nabla_{\theta_i} \ell(W;f_{\theta}(X),p(X)) + \nabla_{\theta_i} f_{\theta}(X)^{\top} \gamma(X)^{\top} (Y-p(X))]$$
$$= \sum_{i \in [d]} \text{Var}[\nabla_{\theta_i} \ell(W;f_{\theta}(X),p_0(X)) + \Delta_i]$$
$$= \sigma_0(\theta,p_0)^2 + \sum_{i \in [d]} \text{Var}[\Delta_i] + 2\text{Cov}(\nabla_{\theta_i} \ell(W;f_{\theta}(X),p_0(X)),\Delta_i)$$
$$\leq \sigma_0(\theta,p_0)^2 + \sum_{i \in [d]} \text{Var}[\Delta_i] + 2\sqrt{\text{Var}[\nabla_{\theta_i} \ell(W;f_{\theta}(X),p_0(X))]}\text{Var}[\Delta_i]$$
$$\leq \sigma_0(\theta,p_0)^2 + (\sum_{i \in [d]} \text{Var}[\Delta_i])^{1/2} \sum_{i \in [d]} \text{Var}[\Delta_i]$$
$$= (\sigma_0(\theta,p_0) + \sqrt{\sum_{i \in [d]} \text{Var}[\Delta_i]})^{1/2}.$$

We next employ the law of total variance to rewrite the variance terms:
$$\sum_{i \in [d]} \text{Var}[\Delta_i] = \sum_{i \in [d]} \text{Var}[Z_i + \nabla_{\theta_i} f_{\theta}(X)^{\top} \gamma(X)^{\top} (Y-p_0(X))]$$
$$= \mathbb{E}[\|\nabla_{\theta} f_{\theta}(X)^{\top} \gamma(X)^{\top} (Y-p_0(X))\|_{2}^{2}] + \sum_{i \in [d]} \text{Var}[Z_i].$$

Finally, we control \(\text{Var}[Z_i]\) using Cauchy-Schwarz
$$\sqrt{\sum_{i \in [d]} \text{Var}[Z_i]} \leq \|\nabla_{\theta} f_{\theta}^{\top} (\gamma_{p_0-p} - \gamma)^{\top} (p-p_0)\|_{2,2}.$$

The two claims of Thm. 4 now follow from Theorems 2 and 3 of Ajalleloian & Stich (2020) respectively, with the parameters \(\sigma^2\) and \(\zeta\) 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 \(p_0(x) = p(y = 1 \mid x) = \frac{1}{1+e^{\theta^\top x}}\) 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 \(\lambda\) hyperparameter of the \(\gamma\)-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 \(\lambda\) hyperparameter of the \(\gamma\)-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 \(\log(\tilde{p}(x))\) of the teacher. When \(\tilde{p}(x) = 0\), we clip its value to \(10^{-3}\).

We also evaluate the impact of \(\gamma\) 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, \(\gamma\) 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.
Figure 4: Comparing knowledge distillation, with and without cross-fitting.

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 $\alpha$. The hyperparameter $\alpha$ controls the tradeoff between bias and variance. When $\alpha$ is very small, the objective is close to the vanilla KD objective. When $\alpha$ is large, the objective is closer to the Neyman-orthogonal loss. In Figure 5, we show the effect of varying $\alpha$, with ResNet-8 as the student and ResNet-20 as the teacher, on the CIFAR-10 dataset. Large values of $\alpha$ lead to high variance and thus lower test accuracy. Intermediate values of $\alpha$ improves on the vanilla KD objective, which corresponds to $\alpha = 0$. The test accuracy drops sharply beyond some threshold of $\alpha$ as the variance becomes too high (due to the terms $\hat{q}_\beta(x) = \text{diag}\left(\frac{1}{\hat{p}_1(x)}, \ldots, \frac{1}{\hat{p}_K(x)}\right)$), causing training to become unstable.
Figure 6: On CIFAR-10, large values of hyperparameter $\alpha$ (corresponding to orthogonal loss) is unstable, while intermediate values improve on vanilla KD.