SELECTIVE KNOWLEDGE DISTILLATION FROM PARTIALLY EXPERT TEACHERS: AN SGD ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Knowledge Distillation (KD) typically relies on a teacher whose predictions are assumed to provide useful supervision throughout the student's training data. In practice, pretrained teachers can be reliable on some inputs while being misleading on others. We study Stochastic Gradient Descent (SGD)-based KD from such *partially expert teachers* and ask, on a sample-by-sample basis, when the student should learn from the observed one-hot label and when it should learn from the teacher. We derive convergence bounds whose asymptotic error decomposes into competing sample-wise costs associated with the two supervision sources. Our analysis shows that the cost of one-hot supervision is governed by the intrinsic diversity of the Bayes Class Probabilities (BCPs), whereas the cost of teacher supervision increases with the teacher error relative to them, yielding a principled sample-wise routing rule. Since the underlying BCPs are unavailable in practice, we relate these quantities to predictive uncertainty: aleatoric uncertainty provides a proxy for BCP diversity, while epistemic uncertainty indicates limited teacher expertise. Building on this connection, we propose an uncertainty-guided selective distillation method that routes each sample between label- and teacher-based supervision. Experiments across vision, audio, and high-energy physics classification tasks with partially expert teachers demonstrate consistent improvements over state-of-the-art KD approaches.

1 INTRODUCTION

Knowledge Distillation (KD) (Hinton et al., 2015) has become a foundational method for training a (usually smaller) student model under the guidance of an expert teacher. In *response-based* distillation, the student matches the teacher's predictions, which, under the widely adopted Bayesian view (Menon et al., 2021), estimate the true Bayes Class Probabilities (BCPs) and give a richer training signal than one-hot labels. In practice, the teacher is commonly pretrained on a separate dataset, such that the teacher and the student share the same task, but their data distributions and the generalization demands often differ. The teacher may be competent where its own training data was rich and unreliable where it was sparse, guiding the student well on some inputs and misleading it on others. We refer to teachers whose expertise varies across the input space as *partially expert teachers (PETs)*.

To date, most KD work focuses on *how* to transfer the teacher's knowledge as faithfully as possible (Zhao et al., 2022; Wang et al., 2025; Sun et al., 2024), not *whether* teacher supervision is preferable for a particular sample. The same holds more broadly: theoretical works typically characterize when distillation helps or hurts on average (Safaryan et al., 2024; Liu et al., 2025), while others post-process the teacher distribution (Kim et al., 2026), analyze what makes a teacher effective for KD (Kim et al., 2025), or compare candidate teachers (Panigrahi et al., 2026), thus largely assessing the teacher as a whole. Yet limited data, distribution mismatch, and training imperfections may make the same teacher reliable on some inputs and inaccurate on others. For such a PET, student training involves two imperfect supervision signals: the observed one-hot label is a stochastic realization of the unknown BCPs, whereas the teacher provides a soft estimate of these probabilities that may itself be inaccurate. The relevant question is therefore not simply whether the teacher is reliable, but which supervision source is preferable for each sample.

Recent works have begun adapting the teacher's influence across samples using predictive confidence, uncertainty, and other sample-dependent indicators (Guo et al., 2024; Gore et al., 2025). However, such indicators do not by themselves explain *when* teacher supervision should be preferred over the observed label. This is particularly subtle for predictive uncertainty, whose different sources can

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

have distinct implications for KD. Aleatoric uncertainty (AU) reflects intrinsic ambiguity in the data distribution, where a single one-hot observation provides only one realization of the underlying BCPs, whereas an accurate soft teacher prediction can convey richer information. Epistemic uncertainty (EU) reflects the teacher's limited knowledge and may indicate that its soft prediction is unreliable. Thus, neither teacher confidence nor total predictive uncertainty alone provides a principled criterion for sample-wise distillation. This motivates characterizing the trade-off between label- and teacher-based supervision, together with observable quantities that can be used to implement this trade-off in practice.

Contributions In this work, we address these questions by developing a theoretical and practical framework for sample-wise KD from PETs. We first analyze the learning dynamics of Stochastic Gradient Descent (SGD) when each sample may be supervised either by its observed label or by the teacher, revealing the sample-dependent factors that govern the resulting convergence behavior. We then connect these latent factors to predictive uncertainty and use this connection to construct a practical uncertainty-guided distillation rule. Our main contributions are summarized as follows:

- **Sample-wise analysis of SGD-based distillation.** We characterize SGD-based learning from a PET, modeled as a biased and noisy estimate of the true BCPs whose reliability may vary across inputs. Our convergence bounds reveal that the asymptotic error separates into sample-wise contributions associated with one-hot and teacher supervision, yielding a principled routing rule that selects the supervision source with the smaller cost. We further characterize the factors governing these costs: the one-hot contribution increases with the intrinsic diversity of the BCPs, while the teacher contribution reflects both the variability of the teacher prediction and its deviation from the true BCPs. This rigorously explains why different supervision sources can be preferable on different data subsets.
- **Selective KD method.** The quantities governing the routing rule depend on the unknown BCPs and are not observable. We relate them to the predictive uncertainty of a Bayesian teacher, using AU as a proxy for BCP diversity and EU as an indicator of limited teacher knowledge, and validate this correspondence in controlled settings where the BCPs are known. We then propose an uncertainty-guided KD method that selects between one-hot and teacher supervision per sample.
- **Extensive evaluation with PETs.** We evaluate the proposed approach across vision, audio, and high-energy physics classification tasks under diverse PET settings. The resulting students consistently improve upon standard distillation, fixed teacher-label mixing, and state-of-the-art adaptive KD methods. We further demonstrate that the proposed routing mechanism is complementary to methods that modify the distillation objective and can be combined with them for additional gains.

2 Related Works

Theoretical Analysis of KD Several works have theoretically characterized why soft teacher supervision can improve upon one-hot training. From a statistical perspective, Menon et al. (2021) and Dao et al. (2021) relate the benefit of KD to how accurately the teacher approximates the true BCPs, while other analyses explain its generalization benefits through regularization, data geometry, or properties of the distillation objective (Phuong & Lampert, 2019; Yuan et al., 2020; Cha & Cho, 2026; Yeom et al., 2026). More closely related to our focus on learning dynamics: Safaryan et al. (2024) and Morad et al. (2026) analyze gradient-based distillation and show how soft supervision can reduce stochastic gradient variance; Ren et al. (2022) study the learning trajectories of individual samples and relate the quality of teacher supervision to sample difficulty; and Mandal et al. (2025) compare gradient-based learning from soft and hard labels, demonstrating benefits of soft supervision that depend on problem difficulty. These analyses, however, consider a fixed distillation policy and characterize its benefit through aggregate properties of the teacher or the learning problem. In contrast, we allow the distillation weight to vary across samples and analyze *which supervision source should be used for each input*. This exposes two competing sample-wise quantities governing SGD convergence, BCP diversity and teacher error, and yields a pointwise rule for when teacher supervision is preferable to the observed one-hot label.

Uncertainty in KD Uncertainty has been incorporated into KD in two main ways. One line of work aims to transfer the teacher's uncertainty-quantification capabilities to the student, e.g., by distilling Bayesian or ensemble predictive distributions or equipping the student itself with uncertainty estimates (Korattikara Balan et al., 2015; Gurau et al., 2018; Shen et al., 2021; Xu et al., 2023; Fang et al., 2024; Wang et al., 2026b; Huang et al., 2025). A second, more closely related

line uses uncertainty to regulate the distillation process, for instance through teacher entropy or confidence (Cheng et al., 2024; Guo et al., 2024; Zhang et al., 2026; Wang et al., 2026a; Zheng & Yang, 2024), or through Bayesian uncertainty-aware weighting (Sun et al., 2026). While most works treat uncertainty as a single scalar, Hemmatian et al. (2024) separately use EU and AU in sample-wise KD. Their method uses EU alone to reduce the teacher-label mixing coefficient when the teacher is uncertain, while AU independently controls the distillation temperature, with higher AU producing sharper teacher targets. In contrast, our SGD analysis shows that BCP diversity and teacher error jointly determine *whether teacher supervision is preferable to the observed one-hot label*: BCP diversity increases the cost of one-hot supervision, whereas teacher error increases the cost of distillation. We therefore use both AU and EU as proxies for these competing quantities to select the supervision source on a sample-wise basis, rather than assigning them independently to predefined KD hyperparameters.

Adaptive Mixing of Teacher and Ground-Truth Supervision A closely related line of work adapts the use of teacher supervision across samples. Some methods infer teacher reliability from its agreement with the ground-truth label (Zhou & Song, 2021; Mishra et al., 2026; Lan et al., 2025), while others use sample difficulty or the evolving student-teacher relationship (Ganguly et al., 2024; Li, 2025; Luong et al., 2026). Related sample-selection approaches instead decide *which samples* are most useful to distill (Zhang et al., 2020; Chae & Heo, 2025). More directly addressing the teacher-label balance, Lu et al. (2021) meta-learn the loss weights from a validation objective, while Hu et al. (2024) learn a sample-wise fusion ratio from the geometric relations among the teacher, student, and ground-truth predictions. These approaches determine how to use teacher supervision from empirical indicators of teacher quality, sample informativeness, or student learning. Our analysis instead asks what *fundamentally determines which supervision source is preferable for a given sample*: the resulting SGD convergence bounds identify BCP diversity and teacher error as the two competing quantities. Our uncertainty-guided method then uses AU and EU as practical proxies for these theoretically identified quantities.

3 Analysis of SGD-Based Learning from Partially Expert Teachers

This section analyzes SGD-based learning from a PET. We model the teacher predictions as biased and noisy estimates of the BCPs and, under standard regularity conditions, derive convergence bounds that expose the sample-wise trade-off between one-hot and teacher supervision. Our analysis yields a pointwise routing rule and characterizes the factors governing which supervision source is preferable. We conclude by numerically illustrating these findings. All proofs are provided in Appendix B.

3.1 KD Setup

Learning Framework We study classification in the standard supervised setting, where the input $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^n$ belongs to one of K classes. We write $\boldsymbol{y}_k \in \{0,1\}^K$ as the one-hot encoding of class k, i.e., the k-th standard basis vector, so that each label takes a value in $\{\boldsymbol{y}_1, \dots, \boldsymbol{y}_K\}$. A pair $(\boldsymbol{x}, \boldsymbol{y})$ is drawn from a data-generating distribution $\mathcal{P}$ (Shalev-Shwartz & Ben-David, 2014), and we write $\mathcal{P}(\boldsymbol{y}|\boldsymbol{x})$ for the induced conditional distribution over labels and call it the *true BCPs*. We seek a *student model* $\phi_{\boldsymbol{\theta}} : \mathcal{X} \to \mathcal{Y}$ (typically a Neural Network (NN)) with parameters $\boldsymbol{\theta} \in \mathbb{R}^d$, mapping each input onto the K-category simplex $\mathcal{Y}$. We aim to minimize the *generalization error* (risk), namely,

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} f_{\mathcal{P}}(\boldsymbol{\theta}) := \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{P}} \left[\ell(\phi_{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{y}) \right], \tag{1}$$

where $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is the Cross-Entropy (CE) loss. Following (1), the desired student model parameters are defined as $\boldsymbol{\theta}^* \in \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} f_{\mathcal{P}}(\boldsymbol{\theta})$ and their corresponding risk is $f^*_{\mathcal{P}} := f_{\mathcal{P}}(\boldsymbol{\theta}^*)$.

SGD-Based Distillation In standard supervised training, the student is trained to match the one-hot labels $\boldsymbol{y}$. In KD, the student also has access to the probability mass functions predicted by a teacher model, $\Phi(\boldsymbol{x}) \in \mathcal{Y}$, which are used as additional soft labels in the training process. The modified loss used for student training then becomes the linear combination $(1-\lambda)\ell(\phi_{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{y}) + \lambda\ell(\phi_{\boldsymbol{\theta}}(\boldsymbol{x}), \Phi(\boldsymbol{x}))$, where $\lambda \in [0,1]$. Prior analyses usually fix λ to a constant (typically $\lambda = 1$, full distillation, compared against $\lambda = 0$, no distillation). We instead let $\lambda = \lambda(\boldsymbol{x})$ vary per sample, and denote the student parameters trained with such an objective as $\bar{\boldsymbol{\theta}}$. We initially consider the general case $\lambda(\boldsymbol{x}) \in [0,1]$, which includes sample-wise soft mixing of label and teacher supervision. Our main analysis then specializes to *selective distillation*, where $\lambda(\boldsymbol{x}) \in \{0,1\}$ and each sample is supervised either by its observed label or by the teacher.

Since the CE loss is linear in its second argument, the modified loss for data sample $(\boldsymbol{x}, \boldsymbol{y})$ becomes $\bar{f}_{(\boldsymbol{x},\boldsymbol{y}),\lambda}(\bar{\boldsymbol{\theta}}) := \ell(\phi_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}), (1-\lambda(\boldsymbol{x}))\boldsymbol{y} + \lambda(\boldsymbol{x})\Phi(\boldsymbol{x}))$. Given a dataset $\mathcal{D}$ composed of i.i.d. samples drawn from $\mathcal{P}$, the standard SGD-based KD framework is applied by iterating

$$\bar{\boldsymbol{\theta}}^{t+1} = \bar{\boldsymbol{\theta}}^t - \alpha \nabla \bar{f}_{(\boldsymbol{x}_t,\boldsymbol{y}_t),\lambda}(\bar{\boldsymbol{\theta}}^t), \tag{2}$$

where $\alpha > 0$ is the learning rate, and $(\boldsymbol{x}_t, \boldsymbol{y}_t) \sim \mathcal{D}$ is a randomly sampled data point or mini-batch.

Under the learning framework of (2), the risk that is *effectively* being minimized is

$$\min_{\bar{\boldsymbol{\theta}} \in \mathbb{R}^d} \bar{f}_{\mathcal{P},\lambda}(\bar{\boldsymbol{\theta}}) := \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{P}} \left[\ell(\phi_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}), (1-\lambda(\boldsymbol{x}))\boldsymbol{y} + \lambda(\boldsymbol{x})\Phi(\boldsymbol{x}))\right]. \tag{3}$$

We define $\bar{f}_{\mathcal{P},\lambda}(\bar{\boldsymbol{\theta}})$ as the *distillation risk*, with its own minimum $\bar{f}^*_{\mathcal{P},\lambda}$ and minimizer $\bar{\boldsymbol{\theta}}^*_\lambda \in \arg\min_{\bar{\boldsymbol{\theta}}\in\mathbb{R}^d} \bar{f}_{\mathcal{P},\lambda}(\bar{\boldsymbol{\theta}})$. The distillation risk (3) reduces to the true risk (1) when $\lambda(\boldsymbol{x}) = 0$ for all $\boldsymbol{x} \in \mathcal{X}$.

Analytical PET Model We next introduce a tractable probabilistic model for a PET, which serves as an abstraction for separately capturing two forms of teacher imperfection: systematic deviation from the true BCPs and random variability around that deviation. In our analysis, we model

$$\Phi(\boldsymbol{x}) \sim \mathrm{Dir}(\varepsilon(\boldsymbol{x}) \left[\mathcal{P}(\boldsymbol{y}|\boldsymbol{x}) + \mathbf{b}(\boldsymbol{x})\right]). \tag{4}$$

Here, the randomness in $\Phi(\boldsymbol{x})$ represents variability in the teacher's estimate at input $\boldsymbol{x}$, e.g., due to limited training data, model uncertainty, or randomness in the training procedure. Accordingly, whenever an expectation in the analysis involves $\Phi(\boldsymbol{x})$, it is also taken over the conditional randomness in (4). This model is introduced for the purpose of the analysis and is not intended to imply that the predictions of a trained teacher necessarily follow a Dirichlet distribution.

A Dirichlet draw lies on the simplex by construction, so the teacher output is always a valid distribution. Its mean is $\mathcal{P}(\boldsymbol{y}|\boldsymbol{x}) + \mathbf{b}(\boldsymbol{x})$, so $\mathbf{b}(\boldsymbol{x})$ is the teacher's systematic bias and $\varepsilon(\boldsymbol{x})$ controls the spread around that mean, with larger $\varepsilon(\boldsymbol{x})$ meaning a more concentrated, less noisy teacher. The bias is constrained to be zero-sum, $\mathbf{1}^\top \mathbf{b}(\boldsymbol{x}) = 0$, and to keep $\mathcal{P}(\boldsymbol{y}|\boldsymbol{x}) + \mathbf{b}(\boldsymbol{x})$ positive, so it redistributes probability mass without leaving the simplex. Allowing both $\varepsilon(\boldsymbol{x})$ and $\mathbf{b}(\boldsymbol{x})$ to depend on the input yields a PET whose reliability varies across the input space. A perfect teacher is recovered in the limiting case of zero bias and infinite concentration. We say that the teacher is more expert on $\boldsymbol{x}_1$ than on $\boldsymbol{x}_2$ when it is both less biased and less noisy there, i.e., $\|\mathbf{b}(\boldsymbol{x}_1)\| < \|\mathbf{b}(\boldsymbol{x}_2)\|$ and $\varepsilon(\boldsymbol{x}_1) > \varepsilon(\boldsymbol{x}_2)$.

Assumptions To obtain explicit convergence guarantees, we adopt the following regularity and realizability conditions. Variants of these assumptions are standard in analyses of stochastic gradient methods and KD (Garrigos & Gower, 2023; Safaryan et al., 2024; Morad et al., 2026). They are used here as analytical conditions that isolate the effect of the supervision source on the SGD dynamics, rather than as claims that arbitrary NNs satisfy them globally.

AS1 (Strong quasi-convexity). The risk function $\bar{f}_{\mathcal{P},\lambda} : \mathbb{R}^d \to \mathbb{R}$ is differentiable and μ-strongly quasi-convex for some constant $\mu > 0$, i.e., for any $\bar{\boldsymbol{\theta}} \in \mathbb{R}^d$ it holds that

$$\bar{f}^*_{\mathcal{P},\lambda} \geq \bar{f}_{\mathcal{P},\lambda}(\bar{\boldsymbol{\theta}}) + \langle \nabla \bar{f}_{\mathcal{P},\lambda}(\bar{\boldsymbol{\theta}}), \bar{\boldsymbol{\theta}}^* - \bar{\boldsymbol{\theta}} \rangle + \frac{\mu}{2}\|\bar{\boldsymbol{\theta}}^* - \bar{\boldsymbol{\theta}}\|^2.$$

Assumption *AS1* provides sufficient geometry around the minimizer for the SGD iterates to contract toward it. It serves the standard role of enabling a tractable convergence characterization and can also be viewed as describing the relevant optimization region around a solution.

AS2 (Smoothness). The loss $\bar{f}_{\xi,\lambda}(\bar{\boldsymbol{\theta}})$ is differentiable and $\mathcal{L}$-smooth in expectation (for some constant $\mathcal{L}$) for $\xi = (\boldsymbol{x}_t, \boldsymbol{y}_t) \sim \mathcal{P}$, i.e., for any $\bar{\boldsymbol{\theta}} \in \mathbb{R}^d$ it holds that

$$\mathbb{E}_{\xi\sim\mathcal{P}} \left[\|\nabla \bar{f}_{\xi,\lambda}(\bar{\boldsymbol{\theta}}) - \nabla \bar{f}_{\xi,\lambda}(\bar{\boldsymbol{\theta}}^*)\|^2\right] \leq 2\mathcal{L}(\bar{f}_{\mathcal{P},\lambda}(\bar{\boldsymbol{\theta}}) - \bar{f}^*_{\mathcal{P},\lambda}).$$

Moreover, the risk $\bar{f}_{\mathcal{P},\lambda}$ is L-smooth with $L > 0$, i.e., $\|\nabla \bar{f}_{\mathcal{P},\lambda}(\bar{\boldsymbol{\theta}}_1) - \nabla \bar{f}_{\mathcal{P},\lambda}(\bar{\boldsymbol{\theta}}_2)\| \leq L\|\bar{\boldsymbol{\theta}}_1 - \bar{\boldsymbol{\theta}}_2\|$ for any $\bar{\boldsymbol{\theta}}_1, \bar{\boldsymbol{\theta}}_2 \in \mathbb{R}^d$.

Assumption *AS2* controls how rapidly the stochastic gradients can change and is a standard regularity condition in SGD convergence analyses. Such smoothness conditions can be justified, for example, over bounded optimization regions when the model has sufficiently regular activations and bounded

derivatives (Hardt et al., 2016). We adopt it here to characterize the optimization dynamics rather than to assert global smoothness of arbitrary NNs. To compare different sample-wise distillation policies, we assume that the constants hold uniformly over the considered mappings $\lambda : \mathcal{X} \to [0, 1]$. In particular, the same constants μ, $\mathcal{L}$, and L are taken to satisfy Assumptions *AS1–AS2* for all such λ.

AS3 (Student expressiveness). There exists $\boldsymbol{\theta} \in \mathbb{R}^d$ such that $\phi_{\boldsymbol{\theta}}$ attains the pointwise minimizer of the distillation risk (3) over all maps $\mathcal{X} \to \mathcal{Y}$.

Assumption *AS3* is a realizability condition that abstracts away approximation error of the student model. It isolates the effect of choosing between one-hot and teacher supervision from the properties of the student's function class.

3.2 Convergence Analysis

Regardless of whether KD is performed, the desired student model is one that minimizes the true risk (1). Nevertheless, the distillation risk (3) that is effectively being minimized can differ from the true risk in its loss landscape and minimizer. We thus first identify what the student converges to under KD.

Proposition 1. *Let Assumption* AS3 *hold. The inference rule that minimizes the distillation risk (3) is* $\phi(\boldsymbol{x}) \equiv [\mathcal{P}(\boldsymbol{y}_1|\boldsymbol{x}) + \lambda(\boldsymbol{x})\mathbf{b}_1(\boldsymbol{x}), \ldots, \mathcal{P}(\boldsymbol{y}_K|\boldsymbol{x}) + \lambda(\boldsymbol{x})\mathbf{b}_K(\boldsymbol{x})]$.

Proposition 1 shows that distillation shifts the student's prediction away from the true BCPs by $\lambda(\boldsymbol{x})\mathbf{b}(\boldsymbol{x})$. Thus, teacher supervision may introduce bias relative to one-hot training. This observation alone, however, does not determine which supervision is preferable under SGD: while one-hot labels are unbiased realizations of the BCPs, their stochasticity also affects the learning dynamics, as shown next.

We now derive convergence bounds for the weights and for the true risk of the student produced by SGD-based distillation (2) with a PET. As the optimum of (3) is generally biased (Proposition 1), we seek to bound distances from the *true risk minimizer*, i.e., the weight distance $\|\bar{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^*\|^2$ and the excess true risk $f(\bar{\boldsymbol{\theta}}^t) - f(\boldsymbol{\theta}^*)$. We now specialize to the selective-distillation setting introduced above, $\lambda(\boldsymbol{x}) \in \{0, 1\}$, and derive convergence bounds that compare the supervision choices for each input. To that end, define

$$D(\boldsymbol{x}) := \sum_{k=1}^{K} \frac{\mathbf{b}_k(\boldsymbol{x})}{\mathcal{P}(\boldsymbol{y}_k|\boldsymbol{x})^2} \, \|J_{\bar{\boldsymbol{\theta}},k}[\phi_{\boldsymbol{\theta}^*}(\boldsymbol{x})]\|^2 + \varepsilon(\boldsymbol{x}) \Big\| \sum_{k=1}^{K} \frac{\mathbf{b}_k(\boldsymbol{x})}{\mathcal{P}(\boldsymbol{y}_k|\boldsymbol{x})} J_{\bar{\boldsymbol{\theta}},k}[\phi_{\boldsymbol{\theta}^*}(\boldsymbol{x})] \Big\|^2,$$

$$C(\boldsymbol{x}) := \sum_{k=1}^{K} \frac{\|J_{\bar{\boldsymbol{\theta}},k}[\phi_{\boldsymbol{\theta}^*}(\boldsymbol{x})]\|^2}{\mathcal{P}(\boldsymbol{y}_k|\boldsymbol{x})}, \qquad B(\boldsymbol{x}) := \frac{1}{\varepsilon(\boldsymbol{x}) + 1} \big(C(\boldsymbol{x}) + D(\boldsymbol{x})\big),$$

$$A(\boldsymbol{x}) := D_{KL}\big(\mathcal{P}(\boldsymbol{y}|\boldsymbol{x}) \,\|\, \mathcal{P}(\boldsymbol{y}|\boldsymbol{x}) + \mathbf{b}(\boldsymbol{x})\big), \qquad \sigma^*_{\bar{f}}(\lambda) := \mathbb{E}\big[(1 - \lambda(\boldsymbol{x}))\, C(\boldsymbol{x}) + \lambda(\boldsymbol{x})\, B(\boldsymbol{x})\big],$$

where $J_{\bar{\boldsymbol{\theta}},k}[\phi_{\boldsymbol{\theta}^*}(\boldsymbol{x})]$ denotes the k-th column of the Jacobian of the student model evaluated at $\boldsymbol{\theta}^*$, i.e., $\nabla_{\bar{\boldsymbol{\theta}}}\phi(\boldsymbol{x}) = [J_{\bar{\boldsymbol{\theta}},1}[\phi_{\boldsymbol{\theta}^*}(\boldsymbol{x})], \ldots, J_{\bar{\boldsymbol{\theta}},K}[\phi_{\boldsymbol{\theta}^*}(\boldsymbol{x})]]$. We now obtain the following bounds:

Theorem 1. *Let Assumptions* AS1*,* AS2*, and* AS3 *hold. For any step-size* $\alpha \leq \frac{1}{2\mathcal{L}}$*, the weights produced by SGD-based distillation (2) converge as*

$$\mathbb{E}\big[\|\bar{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^*\|^2\big] \leq 2(1 - \alpha\mu)^t \|\bar{\boldsymbol{\theta}}^0 - \bar{\boldsymbol{\theta}}^*\|^2 + \frac{8\alpha}{\mu}\, \sigma^*_{\bar{f}}(\lambda) + \frac{16\mathcal{L}}{\mu^2} \Big(1 + \frac{4L\mathcal{L}\alpha}{\mu}\Big) \mathbb{E}\big[\lambda(\boldsymbol{x})\, A(\boldsymbol{x})\big]. \quad (5)$$

Theorem 2. *Let Assumptions* AS1*,* AS2*, and* AS3 *hold. For any step-size* $\alpha \leq \frac{1}{2\mathcal{L}}$*, the true risk of the student produced by SGD-based distillation (2) converges as*

$$\mathbb{E}\big[f(\bar{\boldsymbol{\theta}}^t) - f(\boldsymbol{\theta}^*)\big] \leq L(1 - \alpha\mu)^t \|\bar{\boldsymbol{\theta}}^0 - \bar{\boldsymbol{\theta}}^*\|^2 + \frac{4L\alpha}{\mu}\, \sigma^*_{\bar{f}}(\lambda) + \frac{8L\mathcal{L}}{\mu^2} \Big(1 + \frac{4L\mathcal{L}\alpha}{\mu}\Big) \mathbb{E}\big[\lambda(\boldsymbol{x})\, A(\boldsymbol{x})\big]. \quad (6)$$

The resulting bounds decompose into a transient convergence term, which decays geometrically in t from the initialization distance to the distillation minimizer $\|\bar{\boldsymbol{\theta}}^0 - \bar{\boldsymbol{\theta}}^*\|^2$, and a persistent asymptotic neighborhood term. Specifically, in Theorems 1–2, $\sigma^*_{\bar{f}}(\lambda)$ encapsulates the *gradient noise* (Garrigos & Gower, 2023), which here depends on the routing λ, with $C(\boldsymbol{x})$ and $B(\boldsymbol{x})$ representing the one-hot and teacher contributions to it, and $A(\boldsymbol{x})$ captures the KL divergence between the true BCPs and the teacher's mean prediction. Letting $\mathbb{1}\{\cdot\}$ be the indicator function, we next show how to choose $\lambda(\boldsymbol{x})$ per sample to minimize the bound's asymptotic error:

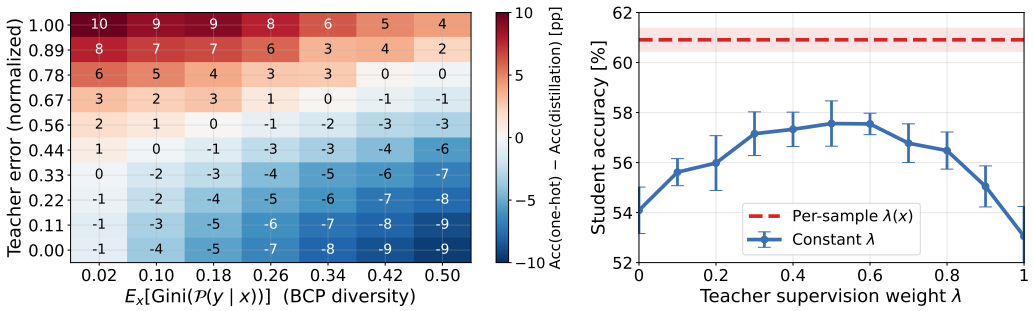

Figure 1: (Left) Each cell reports the test accuracy of a student trained on one-hot labels minus that of a student trained by distillation (pp), on a synthetic task, as expected BCP diversity (x-axis) and teacher error (y-axis) are varied. Red: one-hot student is more accurate; blue: distilled student is more accurate. (Right) KD with a PET under constant values of λ and under the per-sample rule $\lambda(\boldsymbol{x})$ of Proposition 2. Bars and band are one standard deviation over 10 seeds.

Proposition 2. *Let* $c_1 = \frac{4L\alpha}{\mu}$ *and* $c_2 = \frac{8L\mathcal{L}}{\mu^2}\left(1 + \frac{4L\mathcal{L}\alpha}{\mu}\right)$. *For each* $\boldsymbol{x}$ *define the one-hot and teacher costs* $c_{\mathrm{OH}}(\boldsymbol{x}) := c_1\, C(\boldsymbol{x})$ *and* $c_{\mathrm{T}}(\boldsymbol{x}) := c_1\, B(\boldsymbol{x}) + c_2\, A(\boldsymbol{x})$. *The asymptotic error bounds of Theorems 1 and 2 are minimized over routings* $\lambda : \mathcal{X} \to \{0, 1\}$ *by setting* $\lambda^*(\boldsymbol{x}) = \mathbb{1}\{\, c_{\mathrm{T}}(\boldsymbol{x}) < c_{\mathrm{OH}}(\boldsymbol{x}) \,\}$.

3.3 Interpreting the Routing Costs

Proposition 2 sets λ by comparing two costs: $c_{\mathrm{OH}}(\boldsymbol{x})$, associated with one-hot supervision, and $c_{\mathrm{T}}(\boldsymbol{x})$, associated with teacher supervision. We next show that these costs admit intuitive interpretations in terms of BCP diversity and the teacher's expertise at the given input.

BCP Diversity We first examine the one-hot cost $c_{\mathrm{OH}}(\boldsymbol{x})$. For a linear softmax head, its dependence on the true BCP can be characterized explicitly, as stated in the following proposition:

Proposition 3. *Let Assumption* AS3 *hold. For a linear softmax head on fixed features* $h(\boldsymbol{x})$ *independent of* $\bar{\boldsymbol{\theta}}$, $\phi_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}) = \mathrm{softmax}\big(Wh(\boldsymbol{x}) + \beta\big)$ *with* $\bar{\boldsymbol{\theta}} = (W, \beta)$, *at any input with* $\lambda(\boldsymbol{x}) = 0$ *the one-hot cost is*

$$c_{\mathrm{OH}}(\boldsymbol{x}) = c_1\big(\|h(\boldsymbol{x})\|^2 + 1\big) \underbrace{\left(1 - \sum_{k=1}^{K} \mathcal{P}(\boldsymbol{y}_k|\boldsymbol{x})^2\right)}_{\textit{Gini index of } \mathcal{P}(\boldsymbol{y}|\boldsymbol{x})}. \tag{7}$$

The Gini index (Rao, 1982) measures BCP diversity and equals the total conditional variance of the one-hot label, $\sum_k \mathrm{Var}[\boldsymbol{y}_k|\boldsymbol{x}]$. For a fixed representation, $c_{\mathrm{OH}}(\boldsymbol{x})$ thus grows with the intrinsic ambiguity of the sample and vanishes when the true BCP is one-hot. Ambiguous samples therefore incur greater stochasticity under one-hot supervision and offer greater potential for soft teacher supervision to improve convergence. While Proposition 3 considers a linear head, Appendix C shows empirically that the relation between BCP diversity and the one-hot cost persists for multi-layer NNs.

Teacher Error The teacher cost $c_{\mathrm{T}}(\boldsymbol{x}) = c_1 B(\boldsymbol{x}) + c_2 A(\boldsymbol{x})$ reflects two forms of teacher imperfection: $B(\boldsymbol{x})$ captures stochastic variability in its prediction, while $A(\boldsymbol{x})$ measures the systematic deviation of its mean prediction from the BCP. Accordingly, as the teacher becomes more expert on $\boldsymbol{x}$ (smaller $\|\mathbf{b}(\boldsymbol{x})\|$ and larger $\varepsilon(\boldsymbol{x})$), both contributions decrease and $c_{\mathrm{T}}(\boldsymbol{x}) \to 0$. Proposition 2 thus exposes a sample-wise trade-off: distillation is most favorable where the true BCP is diverse, making one-hot supervision costly, yet the teacher remains reliable, making teacher supervision inexpensive. Conversely, decisive samples favor one-hot supervision, while unreliable teacher predictions make distillation harmful. This indicates that a PET calls for selective rather than uniform distillation.

Numerical Validation We illustrate these implications using a synthetic setting where the BCPs are known and both their diversity and teacher expertise can be controlled. Figure 1 (left) shows the transition between regimes where one-hot supervision and distillation are preferable as these two quantities vary, while Figure 1 (right) demonstrates that the sample-wise routing rule of Proposition 2 outperforms every constant routing. Experimental details and additional validation are provided in Appendix C.

4 BAYESIAN UNCERTAINTY-GUIDED KD

The analysis of Section 3 yields a sample-wise routing rule, but its key quantities depend on the unknown BCP diversity and teacher error and are therefore unavailable in practice. Bayesian teachers have been motivated in KD due to their calibrated probabilistic supervision (Kim et al., 2025; Fan et al., 2024; Morad et al., 2026). Here, we exploit an additional advantage: their predictive uncertainty can be decomposed into AU and EU, which can serve as observable proxies for BCP diversity and teacher error, respectively. We use these uncertainty components to construct a practical PET-aware distillation rule.

4.1 UNCERTAINTY ESTIMATES AS PROXIES

Proposition 2 identifies latent factors that govern the sample-wise routing decision: the intrinsic diversity of the BCPs, which increases the cost of one-hot supervision, and the reliability of the teacher prediction, which determines the cost of distillation. Neither quantity is directly observable, since both depend on the unknown BCPs. We therefore seek measurable quantities that reflect their variation across inputs. A Bayesian teacher provides such information through its predictive uncertainty. Given a posterior $q(\boldsymbol{\theta}')$ over the teacher parameters, predictive uncertainty admits the standard decomposition (Kendall & Gal, 2017; Depeweg et al., 2018)

$$\underbrace{\mathrm{H}\big(\Phi(\boldsymbol{x})\big)}_{\mathrm{TU}(\boldsymbol{x})} = \underbrace{\mathbb{E}_{q(\boldsymbol{\theta}')}\left[\mathrm{H}(\Phi_{\boldsymbol{\theta}'}(\boldsymbol{x}))\right]}_{\mathrm{AU}(\boldsymbol{x})} + \underbrace{\mathrm{I}(\boldsymbol{y};\boldsymbol{\theta}' \mid \boldsymbol{x})}_{\mathrm{EU}(\boldsymbol{x})}. \tag{8}$$

Here, AU captures ambiguity that persists across plausible teacher parameters, whereas EU captures disagreement arising from uncertainty about the teacher parameters themselves (Gawlikowski et al., 2023). These quantities naturally relate to the two factors identified by our analysis: for a well-calibrated teacher, AU tends to be larger where the underlying class distribution is intrinsically diverse, while EU tends to increase where the teacher has limited knowledge and its prediction is therefore less reliable. We consequently use AU and EU as observable proxies for BCP diversity and teacher error, respectively, and validate this correspondence empirically below.

The required predictive distribution can be obtained from a teacher trained as a Bayesian Neural Network (BNN), or approximated from an existing deterministic teacher. In our experiments we consider Laplace Approximations (LAs) as well as Variational Inference (VI) and deep ensembles (ens); notably, LA enables converting a pretrained deterministic teacher into an approximate Bayesian one without retraining it from scratch (Ritter et al., 2018). Given M parameter samples $\boldsymbol{\theta}'_m \sim q(\boldsymbol{\theta}')$ and predictions $\Phi_m(\boldsymbol{x}) = \Phi_{\boldsymbol{\theta}'_m}(\boldsymbol{x})$, we estimate

$$\widehat{\mathrm{TU}}(\boldsymbol{x}) = \mathrm{H}\Big(\tfrac{1}{M}\sum_{m=1}^{M}\Phi_m(\boldsymbol{x})\Big),\ \widehat{\mathrm{AU}}(\boldsymbol{x}) = \tfrac{1}{M}\sum_{m=1}^{M}\mathrm{H}\left(\Phi_m(\boldsymbol{x})\right),\ \widehat{\mathrm{EU}}(\boldsymbol{x}) = \widehat{\mathrm{TU}}(\boldsymbol{x}) - \widehat{\mathrm{AU}}(\boldsymbol{x}). \tag{9}$$

Further details on the Bayesian approximations are provided in Appendix D.

The uncertainty measures are observable proxies for the two latent factors that drive the analytical costs $c_{\mathrm{OH}}(\boldsymbol{x})$ and $c_{\mathrm{T}}(\boldsymbol{x})$ across inputs, BCP diversity and teacher error. The costs also carry model-dependent factors, as in Proposition 3, which is why the decision point is set from data in Section 4.2. We examine this correspondence in controlled synthetic settings where the true BCPs are known. Figure 2 illustrates such a setting, in which the training density of an LA teacher decreases away from the center of the input domain. The estimated AU tracks the spatial variation of the BCP diversity, while the estimated EU increases in regions where the teacher deviates more strongly from the BCPs. Additional settings and quantitative validation are reported in Appendix E.

4.2 METHOD

The analysis in Section 3 determines the direction of the sample-wise trade-off, but does not provide an observable decision boundary: larger BCP diversity favors teacher supervision, whereas lower teacher reliability favors the label. Following Section 4.1, we use AU and EU as observable indicators of these two factors. We combine them through the *EU fraction*

$$r(\boldsymbol{x}) = \frac{\widehat{\mathrm{EU}}(\boldsymbol{x})}{\widehat{\mathrm{EU}}(\boldsymbol{x}) + \widehat{\mathrm{AU}}(\boldsymbol{x})}. \tag{10}$$

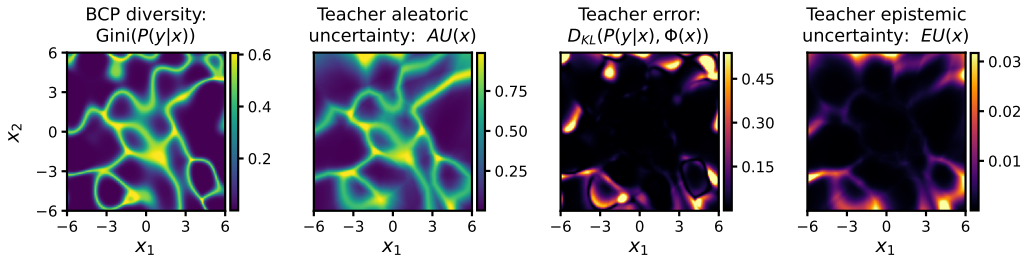

Figure 2: BNN trained with data density leading to unreliable predictions at the edges. Left to right: latent BCP diversity; teacher AU; latent teacher error (larger for low training density); and teacher EU.

Algorithm 1 Uncertainty-Guided Distillation

for each $\boldsymbol{x} \in \mathcal{D}$ **do**
  for $m = 1$ **to** M **do**
    $\Phi_m(\boldsymbol{x}) \leftarrow$ forward pass with $\boldsymbol{\theta}'_m \sim q(\boldsymbol{\theta}')$
  compute $\Phi(\boldsymbol{x})$, $\mathrm{AU}(\boldsymbol{x})$, $\mathrm{EU}(\boldsymbol{x})$ from $\{\Phi_m(\boldsymbol{x})\}_{m=1}^{M}$ ▷ equation 9
  $r(\boldsymbol{x}) \leftarrow \mathrm{EU}(\boldsymbol{x}) / \big(\mathrm{EU}(\boldsymbol{x}) + \mathrm{AU}(\boldsymbol{x})\big)$ ▷ equation 10
Fit a two-component GMM to $\{r(\boldsymbol{x})\}_{\boldsymbol{x}\in\mathcal{D}}$ and denote its lower-mean component by $\mathcal{C}_{\mathrm{T}}$
for epoch $= 1$ **to** E **do**
  for each $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{D}$ **do**
    Set $\lambda(\boldsymbol{x}) \leftarrow \mathbb{1}\{r(\boldsymbol{x}) \in \mathcal{C}_{\mathrm{T}}\}$
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \nabla_{\boldsymbol{\theta}} \big[(1 - \lambda(\boldsymbol{x}))\, \ell_{\mathrm{CE}}\big(\phi_{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{y}\big) + \lambda(\boldsymbol{x})\, \ell_{\mathrm{KD}}\big(\phi_{\boldsymbol{\theta}}(\boldsymbol{x}), \Phi(\boldsymbol{x})\big)\big]$ ▷ equation 2

As $r(\boldsymbol{x})$ increases with EU and decreases with AU, low values favor distillation, while high ones favor one-hot supervision. The theory does not prescribe a threshold on $r(\boldsymbol{x})$, as the uncertainty measures are proxies for the latent costs that also depend on unknown model-dependent factors. We thus use a simple unsupervised rule to set the threshold from the student's training set: a two-component Gaussian mixture model (GMM) is fitted to $\{r(\boldsymbol{x})\}_{\boldsymbol{x}\in\mathcal{D}}$, and the component with the lower mean is assigned to teacher supervision and the other to one-hot supervision. This yields the selective routing $\lambda(\boldsymbol{x}) \in \{0, 1\}$, as summarized in Algorithm 1. The GMM serves as an adaptive thresholding mechanism rather than as a consequence of the theoretical analysis; alternative thresholding rules are considered in Appendix G.5.

4.3 Numerical Study

We evaluate our method on three classification tasks with PETs, training student networks via SGD. In each task the student's data spans two conditions while the teacher, shared by all methods within a column, is trained on only one of them, making it a PET. Full experimental settings and baseline implementations are in Appendix F, and the code is available at `https://github.com/Selective-KD`.

Datasets and PETs For vision, we use CIFAR-10 with two CIFAR-10-C corruptions (Hendrycks & Dietterich, 2019) as the conditions. For audio, we use UrbanSound8K (Salamon et al., 2014), where the conditions differ by a degradation applied to the input. For high-energy physics, we use JetClass (Qu et al., 2022), where the conditions are two bands of jet transverse momentum.

Baselines We divide the state-of-the-art KD methods we compare against into two groups. The first group improves how the teacher's prediction is transferred, with one distillation weight shared by all samples: DKD decouples the target-class and non-target-class terms of the distillation loss; ABKD replaces the Kullback–Leibler (KL) divergence with an α-β-divergence; LSD standardizes the logits of both models before the softmax; and Beta-KD learns the distillation weight as the precision of a teacher-informed prior. The second group, like ours, sets $\lambda(\boldsymbol{x})$ from a per-sample signal: teacher confidence in LUD, the teacher's EU in UBKD, teacher entropy in WTTM, agreement with the label in MCD, the student-to-teacher loss ratio in WLS, a learned map of the student, teacher, and label geometry in TGeo-KD, and gradient alignment in RW-KD. Our comparison also includes one-hot training ($\lambda = 0$), pure distillation ($\lambda = 1$), and the best constant λ from a grid search.

Table 1: Accuracy (%) of the student and the PET on the student's test set, for teacher-student pairs with matching architectures. All methods within a column use the same PET. Results are mean (std) over 3 runs on CIFAR-10 and 10 otherwise. Best results are in **bold** and second best are underlined.

Modality	JetClass	CIFAR-10					UrbanSound8K		
Partially expert teacher	MLP-256	WRN-16-2	ResNet-32	ResNet-20	WRN-40-2	ConvNet-6	CNN-8	ResNet-14	CNN-8-N
Student	MLP-64	WRN-16-1	ResNet-20	ResNet-20	WRN-40-1	ConvNet-3	CNN-4	ResNet-8	CNN-4-N
Partially expert teacher	58.67 (1.09)	68.08 (0.58)	68.50 (0.44)	67.44 (0.70)	69.72 (0.41)	65.16 (0.17)	60.58 (2.01)	60.61 (2.33)	61.77 (2.46)
Standard one-hot training	58.14 (0.78)	78.66 (0.16)	80.40 (0.28)	80.30 (0.62)	82.01 (0.24)	69.26 (0.16)	60.10 (2.43)	51.30 (1.37)	57.13 (2.76)
KD Hinton et al. (2015)	56.35 (0.90)	65.27 (0.12)	66.27 (0.15)	65.90 (0.19)	66.23 (0.07)	58.92 (0.36)	56.69 (0.65)	49.47 (1.19)	53.82 (1.25)
Best constant λ	60.56 (0.53)	78.66 (0.16)	80.40 (0.28)	80.30 (0.62)	84.06 (0.32)	69.49 (0.21)	62.72 (1.62)	54.75 (2.15)	58.78 (1.76)
DKD Zhao et al. (2022)	60.76 (0.88)	72.52 (0.19)	75.29 (0.32)	72.17 (1.38)	81.76 (0.28)	67.01 (0.66)	61.81 (1.60)	53.02 (1.45)	58.54 (1.03)
ABKD Wang et al. (2025)	61.14 (0.55)	78.72 (0.36)	81.17 (0.17)	79.88 (0.35)	83.60 (0.38)	68.42 (0.31)	62.65 (0.84)	53.09 (1.76)	59.24 (1.35)
LSD Sun et al. (2024)	60.44 (0.41)	78.79 (0.30)	80.19 (0.37)	80.42 (0.41)	82.98 (0.30)	69.77 (0.43)	62.29 (0.88)	53.97 (1.24)	59.67 (1.83)
Beta-KD Sun et al. (2026)	54.41 (1.23)	63.76 (0.11)	65.23 (0.31)	64.69 (0.25)	65.10 (0.01)	58.91 (0.14)	52.92 (1.20)	46.98 (1.29)	51.39 (0.75)
LUD Guo et al. (2024)	57.69 (0.71)	79.25 (0.36)	82.09 (0.42)	80.83 (0.11)	83.58 (0.32)	69.16 (0.24)	59.96 (1.74)	52.75 (1.19)	56.22 (2.01)
UBKD Hemmatian et al. (2024)	61.39 (0.45)	78.23 (0.37)	80.46 (0.06)	80.42 (0.30)	82.23 (0.26)	69.10 (0.42)	63.04 (1.01)	54.56 (2.27)	59.73 (2.54)
WTTM Zheng & Yang (2024)	60.55 (0.68)	68.16 (0.19)	69.61 (0.13)	67.83 (0.18)	72.31 (0.23)	63.00 (0.26)	59.93 (0.88)	52.84 (1.49)	57.55 (1.52)
MCD Mishra et al. (2026)	59.93 (1.01)	76.18 (0.40)	78.43 (0.48)	78.13 (0.26)	80.67 (0.37)	69.33 (0.95)	57.94 (2.47)	53.11 (1.45)	58.11 (2.38)
WLS Zhou & Song (2021)	61.98 (0.54)	80.65 (0.25)	81.04 (0.26)	81.49 (0.11)	83.11 (0.10)	71.01 (0.33)	62.94 (2.16)	54.10 (1.66)	59.32 (2.82)
TGeo-KD Hu et al. (2024)	61.21 (0.39)	75.63 (0.21)	77.16 (0.55)	76.52 (0.05)	82.78 (0.15)	68.97 (0.25)	62.46 (0.81)	52.55 (1.32)	58.64 (1.43)
RW-KD Lu et al. (2021)	60.14 (0.41)	72.86 (0.30)	77.00 (0.45)	75.67 (0.46)	80.19 (0.09)	65.21 (0.21)	61.39 (1.23)	54.99 (2.40)	57.97 (2.42)
Ours	**64.43** (0.35)	**83.58** (0.35)	**83.39** (0.16)	**83.26** (0.44)	**84.79** (0.23)	**73.18** (0.27)	**65.95** (0.91)	**57.31** (1.43)	**60.34** (1.33)

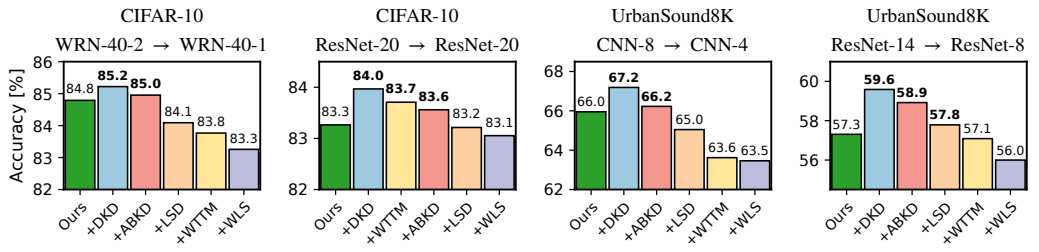

Figure 3: Test accuracy (%) of the student trained by our router combined with existing KD objectives on CIFAR-10 and UrbanSound8K, averaged over 3 and 10 runs, respectively.

Results Table 1 shows that our method achieves the highest accuracy for all nine teacher-student pairs, improving over the strongest published baseline in each column by 0.61–2.93 percentage points and over one-hot training by 2.78–6.29 points. Pure distillation falls below one-hot training by 1.79–15.78 points, since the student inherits the bias of Proposition 1 on the inputs where the teacher is unreliable. The shared-weight methods distill on every sample, so the best they can do is balance the teacher's help against this bias with a single weight, and they gain at most 2.55 points over one-hot training, similarly to the best constant λ with a maximum gain of 2.62 points. The strongest baselines across most columns are adaptive methods such as WLS, LUD, or UBKD, which adjust or withhold the teacher where it appears unreliable, yet they gain at most 3.84 points. Each of their signals captures at most one of the two costs of Proposition 2, and Table 9 shows that replacing our signal with any of them loses up to 7 points in student accuracy. Figure 3 shows that the routing mechanism can be combined with complementary KD objectives, yielding further gains over either approach alone. DKD on its own falls below one-hot training on every CIFAR-10 pair of Table 1, so applying it to every sample can hurt in PET settings, yet fusing it with our method achieves the highest accuracy on every pair of Figure 3. Whether to distill a sample and how to distill it are thus separate choices, and both matter. Appendix G reports results on additional teacher-student pairs, types and amount of teacher expertise, and types of Bayesian teachers, as well as ablations.

5 Conclusion

We studied selective KD from PETs by characterizing when SGD favors one-hot versus teacher supervision on a per-sample basis. Our analysis reveals competing supervision costs governed by BCP diversity and teacher error, motivating an uncertainty-guided routing method that uses AU and EU as practical proxies for these latent factors. Experiments across vision, audio, and high-energy physics classification tasks demonstrate consistent gains over standard and adaptive KD baselines, and show that the proposed routing is complementary to improved distillation objectives.

AI USE STATEMENT

In this work, we used generative AI tools, including Gemini and Claude, solely for manuscript readability polishing and coding refinement, notably for implementing benchmark methods that lacked public repositories. Generative AI did not contribute to determining the datasets, experimental design, methodology, or theoretical derivations. All AI-assisted content and code were reviewed and verified by the authors, who take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International conference on machine learning*, pp. 1613–1622. PMLR, 2015.

Sungmin Cha and Kyunghyun Cho. Why knowledge distillation works in generative models: A minimal working explanation. *Advances in Neural Information Processing Systems*, 38:30017–30037, 2026.

Byungchul Chae and Seonyeong Heo. What to distill? fast knowledge distillation with adaptive sampling. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2407–2416, 2025.

Xin Cheng, Zhiqiang Zhang, Wei Weng, Wenxin Yu, and Jinjia Zhou. De-mkd: Decoupled multi-teacher knowledge distillation based on entropy. *Mathematics*, 12(11):1672, 2024.

Tri Dao, Govinda M Kamath, Vasilis Syrgkanis, and Lester Mackey. Knowledge distillation as semiparametric inference. In *International Conference on Learning Representations*, 2021.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux-effortless Bayesian deep learning. *Advances in Neural Information Processing Systems*, 34:20089–20103, 2021.

Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *International conference on machine learning*, pp. 1184–1193. PMLR, 2018.

Wen-Shu Fan, Su Lu, Xin-Chun Li, De-Chuan Zhan, and Le Gan. Revisit the essence of distilling knowledge through calibration. In *International Conference on Machine Learning*, 2024.

Luyang Fang, Yongkai Chen, Wenxuan Zhong, and Ping Ma. Bayesian knowledge distillation: A Bayesian perspective of distillation with uncertainty quantification. In *International Conference on Machine Learning*, 2024.

Vincent Fortuin. Priors in Bayesian deep learning: A review. *International Statistical Review*, 90 (3):563–591, 2022.

Shreyan Ganguly, Roshan Nayak, Rakshith Rao, Ujan Deb, and Prathosh AP. Adakd: Dynamic knowledge distillation of asr models using adaptive loss weighting. *arXiv preprint arXiv:2405.08019*, 2024.

Guillaume Garrigos and Robert M Gower. Handbook of convergence theorems for (stochastic) gradient methods. *arXiv preprint arXiv:2301.11235*, 2023.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1):1513–1589, 2023.

Aakash Gore, Anoushka Dey, and Aryan Mishra. Uncertainty-aware dual-student knowledge distillation for efficient image classification. *arXiv preprint arXiv:2511.18826*, 2025.

Alex Graves. Practical variational inference for neural networks. *Advances in Neural Information Processing Systems*, 24, 2011.

Zhen Guo, Dong Wang, Qiang He, and Pengzhou Zhang. Leveraging logit uncertainty for better knowledge distillation. *Scientific Reports*, 14(1):31249, 2024.

Corina Gurau, Alex Bewley, and Ingmar Posner. Dropout distillation for efficiently estimating model confidence. *arXiv preprint arXiv:1809.10562*, 2018.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pp. 1225–1234. PMLR, 2016.

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.

Mina Hemmatian, Ali Shahzadi, and Saeed Mozaffari. Uncertainty-based knowledge distillation for Bayesian deep neural network compression. *International Journal of Approximate Reasoning*, 175: 109301, 2024.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

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

Chengming Hu, Haolun Wu, Xuan Li, Chen Ma, Xi Chen, Boyu Wang, Jun Yan, and Xue Liu. Less or more from teacher: Exploiting trilateral geometry for knowledge distillation. In *International Conference on Learning Representations*, pp. 24503–24518, 2024.

Jiayi Huang, Sangwoo Park, Nicola Paoletti, and Osvaldo Simeone. Distilling calibration via conformalized credal inference. In *IEEE International Joint Conference on Neural Networks (IJCNN)*, 2025.

Laurent Valentin Jospin, Hamid Laga, Farid Boussaid, Wray Buntine, and Mohammed Bennamoun. Hands-on bayesian neural networks—a tutorial for deep learning users. *IEEE Computational Intelligence Magazine*, 17(2):29–48, 2022.

Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, 30, 2017.

Jeonghyun Kim, SooKyung Kim, Richeng Xuan, and Hyunsoo Cho. Trust the uncertain teacher: distilling dark knowledge via calibrated uncertainty. *arXiv preprint arXiv:2602.12687*, 2026.

Suyoung Kim, Seonguk Park, Junhoo Lee, and Nojun Kwak. The role of teacher calibration in knowledge distillation. *IEEE Access*, 2025.

Anoop Korattikara Balan, Vivek Rathod, Kevin P Murphy, and Max Welling. Bayesian dark knowledge. *Advances in Neural Information Processing Systems*, 28, 2015.

Weichao Lan, Yiu-ming Cheung, Qing Xu, Buhua Liu, Zhikai Hu, Mengke Li, and Zhenghua Chen. Improve knowledge distillation via label revision and data selection. *IEEE Transactions on Cognitive and Developmental Systems*, 2025.

Zhengda Li. Context-aware knowledge distillation with adaptive weighting for image classification. *arXiv preprint arXiv:2509.05319*, 2025.

Chuanhui Liu, Haoyun Yin, and Xiao Wang. Theoretical perspectives on knowledge distillation: A review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 17(4):e70049, 2025.

Peng Lu, Abbas Ghaddar, Ahmad Rashid, Mehdi Rezagholizadeh, Ali Ghodsi, and Philippe Langlais. Rw-kd: Sample-wise loss terms re-weighting for knowledge distillation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 3145–3152, 2021.

Hoang-Chau Luong, Nghia Van Vo, Kaiqi Zhao, and Lingwei Chen. Consistently informative soft-label temperature for knowledge distillation. *arXiv preprint arXiv:2605.20357*, 2026.

Saptarshi Mandal, Xiaojun Lin, and R Srikant. A theoretical analysis of soft-label vs hard-label training in neural networks. *Proceedings of Machine Learning Research vol*, 283:1–12, 2025.

Aditya K Menon, Ankit Singh Rawat, Sashank Reddi, Seungyeon Kim, and Sanjiv Kumar. A statistical perspective on distillation. In *International Conference on Machine Learning*, pp. 7632–7642. PMLR, 2021.

Thomas Minka. Estimating a dirichlet distribution, 2000.

Ishan Mishra, Deepak Mishra, and Jinjun Xiong. Correctness-aware knowledge distillation for enhanced student learning. *Transactions on Machine Learning Research*, 2026.

Itai Morad, Nir Shlezinger, and Yonina C Eldar. SGD-based knowledge distillation with Bayesian teachers: Theory and guidelines. In *International Conference on Learning Representations (ICLR)*, 2026.

Bálint Mucsányi, Michael Kirchhof, and Seong Joon Oh. Benchmarking uncertainty disentanglement: Specialized uncertainties for specialized tasks. *Advances in Neural Information Processing Systems*, 37:50972–51038, 2024.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

Abhishek Panigrahi, Bingbin Liu, Sadhika Malladi, Sham Kakade, and Surbhi Goel. In good graces: Principled teacher selection for knowledge distillation. In *International Conference on Learning Representations*, volume 2026, pp. 54779–54822, 2026.

Mary Phuong and Christoph Lampert. Towards understanding knowledge distillation. In *International conference on machine learning*, pp. 5142–5151. PMLR, 2019.

Huilin Qu, Congqiao Li, and Sitian Qian. Particle transformer for jet tagging. In *International Conference on Machine Learning*, pp. 18281–18292. PMLR, 2022.

C Radhakrishna Rao. Diversity and dissimilarity coefficients: a unified approach. *Theoretical population biology*, 21(1):24–43, 1982.

Yi Ren, Shangmin Guo, and Danica J Sutherland. Better supervisory signals by observing learning paths. In *International Conference on Learning Representations*, 2022.

Hippolyt Ritter, Aleksandar Botev, and David Barber. A scalable laplace approximation for neural networks. In *International Conference on Learning Representations*, 2018.

Mher Safaryan, Alexandra Peste, and Dan Alistarh. Knowledge distillation performs partial variance reduction. *Advances in Neural Information Processing Systems*, 36, 2024.

Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 1041–1044, 2014.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Gehui Shen, Xi Chen, and Zhihong Deng. Variational learning of Bayesian neural networks via Bayesian dark knowledge. In *International Joint Conferences on Artificial Intelligence*, pp. 2037–2043, 2021.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Jingchen Sun, Shaobo Han, Deep Patel, Wataru Kohno, Can Jin, and Changyou Chen. Uncertainty-aware knowledge distillation for multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Shangquan Sun, Wenqi Ren, Jingzhi Li, Rui Wang, and Xiaochun Cao. Logit standardization in knowledge distillation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15731–15740. IEEE, 2024.

Guanghui Wang, Zhiyong Yang, Zitai Wang, Shi Wang, Qianqian Xu, and Qingming Huang. ABKD: Pursuing a proper allocation of the probability mass in knowledge distillation via α-β-divergence. In *International Conference on Machine Learning*, 2025.

Guanghui Wang, Kaiwen Lv Kacuila, Zhiyong Yang, Zitai Wang, Jin-Wen Wu, Longtao Huang, Qianqian Xu, and Qingming Huang. The bridge-garden dilemma in LLM distillation: Why mixing hard and soft labels works. In *International Conference on Machine Learning*, 2026a.

Kaizheng Wang, Fabio Cuzzolin, David Moens, and Hans Hallez. Credal ensemble distillation for uncertainty quantification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 26319–26327, 2026b.

Andrew G Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. *Advances in Neural Information Processing Systems*, 33:4697–4708, 2020.

Jianqing Xu, Shen Li, Ailin Deng, Miao Xiong, Jiaying Wu, Jiaxiang Wu, Shouhong Ding, and Bryan Hooi. Probabilistic knowledge distillation of face ensembles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3489–3498, 2023.

Taesun Yeom, Taehyeok Ha, and Jaeho Lee. Generalization analysis of linear knowledge distillation. In *High-dimensional Learning Dynamics 2026*, 2026. URL https://openreview.net/forum?id=3vFIVSZjQ4.

Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3903–3911, 2020.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Hao Zhang, Zhibin Zhang, Guangxin Wu, Wanyi Ning, Jiafeng Guo, and Xueqi Cheng. EGAD: Entropy-guided adaptive distillation for token-level knowledge transfer. *arXiv preprint arXiv:2605.01732*, 2026.

Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 6848–6856. IEEE, 2018.

Youcai Zhang, Zhonghao Lan, Yuchen Dai, Fangao Zeng, Yan Bai, Jie Chang, and Yichen Wei. Prime-aware adaptive distillation. In *European Conference on Computer Vision*, pp. 658–674. Springer, 2020.

Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 11953–11962, 2022.

Kaixiang Zheng and En-Hui Yang. Knowledge distillation based on transformed teacher matching. In *International Conference on Learning Representations*, volume 2024, pp. 44963–44979, 2024.

Helong Zhou and Liangchen Song. Rethinking soft labels for knowledge distillation: A bias–variance tradeoff perspective. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2021.

APPENDIX

MAIN SECTIONS

Introduction . . . 1
Related Works . . . 2
Analysis of SGD-Based Learning from Partially Expert Teachers . . . 3
KD Setup . . . 3
Convergence Analysis . . . 5
Interpreting the Routing Costs . . . 6
Bayesian Uncertainty-guided KD . . . 7
Uncertainty Estimates as Proxies . . . 7
Method . . . 7
Numerical Study . . . 8
Conclusion . . . 9

APPENDICES

Basic Linear Algebra, Facts, and Inequalities . . . 15
Detailed Proofs . . . 16
Proof of Proposition 1 . . . 16
Proof of Proposition 2 . . . 17
Proof of Proposition 3 . . . 17
Proof of Theorem 1 . . . 18
Proof of Theorem 2 . . . 22
Synthetic Experiments - Theory . . . 22
Bayesian Deep Learning and Uncertainty Estimation . . . 26
Synthetic Experiments - Proxies . . . 28
Experiment Settings . . . 34
Additional Results and Ablations . . . 37
Limitations and Future Work . . . 43

A BASIC LINEAR ALGEBRA, FACTS, AND INEQUALITIES

We first review some basic facts, standard inequalities, and characteristics of the Dirichlet distribution which are used in the proofs.

• The squared norm of a sum of two terms can be bounded by twice the sum of the individual squared norms:

$$\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2. \tag{11}$$

• Our analysis, like other derivations of linear and exponential rates, relies on a technique where a one-step recurrence is converted into a convergence inequality. Let the sequence $\{z_t\}_{t\geq 0}$ satisfy

$$z_{t+1} \leq (1-\eta)z_t + N, \qquad t = 0, 1, 2, \ldots$$

for constants $\eta \in (0, 1]$ and $N \geq 0$. Unrolling the recursion gives

$$z_t \leq (1-\eta)^t z_0 + (1-\eta)^{t-1}N + (1-\eta)^{t-2}N + \cdots + (1-\eta)^0 N.$$

The geometric sum is then bounded by

$$z_t \leq (1-\eta)^t z_0 + N\sum_{j=0}^{\infty}(1-\eta)^j = (1-\eta)^t z_0 + \frac{N}{\eta}. \tag{12}$$

• If $f(\boldsymbol{\theta})$ is L-smooth (*AS2*) then by Lemma 2.25 in Garrigos & Gower (2023), for all $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^d$ it holds that

$$f(\boldsymbol{\theta}_1) \leq f(\boldsymbol{\theta}_2) + \langle \nabla f(\boldsymbol{\theta}_2), \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \rangle + \frac{L}{2}\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^2. \tag{13}$$

• The law of total expectation states that for random variables (X, Y) and any measurable g,

$$\mathbb{E}[g(X,Y)] \;=\; \mathbb{E}_X\big[\,\mathbb{E}_{Y|X}[\,g(X,Y) \mid X\,]\,\big]\,. \tag{14}$$

• It holds that

$$\frac{\partial \log(\phi(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} = \frac{1}{\phi(\boldsymbol{\theta})} \cdot \frac{\partial \phi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \tag{15}$$

for $\phi(\boldsymbol{\theta}) > 0$.

• For any $a \in \mathbb{R}^m$ and $M \in \mathbb{R}^{m\times m}$,

$$a^\top M a \;=\; \mathrm{Tr}(M\, aa^\top), \tag{16}$$

where $\mathrm{Tr}(\cdot)$ is the trace operator. Moreover, the trace and expectation operators are interchangeable, i.e.,

$$\mathbb{E}[\mathrm{Tr}(M)] = \mathrm{Tr}(\mathbb{E}[M]). \tag{17}$$

• For the softmax function $\phi = \mathrm{softmax}(u)$ with logits $u \in \mathbb{R}^K$, the derivative of the k-th output with respect to the jth logit is

$$\frac{\partial \phi_k}{\partial u_j} = \phi_k\,(\delta_{kj} - \phi_j), \tag{18}$$

where δ_{kj} is the Kronecker delta.

• The student outputs always sum to 1 (softmax). Differentiating $\mathbf{1}^\top \phi_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}) = 1$ with respect to $\bar{\boldsymbol{\theta}}$ gives

$$J_{\bar{\boldsymbol{\theta}}}[\phi_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x})]\,\mathbf{1} = \sum_{k=1}^{K} J_{\bar{\boldsymbol{\theta}},k}[\phi_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x})] = 0, \tag{19}$$

where $J_{\bar{\boldsymbol{\theta}},k}[\phi_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x})]$ is the k-th column of the Jacobian of the student model.

• For the generalization error (1), the optimal inference rule in the sense of minimal cross-entropy risk is the true conditional distribution (Shalev-Shwartz & Ben-David, 2014), i.e.,

$$\phi(\boldsymbol{x}) = [\mathcal{P}(\boldsymbol{y}_1|\boldsymbol{x}), \ldots, \mathcal{P}(\boldsymbol{y}_K|\boldsymbol{x})]. \tag{20}$$

• For any vectors $a, b \in \mathbb{R}^d$, the Cauchy–Schwarz inequality states

$$\langle a, b\rangle \le \|a\| \cdot \|b\|. \tag{21}$$

• For any convex function $g : \mathbb{R}^d \to \mathbb{R}$ and any random vector $z \in \mathbb{R}^d$, Jensen's inequality states $g(\mathbb{E}[z]) \le \mathbb{E}[g(z)]$. We apply it to the convex squared norm

$$\|\mathbb{E}[z]\|^2 \le \mathbb{E}\left[\|z\|^2\right]. \tag{22}$$

• Our proofs use the first two conditional moments of the PET model (4). Since the bias is zero-sum, the total concentration of that Dirichlet is $\varepsilon(\boldsymbol{x})\,\mathbf{1}^\top[\mathcal{P}(\boldsymbol{y}|\boldsymbol{x}) + \mathbf{b}(\boldsymbol{x})] = \varepsilon(\boldsymbol{x})$, and its mean parameter is $\mathcal{P}(\boldsymbol{y}|\boldsymbol{x}) + \mathbf{b}(\boldsymbol{x})$. This results in the basic Dirichlet expectation, variance, and covariance. For every class k and every pair $i \neq j$ we have (Minka, 2000),

$$\mathbb{E}\left[\Phi_k(\boldsymbol{x})|\boldsymbol{x}\right] = \mathcal{P}(\boldsymbol{y}_k|\boldsymbol{x}) + \mathbf{b}_k(\boldsymbol{x}), \tag{23}$$

$$\mathrm{Var}[\Phi_k(\boldsymbol{x})|\boldsymbol{x}] = \frac{\big(\mathcal{P}(\boldsymbol{y}_k|\boldsymbol{x}) + \mathbf{b}_k(\boldsymbol{x})\big)\big(1 - \mathcal{P}(\boldsymbol{y}_k|\boldsymbol{x}) - \mathbf{b}_k(\boldsymbol{x})\big)}{\varepsilon(\boldsymbol{x}) + 1}, \tag{24}$$

$$\mathrm{Cov}[\Phi_i(\boldsymbol{x}), \Phi_j(\boldsymbol{x})|\boldsymbol{x}] = -\frac{\big(\mathcal{P}(\boldsymbol{y}_i|\boldsymbol{x}) + \mathbf{b}_i(\boldsymbol{x})\big)\big(\mathcal{P}(\boldsymbol{y}_j|\boldsymbol{x}) + \mathbf{b}_j(\boldsymbol{x})\big)}{\varepsilon(\boldsymbol{x}) + 1}. \tag{25}$$

The presented characteristics are conditional on $\boldsymbol{x}$, taken over the randomness of the draw in (4). The variance and covariance carry the factor $1/(\varepsilon(\boldsymbol{x}) + 1)$, which is the sense in which $\varepsilon(\boldsymbol{x})$ controls the teacher's noise at $\boldsymbol{x}$.

B Detailed Proofs

B.1 Proof of Proposition 1

Proof. Under the analytical PET model, the distillation risk (3) can be written as

$$\bar{f}_{\mathcal{P},\lambda}(\bar{\boldsymbol{\theta}}) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{P}}\left[\ell(\phi_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}), (1-\lambda(\boldsymbol{x}))\boldsymbol{y} + \lambda(\boldsymbol{x})\Phi(\boldsymbol{x}))\right]. \tag{26}$$

The goal here is to prove what inference rule minimizes the distillation risk (3) which is effectively being minimized when the standard SGD-based KD framework (2) is used to train the student model. Recall that the distillation risk (3) takes the expectation of the CE loss between the student model $\phi_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x})$ and the mixed target $(1-\lambda(\boldsymbol{x}))\boldsymbol{y} + \lambda(\boldsymbol{x})\Phi(\boldsymbol{x})$, over $\boldsymbol{x}, \boldsymbol{y}$, and $\Phi(\boldsymbol{x})$. To ease the analysis, instead of using the mixed target, we define the label variable $\boldsymbol{s} \in \{1, \ldots, K\}$ drawn from it

$$\boldsymbol{s} \mid \big(\boldsymbol{x}, \boldsymbol{y}, \Phi(\boldsymbol{x})\big) \sim \mathrm{Cat}\big((1-\lambda(\boldsymbol{x}))\boldsymbol{y} + \lambda(\boldsymbol{x})\Phi(\boldsymbol{x})\big). \tag{27}$$

The new label variable is well defined since the mixed target is a convex combination (with weight $\lambda(\boldsymbol{x})$) of two probability mass functions: the teacher's prediction $\Phi(\boldsymbol{x})$ and the one-hot encoding of the true label $\boldsymbol{y}$. We write $\tilde{p}(\boldsymbol{s} = k|\boldsymbol{x}) = [\phi_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x})]_k$ because $\phi_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x})$ represents a K-dimensional probability mass function. We compute the conditional law of $\boldsymbol{s}$ given $\boldsymbol{x}$:

$$\begin{aligned} q(\boldsymbol{s} = k|\boldsymbol{x}) &:= \Pr(\boldsymbol{s} = k|\boldsymbol{x}) \\ &\overset{(14)}{=} \mathbb{E}\left[(1-\lambda(\boldsymbol{x}))[\boldsymbol{y}]_k + \lambda(\boldsymbol{x})[\Phi(\boldsymbol{x})]_k \mid \boldsymbol{x}\right] \\ &= (1-\lambda(\boldsymbol{x}))\,\mathbb{E}\left[[\boldsymbol{y}]_k|\boldsymbol{x}\right] + \lambda(\boldsymbol{x})\,\mathbb{E}\left[[\Phi(\boldsymbol{x})]_k|\boldsymbol{x}\right] \\ &\overset{(23)}{=} (1-\lambda(\boldsymbol{x}))\mathcal{P}(\boldsymbol{y}_k|\boldsymbol{x}) + \lambda(\boldsymbol{x})\big(\mathcal{P}(\boldsymbol{y}_k|\boldsymbol{x}) + \mathbf{b}_k(\boldsymbol{x})\big) \\ &= \mathcal{P}(\boldsymbol{y}_k|\boldsymbol{x}) + \lambda(\boldsymbol{x})\mathbf{b}_k(\boldsymbol{x}). \end{aligned} \tag{28}$$

Equipped with these notations, the distillation risk (3) with the CE loss becomes

$$\begin{aligned} \bar{f}_{\mathcal{P},\lambda}(\bar{\boldsymbol{\theta}}) &= \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{P},\ \Phi(\boldsymbol{x})\sim\mathrm{Dir}}\left[\ell\big(\phi_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}),\ (1-\lambda(\boldsymbol{x}))\boldsymbol{y} + \lambda(\boldsymbol{x})\Phi(\boldsymbol{x})\big)\right] \\ &= \mathbb{E}_{(\boldsymbol{x},\boldsymbol{s})}\left[\ell\big(\phi_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}),\ q(\boldsymbol{s}|\boldsymbol{x})\big)\right] \\ &= \mathbb{E}_{(\boldsymbol{x},\boldsymbol{s})}\left[-\log\tilde{p}(\boldsymbol{s}|\boldsymbol{x})\right] \\ &= \mathbb{E}_{(\boldsymbol{x},\boldsymbol{s})}\left[-\log\frac{\tilde{p}(\boldsymbol{s}|\boldsymbol{x})}{q(\boldsymbol{s}|\boldsymbol{x})}\right] + \mathbb{E}_{(\boldsymbol{x},\boldsymbol{s})}\left[-\log q(\boldsymbol{s}|\boldsymbol{x})\right] \\ &= \mathbb{E}_{\boldsymbol{x}}\left[D_{\mathrm{KL}}\big(q(\boldsymbol{s}|\boldsymbol{x})\|\tilde{p}(\boldsymbol{s}|\boldsymbol{x})\big)\right] + H(\boldsymbol{s}|\boldsymbol{x}), \end{aligned}$$

where $D_{\mathrm{KL}}(\cdot\|\cdot)$ is the KL divergence (Shalev-Shwartz & Ben-David, 2014) and $H(\boldsymbol{s}|\boldsymbol{x})$ is the conditional entropy of $\boldsymbol{s}$ given $\boldsymbol{x}$. As D_{KL} is non-negative and equals zero only when $\tilde{p}(\boldsymbol{s}|\boldsymbol{x}) \equiv q(\boldsymbol{s}|\boldsymbol{x})$, i.e., $[\phi_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x})]_k = \mathcal{P}(\boldsymbol{y}_k|\boldsymbol{x}) + \lambda(\boldsymbol{x})\mathbf{b}_k(\boldsymbol{x})$, it holds that the distillation risk (3) is minimized when the inference output is the $\lambda(\boldsymbol{x})$-shifted conditional (28). Under Assumption *AS3*, the minimizer is attainable by some $\bar{\boldsymbol{\theta}}^*_\lambda \in \mathbb{R}^d$ and the proof is concluded. □

B.2 Proof of Proposition 2

Proof. The first term of the bound in Theorem 2 decays geometrically in t and vanishes in the limit, so the asymptotic error is

$$\mathcal{E}(\lambda) := c_1\, \sigma^*_{\bar{f}}(\lambda) + c_2\, \mathbb{E}_{\boldsymbol{x}}\big[\lambda(\boldsymbol{x})\, A(\boldsymbol{x})\big].$$

Substituting $\sigma^*_{\bar{f}}(\lambda) = \mathbb{E}_{\boldsymbol{x}}\big[(1-\lambda(\boldsymbol{x}))\, C(\boldsymbol{x}) + \lambda(\boldsymbol{x})\, B(\boldsymbol{x})\big]$ and plugging in the expressions for the costs yields

$$\begin{aligned}
\mathcal{E}(\lambda) &= c_1\, \mathbb{E}_{\boldsymbol{x}}\big[(1-\lambda(\boldsymbol{x}))\, C(\boldsymbol{x}) + \lambda(\boldsymbol{x})\, B(\boldsymbol{x})\big] + c_2\, \mathbb{E}_{\boldsymbol{x}}\big[\lambda(\boldsymbol{x})\, A(\boldsymbol{x})\big] \\
&= \mathbb{E}_{\boldsymbol{x}}\Big[(1-\lambda(\boldsymbol{x}))\, c_1 C(\boldsymbol{x}) + \lambda(\boldsymbol{x})\big(c_1 B(\boldsymbol{x}) + c_2 A(\boldsymbol{x})\big)\Big] \\
&= \mathbb{E}_{\boldsymbol{x}}\Big[(1-\lambda(\boldsymbol{x}))\, c_{\mathrm{OH}}(\boldsymbol{x}) + \lambda(\boldsymbol{x})\, c_{\mathrm{T}}(\boldsymbol{x})\Big].
\end{aligned}$$

Since $\lambda(\boldsymbol{x}) \in \{0,1\}$, the integrand equals $c_{\mathrm{OH}}(\boldsymbol{x})$ when $\lambda(\boldsymbol{x}) = 0$ and $c_{\mathrm{T}}(\boldsymbol{x})$ when $\lambda(\boldsymbol{x}) = 1$. It is therefore bounded below pointwise by $\min\{c_{\mathrm{OH}}(\boldsymbol{x}), c_{\mathrm{T}}(\boldsymbol{x})\}$, and by monotonicity of the expectation,

$$\mathcal{E}(\lambda) \;\geq\; \mathbb{E}_{\boldsymbol{x}}\big[\min\{c_{\mathrm{OH}}(\boldsymbol{x}), c_{\mathrm{T}}(\boldsymbol{x})\}\big] \qquad \text{for every routing } \lambda(\boldsymbol{x}).$$

The lower bound is attained by minimizing the integrand separately at each $\boldsymbol{x}$, which yields $\lambda^*(\boldsymbol{x}) = \mathbb{1}\{c_{\mathrm{T}}(\boldsymbol{x}) < c_{\mathrm{OH}}(\boldsymbol{x})\}$. The same routing minimizes the bound of Theorem 1, which has the asymptotic error

$$\frac{8\alpha}{\mu}\, \sigma^*_{\bar{f}}(\lambda) + \frac{16\mathcal{L}}{\mu^2}\Big(1 + \frac{4L\mathcal{L}\alpha}{\mu}\Big)\, \mathbb{E}_{\boldsymbol{x}}\big[\lambda(\boldsymbol{x})\, A(\boldsymbol{x})\big] = \frac{2}{L}\, \mathcal{E}(\lambda).$$

□

B.3 Proof of Proposition 3

Proof. Write $z := [h(\boldsymbol{x})^\top,\, 1]^\top$ and $\bar{\boldsymbol{\theta}}_k := [W_{k,:},\, \beta_k]^\top$, so that the k-th logit is $W_{k,:}h(\boldsymbol{x}) + \beta_k = \bar{\boldsymbol{\theta}}_k^\top z$ and $\bar{\boldsymbol{\theta}} = (\bar{\boldsymbol{\theta}}_1, \ldots, \bar{\boldsymbol{\theta}}_K)$. Since the features are frozen, z does not depend on $\bar{\boldsymbol{\theta}}$, and the jth logit depends on $\bar{\boldsymbol{\theta}}$ only through $\bar{\boldsymbol{\theta}}_j$, with $\partial(\bar{\boldsymbol{\theta}}_j^\top z)/\partial\bar{\boldsymbol{\theta}}_j = z$. Hence

$$\frac{\partial[\phi_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x})]_k}{\partial\bar{\boldsymbol{\theta}}_j} \overset{(18)}{=} [\phi_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x})]_k\big(\delta_{kj} - [\phi_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x})]_j\big)\, z,$$

and since the k-th Jacobian column collects these blocks over j, its squared norm is the sum of their squared norms,

$$\|J_{\bar{\boldsymbol{\theta}},k}[\phi_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x})]\|^2 = \|z\|^2\, [\phi_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x})]_k^2\, \|\boldsymbol{y}_k - \phi_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x})\|^2.$$

Substituting into the definition of $C(\boldsymbol{x})$,

$$\begin{aligned}
C(\boldsymbol{x}) &= \sum_{k=1}^{K} \frac{1}{\mathcal{P}(\boldsymbol{y}_k|\boldsymbol{x})}\, \|J_{\bar{\boldsymbol{\theta}},k}[\phi_{\boldsymbol{\theta}^*}(\boldsymbol{x})]\|^2 \overset{(20)}{=} \|z\|^2 \sum_{k=1}^{K} \mathcal{P}(\boldsymbol{y}_k|\boldsymbol{x})\, \|\boldsymbol{y}_k - \mathcal{P}(\boldsymbol{y}|\boldsymbol{x})\|^2 \\
&= \|z\|^2\, \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\big[\|\boldsymbol{y} - \mathbb{E}[\boldsymbol{y}|\boldsymbol{x}]\|^2\big] = \|z\|^2 \sum_{k=1}^{K} \mathrm{Var}[\boldsymbol{y}_k|\boldsymbol{x}] = \|z\|^2\Big(1 - \sum_{k=1}^{K} \mathcal{P}(\boldsymbol{y}_k|\boldsymbol{x})^2\Big),
\end{aligned}$$

where the third equality uses that $\boldsymbol{y}$ equals $\boldsymbol{y}_k$ with probability $\mathcal{P}(\boldsymbol{y}_k|\boldsymbol{x})$ and that $\mathbb{E}[\boldsymbol{y}|\boldsymbol{x}] = \mathcal{P}(\boldsymbol{y}|\boldsymbol{x})$, and the last uses $\mathrm{Var}[\boldsymbol{y}_k|\boldsymbol{x}] = \mathcal{P}(\boldsymbol{y}_k|\boldsymbol{x})(1 - \mathcal{P}(\boldsymbol{y}_k|\boldsymbol{x}))$ since $\boldsymbol{y}_k \in \{0,1\}$. Finally $\|z\|^2 = \|h(\boldsymbol{x})\|^2 + 1$, and multiplying by c_1 concludes the proof. □

B.4 Proof of Theorem 1

Proof. In order to bound the error on the weights, we start by conditioning the expected error on the weights in the last iteration. We denote by $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot|\bar{\boldsymbol{\theta}}^t]$ the conditional expectation w.r.t. $\bar{\boldsymbol{\theta}}^t$, and obtain

$$
\begin{aligned}
&\mathbb{E}_t[\|\bar{\boldsymbol{\theta}}^{t+1} - \bar{\boldsymbol{\theta}}^*\|^2] \overset{(2)}{=} \mathbb{E}_t[\|\bar{\boldsymbol{\theta}}^t - \alpha\nabla\bar{f}_{\xi,\lambda}(\bar{\boldsymbol{\theta}}^t) - \bar{\boldsymbol{\theta}}^*\|^2] \\
&= \|\bar{\boldsymbol{\theta}}^t - \bar{\boldsymbol{\theta}}^*\|^2 - 2\alpha\langle\bar{\boldsymbol{\theta}}^t - \bar{\boldsymbol{\theta}}^*, \nabla\bar{f}_{\mathcal{P},\lambda}(\bar{\boldsymbol{\theta}}^t)\rangle + \alpha^2\mathbb{E}_t[\|\nabla\bar{f}_{\xi,\lambda}(\bar{\boldsymbol{\theta}}^t)\|^2] \\
&\overset{(11)}{\leq} \|\bar{\boldsymbol{\theta}}^t - \bar{\boldsymbol{\theta}}^*\|^2 - 2\alpha\langle\bar{\boldsymbol{\theta}}^t - \bar{\boldsymbol{\theta}}^*, \nabla\bar{f}_{\mathcal{P},\lambda}(\bar{\boldsymbol{\theta}}^t)\rangle + 2\alpha^2\mathbb{E}_t[\|\nabla\bar{f}_{\xi,\lambda}(\bar{\boldsymbol{\theta}}^t) - \nabla\bar{f}_{\xi,\lambda}(\bar{\boldsymbol{\theta}}^*)\|^2] \\
&\qquad + 2\alpha^2\mathbb{E}_t[\|\nabla\bar{f}_{\xi,\lambda}(\bar{\boldsymbol{\theta}}^*)\|^2] \\
&\overset{(AS1)}{\leq} (1-\alpha\mu)\|\bar{\boldsymbol{\theta}}^t - \bar{\boldsymbol{\theta}}^*\|^2 - 2\alpha(\bar{f}_{\mathcal{P},\lambda}(\bar{\boldsymbol{\theta}}^t) - \bar{f}_{\mathcal{P},\lambda}(\bar{\boldsymbol{\theta}}^*)) + 2\alpha^2\mathbb{E}_t[\|\nabla\bar{f}_{\xi,\lambda}(\bar{\boldsymbol{\theta}}^t) - \nabla\bar{f}_{\xi,\lambda}(\bar{\boldsymbol{\theta}}^*)\|^2] \\
&\qquad + 2\alpha^2\mathbb{E}_{\xi\sim\mathcal{P},\,\Phi(\boldsymbol{x})\sim\mathrm{Dir}}\big[\|\nabla\bar{f}_{\xi,\lambda}(\bar{\boldsymbol{\theta}}^*)\|^2\big] \\
&\overset{(AS2)}{\leq} (1-\alpha\mu)\|\bar{\boldsymbol{\theta}}^t - \bar{\boldsymbol{\theta}}^*\|^2 - 2\alpha(\bar{f}_{\mathcal{P},\lambda}(\bar{\boldsymbol{\theta}}^t) - \bar{f}_{\mathcal{P},\lambda}(\bar{\boldsymbol{\theta}}^*)) + 4\alpha^2\mathcal{L}(\bar{f}_{\mathcal{P},\lambda}(\bar{\boldsymbol{\theta}}^t) - \bar{f}_{\mathcal{P},\lambda}(\bar{\boldsymbol{\theta}}^*)) \\
&\qquad + 2\alpha^2\mathbb{E}_{\xi\sim\mathcal{P},\,\Phi(\boldsymbol{x})\sim\mathrm{Dir}}\big[\|\nabla\bar{f}_{\xi,\lambda}(\bar{\boldsymbol{\theta}}^*)\|^2\big] \\
&= (1-\alpha\mu)\|\bar{\boldsymbol{\theta}}^t - \bar{\boldsymbol{\theta}}^*\|^2 + 2\alpha(2\alpha\mathcal{L} - 1)(\bar{f}_{\mathcal{P},\lambda}(\bar{\boldsymbol{\theta}}^t) - \bar{f}_{\mathcal{P},\lambda}(\bar{\boldsymbol{\theta}}^*)) \\
&\qquad + 2\alpha^2\mathbb{E}_{\xi\sim\mathcal{P},\,\Phi(\boldsymbol{x})\sim\mathrm{Dir}}\big[\|\nabla\bar{f}_{\xi,\lambda}(\bar{\boldsymbol{\theta}}^*)\|^2\big] \\
&\leq (1-\alpha\mu)\|\bar{\boldsymbol{\theta}}^t - \bar{\boldsymbol{\theta}}^*\|^2 + 2\alpha^2\mathbb{E}_{\xi\sim\mathcal{P},\,\Phi(\boldsymbol{x})\sim\mathrm{Dir}}\big[\|\nabla\bar{f}_{\xi,\lambda}(\bar{\boldsymbol{\theta}}^*)\|^2\big],
\end{aligned}
$$

where in the last inequality we used the bound on the step-size $\alpha \leq \frac{1}{2\mathcal{L}}$. Next, we apply expectation and unroll the recursion to get

$$
\begin{aligned}
\mathbb{E}[\|\bar{\boldsymbol{\theta}}^t - \bar{\boldsymbol{\theta}}^*\|^2] &\leq (1-\alpha\mu)\mathbb{E}[\|\bar{\boldsymbol{\theta}}^{t-1} - \bar{\boldsymbol{\theta}}^*\|^2] + 2\alpha^2\mathbb{E}_{\xi\sim\mathcal{P},\,\Phi(\boldsymbol{x})\sim\mathrm{Dir}}\big[\|\nabla\bar{f}_{\xi,\lambda}(\bar{\boldsymbol{\theta}}^*)\|^2\big] \\
&\overset{(12)}{\leq} (1-\alpha\mu)^t\|\bar{\boldsymbol{\theta}}^0 - \bar{\boldsymbol{\theta}}^*\|^2 + \frac{2\alpha}{\mu}\,\mathbb{E}_{\xi\sim\mathcal{P},\,\Phi(\boldsymbol{x})\sim\mathrm{Dir}}\big[\|\nabla\bar{f}_{\xi,\lambda}(\bar{\boldsymbol{\theta}}^*)\|^2\big].
\end{aligned}
$$

The recursion bounds the distance of the iterate $\bar{\boldsymbol{\theta}}^t$ from the minimizer $\bar{\boldsymbol{\theta}}^*$ of the distillation risk (3), whose inference rule is given in Proposition 1. To relate the distillation iterates to the true optimum $\boldsymbol{\theta}^*$ that minimizes the true risk (1), we apply (11) and take expectation over both sides:

$$
\mathbb{E}[\|\bar{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^*\|^2] = \mathbb{E}[\|(\bar{\boldsymbol{\theta}}^t - \bar{\boldsymbol{\theta}}^*) + (\bar{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*)\|^2] \leq 2\,\mathbb{E}[\|\bar{\boldsymbol{\theta}}^t - \bar{\boldsymbol{\theta}}^*\|^2] + 2\|\bar{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*\|^2.
$$

After bounding $\mathbb{E}[\|\bar{\boldsymbol{\theta}}^t - \bar{\boldsymbol{\theta}}^*\|^2]$, we now bound $\|\bar{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*\|^2$. We show that for any $\lambda(\boldsymbol{x})$, it holds that:

$$
\|\bar{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*\|^2 \leq \frac{8\mathcal{L}}{\mu^2}\,\mathbb{E}_{\boldsymbol{x}}\left[D_{KL}\big(\mathcal{P}(\boldsymbol{y}|\boldsymbol{x}) \parallel \mathcal{P}(\boldsymbol{y}|\boldsymbol{x}) + \lambda(\boldsymbol{x})\mathbf{b}(\boldsymbol{x})\big)\right].
$$

By plugging $\bar{\boldsymbol{\theta}} = \bar{\boldsymbol{\theta}}^*$ into Assumption *AS1*, applied to the true risk (1), we get

$$
f^*_{\mathcal{P}} \geq f_{\mathcal{P}}(\bar{\boldsymbol{\theta}}^*) + \langle\nabla f_{\mathcal{P}}(\bar{\boldsymbol{\theta}}^*), \boldsymbol{\theta}^* - \bar{\boldsymbol{\theta}}^*\rangle + \frac{\mu}{2}\|\boldsymbol{\theta}^* - \bar{\boldsymbol{\theta}}^*\|^2.
$$

Rearranging both sides leads to

$$
\begin{aligned}
\frac{\mu}{2}\|\boldsymbol{\theta}^* - \bar{\boldsymbol{\theta}}^*\|^2 &\leq \langle\nabla f_{\mathcal{P}}(\bar{\boldsymbol{\theta}}^*), \bar{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*\rangle + \big(f^*_{\mathcal{P}} - f_{\mathcal{P}}(\bar{\boldsymbol{\theta}}^*)\big) \\
&\leq \langle\nabla f_{\mathcal{P}}(\bar{\boldsymbol{\theta}}^*), \bar{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*\rangle \\
&\overset{(21)}{\leq} \|\nabla f_{\mathcal{P}}(\bar{\boldsymbol{\theta}}^*)\| \cdot \|\bar{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*\|,
\end{aligned}
$$

where the second inequality uses that $f^*_{\mathcal{P}}$ is the minimal value of the true risk, so that $f^*_{\mathcal{P}} - f_{\mathcal{P}}(\bar{\boldsymbol{\theta}}^*) \leq 0$. Dividing both sides by $\|\bar{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*\|$ and squaring yields

$$
\frac{\mu^2}{4}\|\boldsymbol{\theta}^* - \bar{\boldsymbol{\theta}}^*\|^2 \leq \|\nabla f_{\mathcal{P}}(\bar{\boldsymbol{\theta}}^*)\|^2.
$$

We continue by expanding and bounding the right-hand side:

$$\begin{aligned}\|\boldsymbol{\theta}^* - \bar{\boldsymbol{\theta}}^*\|^2 &\le \frac{4}{\mu^2}\|\nabla f_{\mathcal{P}}(\bar{\boldsymbol{\theta}}^*)\|^2 \\ &= \frac{4}{\mu^2}\|\mathbb{E}_{\xi\sim\mathcal{P}}\left[\nabla f_\xi(\bar{\boldsymbol{\theta}}^*)\right] - \mathbb{E}_{\xi\sim\mathcal{P}}\left[\nabla f_\xi(\boldsymbol{\theta}^*)\right]\|^2 \\ &\overset{(22)}{\le} \frac{4}{\mu^2}\mathbb{E}_{\xi\sim\mathcal{P}}\left[\|\nabla f_\xi(\bar{\boldsymbol{\theta}}^*) - \nabla f_\xi(\boldsymbol{\theta}^*)\|^2\right] \\ &\overset{(AS2)}{\le} \frac{8\mathcal{L}}{\mu^2}\big(f_{\mathcal{P}}(\bar{\boldsymbol{\theta}}^*) - f^*_{\mathcal{P}}\big),\end{aligned}$$

where the equality uses $\nabla f_{\mathcal{P}}(\boldsymbol{\theta}^*) = 0$. It remains to evaluate the excess true risk at $\bar{\boldsymbol{\theta}}^*$. By Proposition 1, $\phi_{\bar{\boldsymbol{\theta}}^*}(\boldsymbol{x}) = \mathcal{P}(\boldsymbol{y}|\boldsymbol{x}) + \lambda(\boldsymbol{x})\mathbf{b}(\boldsymbol{x})$, while $\phi_{\boldsymbol{\theta}^*}(\boldsymbol{x}) = \mathcal{P}(\boldsymbol{y}|\boldsymbol{x})$ is the same statement for $\lambda(\boldsymbol{x}) \equiv 0$ (also (20)). Hence

$$\begin{aligned}f_{\mathcal{P}}(\bar{\boldsymbol{\theta}}^*) - f^*_{\mathcal{P}} &= \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{P}}\left[-\boldsymbol{y}^\top \log\big(\mathcal{P}(\boldsymbol{y}|\boldsymbol{x}) + \lambda(\boldsymbol{x})\mathbf{b}(\boldsymbol{x})\big)\right] - \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{P}}\left[-\boldsymbol{y}^\top \log \mathcal{P}(\boldsymbol{y}|\boldsymbol{x})\right] \\ &= \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{P}}\left[-\boldsymbol{y}^\top \log \frac{\mathcal{P}(\boldsymbol{y}|\boldsymbol{x}) + \lambda(\boldsymbol{x})\mathbf{b}(\boldsymbol{x})}{\mathcal{P}(\boldsymbol{y}|\boldsymbol{x})}\right] \\ &\overset{(14)}{=} \mathbb{E}_{\boldsymbol{x}}\left[\sum_{k=1}^{K}\mathcal{P}(\boldsymbol{y}_k|\boldsymbol{x})\log\frac{\mathcal{P}(\boldsymbol{y}_k|\boldsymbol{x})}{\mathcal{P}(\boldsymbol{y}_k|\boldsymbol{x}) + \lambda(\boldsymbol{x})\mathbf{b}_k(\boldsymbol{x})}\right] \\ &= \mathbb{E}_{\boldsymbol{x}}\left[D_{KL}\big(\mathcal{P}(\boldsymbol{y}|\boldsymbol{x}) \,\|\, \mathcal{P}(\boldsymbol{y}|\boldsymbol{x}) + \lambda(\boldsymbol{x})\mathbf{b}(\boldsymbol{x})\big)\right] \\ &= \mathbb{E}_{\boldsymbol{x}}\left[\lambda(\boldsymbol{x})A(\boldsymbol{x})\right],\end{aligned}$$

where the last equality uses the hard routing $\lambda(\boldsymbol{x}) \in \{0,1\}$, under which $D_{KL}\big(\mathcal{P}(\boldsymbol{y}|\boldsymbol{x}) \,\|\, \mathcal{P}(\boldsymbol{y}|\boldsymbol{x}) + \lambda(\boldsymbol{x})\mathbf{b}(\boldsymbol{x})\big) = \lambda(\boldsymbol{x})A(\boldsymbol{x})$ with $A(\boldsymbol{x})$ as defined in the main body. We thus obtain

$$\|\bar{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*\|^2 \le \frac{8\mathcal{L}}{\mu^2}\,\mathbb{E}_{\boldsymbol{x}}\left[\lambda(\boldsymbol{x})A(\boldsymbol{x})\right],$$

which, substituted into the triangle inequality together with the recursion bound, gives

$$\mathbb{E}[\|\bar{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^*\|^2] \le 2(1-\alpha\mu)^t\|\bar{\boldsymbol{\theta}}^0 - \bar{\boldsymbol{\theta}}^*\|^2 + \frac{4\alpha}{\mu}\,\mathbb{E}_{\xi\sim\mathcal{P},\,\Phi(\boldsymbol{x})\sim\mathrm{Dir}}\left[\|\nabla\bar{f}_{\xi,\lambda}(\bar{\boldsymbol{\theta}}^*)\|^2\right] + \frac{16\mathcal{L}}{\mu^2}\,\mathbb{E}_{\boldsymbol{x}}\left[\lambda(\boldsymbol{x})A(\boldsymbol{x})\right].$$

In the final step, we bound $\mathbb{E}_{\xi\sim\mathcal{P},\,\Phi(\boldsymbol{x})\sim\mathrm{Dir}}\left[\|\nabla\bar{f}_{\xi,\lambda}(\bar{\boldsymbol{\theta}}^*)\|^2\right]$:

$$\begin{aligned}\mathbb{E}_{\xi\sim\mathcal{P},\,\Phi(\boldsymbol{x})\sim\mathrm{Dir}}\left[\|\nabla\bar{f}_{\xi,\lambda}(\bar{\boldsymbol{\theta}}^*)\|^2\right] &\overset{(11)}{\le} 2\,\mathbb{E}_{\xi,\Phi}\left[\|\nabla\bar{f}_{\xi,\lambda}(\boldsymbol{\theta}^*)\|^2\right] + 2\,\mathbb{E}_{\xi,\Phi}\left[\|\nabla\bar{f}_{\xi,\lambda}(\boldsymbol{\theta}^*) - \nabla\bar{f}_{\xi,\lambda}(\bar{\boldsymbol{\theta}}^*)\|^2\right] \\ &\overset{(AS2)}{\le} 2\,\mathbb{E}_{\xi,\Phi}\left[\|\nabla\bar{f}_{\xi,\lambda}(\boldsymbol{\theta}^*)\|^2\right] + 4\mathcal{L}\big(\bar{f}_{\mathcal{P},\lambda}(\boldsymbol{\theta}^*) - \bar{f}^*_{\mathcal{P},\lambda}\big).\end{aligned}$$

We continue with:

$$\begin{aligned}\bar{f}_{\mathcal{P},\lambda}(\boldsymbol{\theta}^*) - \bar{f}^*_{\mathcal{P},\lambda} &\overset{(13)}{\le} \langle\nabla\bar{f}_{\mathcal{P},\lambda}(\bar{\boldsymbol{\theta}}^*), \boldsymbol{\theta}^* - \bar{\boldsymbol{\theta}}^*\rangle + \frac{L}{2}\|\boldsymbol{\theta}^* - \bar{\boldsymbol{\theta}}^*\|^2 \\ &= \frac{L}{2}\|\boldsymbol{\theta}^* - \bar{\boldsymbol{\theta}}^*\|^2.\end{aligned}$$

Substituting the bound on $\|\bar{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*\|^2$ derived earlier in the proof, we obtain:

$$\bar{f}_{\mathcal{P},\lambda}(\boldsymbol{\theta}^*) - \bar{f}^*_{\mathcal{P},\lambda} \le \frac{4L\mathcal{L}}{\mu^2}\,\mathbb{E}_{\boldsymbol{x}}\left[\lambda(\boldsymbol{x})A(\boldsymbol{x})\right].$$

This bounds the second term above by $\frac{16L\mathcal{L}^2}{\mu^2}\mathbb{E}_{\boldsymbol{x}}[\lambda(\boldsymbol{x})A(\boldsymbol{x})]$.

We now turn to the first term:

$$\begin{aligned}\mathbb{E}_{\xi\sim\mathcal{P},\,\Phi(\boldsymbol{x})\sim\mathrm{Dir}}\left[\|\nabla\bar{f}_{\xi,\lambda}(\boldsymbol{\theta}^*)\|^2\right] &= \mathbb{E}_{\xi\sim\mathcal{P},\,\Phi(\boldsymbol{x})\sim\mathrm{Dir}}\Big[\|\nabla_{\bar{\boldsymbol{\theta}}}\ell\big(\phi_{\boldsymbol{\theta}^*}(\boldsymbol{x}),\,(1-\lambda(\boldsymbol{x}))\boldsymbol{y} + \lambda(\boldsymbol{x})\Phi(\boldsymbol{x})\big)\|^2\Big] \\ &\overset{(14)}{=} \mathbb{E}_{\boldsymbol{x}}\Big[\mathbb{E}_{\boldsymbol{y},\,\Phi(\boldsymbol{x})|\boldsymbol{x}}\left[\|\nabla_{\bar{\boldsymbol{\theta}}}\ell\big(\phi_{\boldsymbol{\theta}^*}(\boldsymbol{x}),\,(1-\lambda(\boldsymbol{x}))\boldsymbol{y} + \lambda(\boldsymbol{x})\Phi(\boldsymbol{x})\big)\|^2\right]\Big].\end{aligned}$$

Since $\lambda(\boldsymbol{x}) \in \{0, 1\}$, the supervisory signal is always either the one-hot label $\boldsymbol{y}$ when $\lambda(\boldsymbol{x}) = 0$ or the teacher prediction $\Phi(\boldsymbol{x})$ when $\lambda(\boldsymbol{x}) = 1$. This means that for each $\boldsymbol{x}$, the inner conditional expectation is either

$$\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\big[\|\nabla_{\bar{\boldsymbol{\theta}}}\ell\big(\phi_{\boldsymbol{\theta}^*}(\boldsymbol{x}),\, \boldsymbol{y}\big)\|^2\big]$$

when $\lambda(\boldsymbol{x}) = 0$, or

$$\mathbb{E}_{\Phi(\boldsymbol{x})|\boldsymbol{x}}\big[\|\nabla_{\bar{\boldsymbol{\theta}}}\ell\big(\phi_{\boldsymbol{\theta}^*}(\boldsymbol{x}),\, \Phi(\boldsymbol{x})\big)\|^2\big]$$

when $\lambda(\boldsymbol{x}) = 1$. We begin with the first expression. Since $\ell\big(\phi_{\boldsymbol{\theta}^*}(\boldsymbol{x}),\, \boldsymbol{y}\big) = -\boldsymbol{y}^\top \log \phi_{\boldsymbol{\theta}^*}(\boldsymbol{x})$, its ith element is

$$\begin{aligned}
\frac{\partial \ell(\phi_{\boldsymbol{\theta}^*}(\boldsymbol{x}), \boldsymbol{y})}{\partial \bar{\boldsymbol{\theta}}_i} &= -\sum_{k=1}^{K} [\boldsymbol{y}]_k \cdot \frac{\partial [\log \phi_{\boldsymbol{\theta}^*}(\boldsymbol{x})]_k}{\partial \bar{\boldsymbol{\theta}}_i} \\
&\overset{(15)}{=} -\sum_{k=1}^{K} [\boldsymbol{y}]_k \cdot \frac{1}{[\phi_{\boldsymbol{\theta}^*}(\boldsymbol{x})]_k} \cdot \frac{\partial [\phi_{\boldsymbol{\theta}^*}(\boldsymbol{x})]_k}{\partial \bar{\boldsymbol{\theta}}_i} \\
&\overset{(20)}{=} -\sum_{k=1}^{K} [\boldsymbol{y}]_k \cdot \frac{1}{\mathcal{P}(\boldsymbol{y}_k|\boldsymbol{x})} \cdot \frac{\partial [\phi_{\boldsymbol{\theta}^*}(\boldsymbol{x})]_k}{\partial \bar{\boldsymbol{\theta}}_i},
\end{aligned}$$

which means that $\nabla_{\bar{\boldsymbol{\theta}}}\ell(\phi_{\boldsymbol{\theta}^*}(\boldsymbol{x}), \boldsymbol{y}) = -J_{\bar{\boldsymbol{\theta}}}[\phi_{\boldsymbol{\theta}^*}(\boldsymbol{x})] \cdot \operatorname{diag}\Big(\frac{1}{\mathcal{P}(\boldsymbol{y}|\boldsymbol{x})}\Big) \cdot \boldsymbol{y}$. We are now able to compute:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\big[\|\nabla_{\bar{\boldsymbol{\theta}}}\ell(\phi_{\boldsymbol{\theta}^*}(\boldsymbol{x}), \boldsymbol{y})\|^2\big] &= \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[\|-J_{\bar{\boldsymbol{\theta}}}[\phi_{\boldsymbol{\theta}^*}(\boldsymbol{x})] \cdot \operatorname{diag}\Big(\frac{1}{\mathcal{P}(\boldsymbol{y}|\boldsymbol{x})}\Big) \cdot \boldsymbol{y}\|^2\right] \\
&= \sum_{k=1}^{K} \mathcal{P}(\boldsymbol{y}_k|\boldsymbol{x})\, \|\frac{1}{\mathcal{P}(\boldsymbol{y}_k|\boldsymbol{x})}\, J_{\bar{\boldsymbol{\theta}},k}[\phi_{\boldsymbol{\theta}^*}(\boldsymbol{x})]\|^2 \\
&= \sum_{k=1}^{K} \frac{1}{\mathcal{P}(\boldsymbol{y}_k|\boldsymbol{x})}\, \|J_{\bar{\boldsymbol{\theta}},k}[\phi_{\boldsymbol{\theta}^*}(\boldsymbol{x})]\|^2 = C(\boldsymbol{x}).
\end{aligned}$$

We now turn to the second expression ($\lambda(\boldsymbol{x}) = 1$). Since $\ell\big(\phi_{\boldsymbol{\theta}^*}(\boldsymbol{x}),\, \Phi(\boldsymbol{x})\big) = -\Phi(\boldsymbol{x})^\top \log \phi_{\boldsymbol{\theta}^*}(\boldsymbol{x})$, its ith element is

$$\begin{aligned}
\frac{\partial \ell(\phi_{\boldsymbol{\theta}^*}(\boldsymbol{x}), \Phi(\boldsymbol{x}))}{\partial \bar{\boldsymbol{\theta}}_i} &= -\sum_{k=1}^{K} [\Phi(\boldsymbol{x})]_k \cdot \frac{\partial [\log \phi_{\boldsymbol{\theta}^*}(\boldsymbol{x})]_k}{\partial \bar{\boldsymbol{\theta}}_i} \\
&\overset{(15)}{=} -\sum_{k=1}^{K} [\Phi(\boldsymbol{x})]_k \cdot \frac{1}{[\phi_{\boldsymbol{\theta}^*}(\boldsymbol{x})]_k} \cdot \frac{\partial [\phi_{\boldsymbol{\theta}^*}(\boldsymbol{x})]_k}{\partial \bar{\boldsymbol{\theta}}_i} \\
&\overset{(20)}{=} -\sum_{k=1}^{K} [\Phi(\boldsymbol{x})]_k \cdot \frac{1}{\mathcal{P}(\boldsymbol{y}_k|\boldsymbol{x})} \cdot \frac{\partial [\phi_{\boldsymbol{\theta}^*}(\boldsymbol{x})]_k}{\partial \bar{\boldsymbol{\theta}}_i},
\end{aligned}$$

which means that $\nabla_{\bar{\boldsymbol{\theta}}}\ell(\phi_{\boldsymbol{\theta}^*}(\boldsymbol{x}), \Phi(\boldsymbol{x})) = -J_{\bar{\boldsymbol{\theta}}}[\phi_{\boldsymbol{\theta}^*}(\boldsymbol{x})] \cdot \operatorname{diag}\Big(\frac{1}{\mathcal{P}(\boldsymbol{y}|\boldsymbol{x})}\Big) \cdot \Phi(\boldsymbol{x})$. The teacher output is not one-hot, so rather than selecting a single column of the Jacobian we expand the squared norm as a quadratic form. Writing $J := J_{\bar{\boldsymbol{\theta}}}[\phi_{\boldsymbol{\theta}^*}(\boldsymbol{x})]$ and $p := \mathcal{P}(\boldsymbol{y}|\boldsymbol{x})$, we have:

$$\begin{aligned}
\mathbb{E}_{\Phi(\boldsymbol{x})|\boldsymbol{x}}\big[\|\nabla_{\bar{\boldsymbol{\theta}}}\ell(\phi_{\boldsymbol{\theta}^*}(\boldsymbol{x}), \Phi(\boldsymbol{x}))\|^2\big] &= \mathbb{E}_{\Phi(\boldsymbol{x})|\boldsymbol{x}}\Big[\Phi(\boldsymbol{x})^\top \operatorname{diag}\Big(\frac{1}{p}\Big) J^\top J \operatorname{diag}\Big(\frac{1}{p}\Big) \Phi(\boldsymbol{x})\Big] \\
&\overset{(16)}{=} \mathbb{E}_{\Phi(\boldsymbol{x})|\boldsymbol{x}}\Big[\operatorname{Tr}\Big(\operatorname{diag}\Big(\frac{1}{p}\Big) J^\top J \operatorname{diag}\Big(\frac{1}{p}\Big)\, \Phi(\boldsymbol{x})\Phi(\boldsymbol{x})^\top\Big)\Big] \\
&\overset{(17)}{=} \operatorname{Tr}\Big(\operatorname{diag}\Big(\frac{1}{p}\Big) J^\top J \operatorname{diag}\Big(\frac{1}{p}\Big)\, \mathbb{E}_{\Phi(\boldsymbol{x})|\boldsymbol{x}}\big[\Phi(\boldsymbol{x})\Phi(\boldsymbol{x})^\top\big]\Big).
\end{aligned}$$

We compute the second moment of the teacher:

$$\begin{aligned}\mathbb{E}_{\Phi(\boldsymbol{x})|\boldsymbol{x}}\big[\Phi(\boldsymbol{x})\Phi(\boldsymbol{x})^\top\big] &\overset{(23)}{=} \mathrm{Cov}\big[\Phi(\boldsymbol{x})|\boldsymbol{x}\big] + \big(p+\mathbf{b}(\boldsymbol{x})\big)\big(p+\mathbf{b}(\boldsymbol{x})\big)^\top \\ &\overset{(24+25)}{=} \frac{1}{\varepsilon(\boldsymbol{x})+1}\Big(\mathrm{diag}\big(p+\mathbf{b}(\boldsymbol{x})\big) - \big(p+\mathbf{b}(\boldsymbol{x})\big)\big(p+\mathbf{b}(\boldsymbol{x})\big)^\top\Big) \\ &\quad + \big(p+\mathbf{b}(\boldsymbol{x})\big)\big(p+\mathbf{b}(\boldsymbol{x})\big)^\top \\ &= \frac{1}{\varepsilon(\boldsymbol{x})+1}\Big(\mathrm{diag}\big(p+\mathbf{b}(\boldsymbol{x})\big) + \varepsilon(\boldsymbol{x})\,\big(p+\mathbf{b}(\boldsymbol{x})\big)\big(p+\mathbf{b}(\boldsymbol{x})\big)^\top\Big).\end{aligned}$$

Substituting this back separates the expectation into two terms:

$$\begin{aligned}&\mathbb{E}_{\Phi(\boldsymbol{x})|\boldsymbol{x}}\big[\|\nabla_{\bar{\boldsymbol{\theta}}}\ell(\phi_{\boldsymbol{\theta}^*}(\boldsymbol{x}),\Phi(\boldsymbol{x}))\|^2\big] \\ &\quad= \frac{1}{\varepsilon(\boldsymbol{x})+1}\mathrm{Tr}\Big(\mathrm{diag}\Big(\frac{1}{p}\Big)J^\top J\,\mathrm{diag}\Big(\frac{1}{p}\Big)\mathrm{diag}\big(p+\mathbf{b}(\boldsymbol{x})\big)\Big) \\ &\qquad+ \frac{\varepsilon(\boldsymbol{x})}{\varepsilon(\boldsymbol{x})+1}\mathrm{Tr}\Big(\mathrm{diag}\Big(\frac{1}{p}\Big)J^\top J\,\mathrm{diag}\Big(\frac{1}{p}\Big)\big(p+\mathbf{b}(\boldsymbol{x})\big)\big(p+\mathbf{b}(\boldsymbol{x})\big)^\top\Big).\end{aligned}$$

In the first term, the product of the three diagonal matrices simplifies to $\mathrm{diag}\Big(\frac{p+\mathbf{b}(\boldsymbol{x})}{p^2}\Big)$. The trace of a matrix multiplied by a diagonal matrix is then the sum of its scaled diagonal elements:

$$\begin{aligned}&\frac{1}{\varepsilon(\boldsymbol{x})+1}\mathrm{Tr}\Big(J^\top J\,\mathrm{diag}\Big(\frac{p+\mathbf{b}(\boldsymbol{x})}{p^2}\Big)\Big) \\ &\quad= \frac{1}{\varepsilon(\boldsymbol{x})+1}\sum_{k=1}^{K}\frac{\mathcal{P}(\boldsymbol{y}_k|\boldsymbol{x})+\mathbf{b}_k(\boldsymbol{x})}{\mathcal{P}(\boldsymbol{y}_k|\boldsymbol{x})^2}\,\|J_{\bar{\boldsymbol{\theta}},k}[\phi_{\boldsymbol{\theta}^*}(\boldsymbol{x})]\|^2 \\ &\quad= \frac{1}{\varepsilon(\boldsymbol{x})+1}\left(\sum_{k=1}^{K}\frac{1}{\mathcal{P}(\boldsymbol{y}_k|\boldsymbol{x})}\,\|J_{\bar{\boldsymbol{\theta}},k}[\phi_{\boldsymbol{\theta}^*}(\boldsymbol{x})]\|^2 + \sum_{k=1}^{K}\frac{\mathbf{b}_k(\boldsymbol{x})}{\mathcal{P}(\boldsymbol{y}_k|\boldsymbol{x})^2}\,\|J_{\bar{\boldsymbol{\theta}},k}[\phi_{\boldsymbol{\theta}^*}(\boldsymbol{x})]\|^2\right) \\ &\quad= \frac{1}{\varepsilon(\boldsymbol{x})+1}\left(C(\boldsymbol{x}) + \sum_{k=1}^{K}\frac{\mathbf{b}_k(\boldsymbol{x})}{\mathcal{P}(\boldsymbol{y}_k|\boldsymbol{x})^2}\,\|J_{\bar{\boldsymbol{\theta}},k}[\phi_{\boldsymbol{\theta}^*}(\boldsymbol{x})]\|^2\right).\end{aligned}$$

We continue with the second term:

$$\begin{aligned}&\frac{\varepsilon(\boldsymbol{x})}{\varepsilon(\boldsymbol{x})+1}\mathrm{Tr}\Big(\mathrm{diag}\Big(\frac{1}{p}\Big)J^\top J\,\mathrm{diag}\Big(\frac{1}{p}\Big)\big(p+\mathbf{b}(\boldsymbol{x})\big)\big(p+\mathbf{b}(\boldsymbol{x})\big)^\top\Big) \\ &\quad\overset{(16)}{=} \frac{\varepsilon(\boldsymbol{x})}{\varepsilon(\boldsymbol{x})+1}\|J\,\mathrm{diag}\Big(\frac{1}{p}\Big)\big(p+\mathbf{b}(\boldsymbol{x})\big)\|^2 \\ &\quad= \frac{\varepsilon(\boldsymbol{x})}{\varepsilon(\boldsymbol{x})+1}\|J\left(\mathbf{1}+\mathrm{diag}\Big(\frac{1}{p}\Big)\mathbf{b}(\boldsymbol{x})\right)\|^2 \\ &\quad= \frac{\varepsilon(\boldsymbol{x})}{\varepsilon(\boldsymbol{x})+1}\|J\mathbf{1}+J\,\mathrm{diag}\Big(\frac{1}{p}\Big)\mathbf{b}(\boldsymbol{x})\|^2 \\ &\quad\overset{(19)}{=} \frac{\varepsilon(\boldsymbol{x})}{\varepsilon(\boldsymbol{x})+1}\|J\,\mathrm{diag}\Big(\frac{1}{p}\Big)\mathbf{b}(\boldsymbol{x})\|^2 \\ &\quad= \frac{\varepsilon(\boldsymbol{x})}{\varepsilon(\boldsymbol{x})+1}\|\sum_{k=1}^{K}\frac{\mathbf{b}_k(\boldsymbol{x})}{\mathcal{P}(\boldsymbol{y}_k|\boldsymbol{x})}J_{\bar{\boldsymbol{\theta}},k}[\phi_{\boldsymbol{\theta}^*}(\boldsymbol{x})]\|^2.\end{aligned}$$

Combining the two terms yields

$$\mathbb{E}_{\Phi(\boldsymbol{x})|\boldsymbol{x}}\big[\|\nabla_{\bar{\boldsymbol{\theta}}}\ell\big(\phi_{\boldsymbol{\theta}^*}(\boldsymbol{x}),\,\Phi(\boldsymbol{x})\big)\|^2\big] = \frac{1}{\varepsilon(\boldsymbol{x})+1}\big(C(\boldsymbol{x})+D(\boldsymbol{x})\big) = B(\boldsymbol{x}).$$

Since $\lambda(\boldsymbol{x})$ selects between the two cases, the inner conditional expectation equals $(1-\lambda(\boldsymbol{x}))\,C(\boldsymbol{x}) + \lambda(\boldsymbol{x})\,B(\boldsymbol{x})$, and therefore

$$\mathbb{E}_{\xi\sim\mathcal{P},\,\Phi(\boldsymbol{x})\sim\mathrm{Dir}}\big[\|\nabla\bar{f}_{\xi,\lambda}(\boldsymbol{\theta}^*)\|^2\big] = \mathbb{E}_{\boldsymbol{x}}\big[(1-\lambda(\boldsymbol{x}))\,C(\boldsymbol{x}) + \lambda(\boldsymbol{x})\,B(\boldsymbol{x})\big] = \sigma^*_{\bar{f}}(\lambda).$$

Substituting this and the bounded loss difference back into the overall iteration bound concludes the proof.

□

B.5 Proof of Theorem 2

Proof. Since the true risk f is L-smooth (*AS2*), applying (13) with $\boldsymbol{\theta}_1 = \bar{\boldsymbol{\theta}}^t$ and $\boldsymbol{\theta}_2 = \boldsymbol{\theta}^*$ gives

$$f(\bar{\boldsymbol{\theta}}^t) \leq f(\boldsymbol{\theta}^*) + \langle \nabla f(\boldsymbol{\theta}^*), \bar{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^* \rangle + \frac{L}{2}\|\bar{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^*\|^2,$$

and since $\nabla f(\boldsymbol{\theta}^*) = 0$ the inequality reduces to:

$$f(\bar{\boldsymbol{\theta}}^t) - f(\boldsymbol{\theta}^*) \leq \frac{L}{2}\|\bar{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^*\|^2.$$

Taking expectation over both sides and applying the bound of Theorem 1 concludes the proof.

□

C Synthetic Experiments - Theory

Our theoretical analysis in Section 3 examines how supervision from one-hot labels or teacher predictions in PET settings affects SGD learning dynamics. The analysis identifies the per-sample cost of each supervision source and the two key quantities governing these costs: BCP diversity and teacher error. Furthermore, it establishes a sample-wise routing rule that minimizes the resulting bounds.

In this section, we demonstrate these implications for the SGD-based learning dynamics of student models through a synthetic experiment where the true BCPs are known. Compared to real-world datasets like MNIST or CIFAR-10, this synthetic setting offers two primary advantages: (i) the exact data-generating process is known, allowing us to compute the true BCP for every input; and (ii) it allows us to explicitly control both task difficulty (which dictates BCP diversity) and teacher error. Teacher error for different samples is controlled by generating teacher predictions via our analytical PET model (4) with varying $\mathbf{b}(\boldsymbol{x})$ and $\varepsilon(\boldsymbol{x})$.

C.1 Experimental Setup

The experiment is a classification task with $K = 8$ classes in a 30-dimensional input space, where the inputs of each class are drawn from a Gaussian. This setting allows us to control task difficulty (i.e., average BCP diversity across the domain) by moving the centers closer to or further from each other. For fixed Gaussian variance, as the class centers come closer together, each sample has a higher probability of coming from other Gaussians, making the BCPs more diverse and the task more difficult. As the centers are moved further apart, it becomes clearer which Gaussian each sample was drawn from, so the task is easier, with lower BCP diversity and true BCPs closer to one-hot. Moreover, this allows us to calculate the true BCP for each sample drawn, enabling us to create varying teacher expertise by controlling the magnitude of the bias and the noise added to the true BCPs.

As explained, each input $\boldsymbol{x}_n$ is a 30×1 vector ($\mathcal{X} = \mathbb{R}^{30}$) drawn from a Gaussian associated with a label $\boldsymbol{y}_n$ that takes its value from one of $K = 8$ classes. Class k's Gaussian is $\mathcal{N}(\nu\boldsymbol{\mu}_k, I)$, where each entry of the center $\boldsymbol{\mu}_k \in \mathbb{R}^{30}$ is drawn uniformly at random from $\{-1, 0, 1\}$. The scalar $\nu > 0$ controls the separation between the classes, i.e., as ν decreases, class centers are closer together, increasing the expected BCP diversity across the domain and making the task more difficult. The data is therefore generated by first drawing a label $\boldsymbol{y}_n$ uniformly from the K classes, and then drawing the input $\boldsymbol{x}_n$ from the Gaussian centered at that class's mean. The true BCP for each sample then follows from Bayes' rule and is given by

$$\mathcal{P}(\boldsymbol{y}_k|\boldsymbol{x}) = \frac{\exp\big(-\frac{1}{2}\|\boldsymbol{x} - \nu\boldsymbol{\mu}_k\|^2\big)}{\sum_{j=1}^{K}\exp\big(-\frac{1}{2}\|\boldsymbol{x} - \nu\boldsymbol{\mu}_j\|^2\big)}. \tag{29}$$

For a given experiment (i.e., fixed class centers via fixed ν), we train many student models. All student models are trained and evaluated in the same manner, differing only in the supervisory signal, which

can be one-hot labels, teacher predictions, or a sample-wise combination, as explained later. In each experiment we use 5×10^3 training, 2×10^3 validation, and 10^4 test samples. Student models are trained for 20,000 iterations of the SGD update (2) with a learning rate of 5×10^{-2}. The architecture is a standard MLP with two hidden layers of 128 units and ReLU activations.

Teacher predictions are computed from the true BCPs and follow the analytical model (4) used in our analysis, governed by a single scalar $s \geq 0$ that sets how expert the teacher is. Specifically, s controls both the bias and the concentration, with the bias moving probability mass s from the most likely class under $\mathcal{P}(\boldsymbol{y}|\boldsymbol{x})$ to the second most likely one, and the concentration being $\varepsilon(\boldsymbol{x}) = 2/s$. Thus, a perfect teacher corresponds to $s = 0$, where it predicts the exact BCPs. As s grows, the teacher becomes more unreliable, eventually moving all the mass of the most likely class to a wrong class in the limit $s \to \infty$. When the mass to be moved exceeds the top probability, the shift is truncated so that the mean stays on the simplex. We therefore refer to s as the teacher error parameter. When a student is trained with the teacher as the supervisory signal, the teacher prediction is a fresh draw from (4) at each step.

Finally, we evaluate each student's SGD learning dynamics and performance through the accuracy and the generalization error (1) it achieves on the test set, as well as the stability of its learning curves. Here, the generalization error is the CE loss with respect to the observed labels. We report each student at the checkpoint with the highest validation accuracy and refer to its test accuracy and test error there as its accuracy and error. To quantify stability, we take the standard deviation of the test accuracy or test error over the last 10^4 iterations. Except for training curves plotted for illustration, all students are trained with 10 seeds, and results report mean values with one standard deviation.

C.2 Teacher and One-hot Costs

We now show the costs of learning from one-hot labels and from teacher predictions in different settings. Note that the asymptotic error in the bounds of Theorems 1 and 2 vanishes when the student is supervised by the true BCPs. In this case, both $A(\boldsymbol{x})$ and $B(\boldsymbol{x})$ are 0, while $C(\boldsymbol{x})$ is only incurred when the one-hot labels are used as the supervision source. Thus, for a given setting, a student trained from the true BCPs serves as a zero-cost reference. Students trained under any other supervision can then be compared both against this reference and against each other.

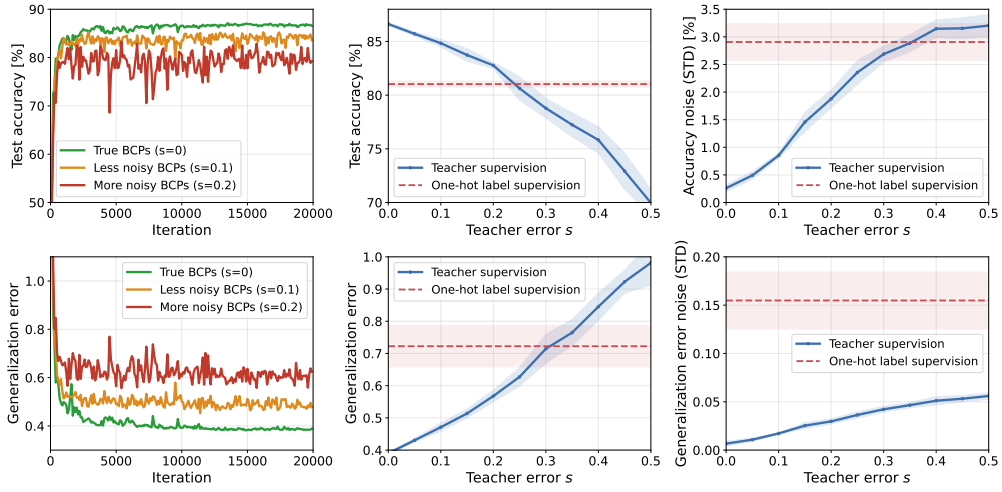

Figure 4: Effect of teacher error s at a fixed setting ($\nu = 0.616$, giving $\mathbb{E}_{\boldsymbol{x}}[\mathrm{Gini}(\mathcal{P}(\boldsymbol{y}|\boldsymbol{x}))] \approx 0.2$). Left: test accuracy (top) and generalization error (bottom) curves for students distilled from teachers at $s = 0, 0.1$ and 0.2. Middle: the distilled student's best-validation checkpoint as a function of s. Right: the stability of each student, the standard deviation of its test curve plateau.

Teacher Supervision Cost For a fixed setting, we train one student from one-hot labels, one student from a perfectly expert teacher (i.e., the true BCPs), and additional students from teachers with different levels of expertise. We sweep s from 0 to 0.5 in steps of 0.05 to control how expert the teacher is. The

results are presented in Figure 4. The top left plot shows the test accuracy curves of students trained from teachers at three levels of expertise ($s = 0$, 0.1 and 0.2), while the bottom left plot shows the corresponding generalization error curves. As the teacher becomes less expert, both curves become less stable and converge to a worse value. The middle and right columns then show how the test accuracy and the generalization error, as well as their stability, behave as a function of teacher error, with the one-hot student shown for reference. As expected, the student's accuracy decreases and its error increases with teacher error. While an expert teacher yields a more accurate student than one-hot labels do, this margin shrinks as teacher expertise decreases, up until a flipping point at $s = 0.24$, beyond which one-hot labels are preferable to the teacher. Beyond this flipping point the student distilled from the noisy teacher is less accurate than the one-hot student, yet over a range of s its learning curves remain more stable, which can be attributed to the teacher's bias shifting the target the student converges to without making its trajectory noisier, while its predictions still vary less between iterations than one-hot samples of the label.

One-hot Supervision Cost We characterize the one-hot cost by first comparing a student trained from one-hot labels to a student trained from a perfect teacher. This allows us to quantify the extent to which an expert teacher can improve upon one-hot learning. Specifically, we consider many settings of the experiment, each with a different value of ν and hence a different task difficulty. We sweep ν from values that yield $\mathbb{E}_{\boldsymbol{x}}[\mathrm{Gini}(\mathcal{P}(\boldsymbol{y}|\boldsymbol{x}))] \approx 0.01$, a very easy and separable task with low BCP diversity, up to $\mathbb{E}_{\boldsymbol{x}}[\mathrm{Gini}(\mathcal{P}(\boldsymbol{y}|\boldsymbol{x}))] \approx 0.5$, a difficult task with high BCP diversity. Results are presented in Figure 5. The top row shows the accuracy curves of the two students at four different difficulties, each subplot harder than the one to its left. In the top left, which is the easiest task both students reach near-perfect accuracy, over 99.7%. As the task becomes more difficult and the average BCP diversity increases, the gap between the two students grows, leaving more room for an expert teacher to enhance student performance. This gap is illustrated in the two bottom left subplots, in which each point is the gap between the two students at that task difficulty, showing that the gap in both accuracy and generalization error increases with BCP diversity. This demonstrates that the relation between BCP diversity and the one-hot cost identified in Proposition 3 persists for multi-layer NNs. If the gap were shown relative to the accuracy of the student trained from the perfect teacher, which itself decreases with task difficulty, it would increase even more. Moreover, the two bottom right plots show the noise ratio between the two students, which is the standard deviation of the one-hot student's test curve divided by that of the perfect-teacher student. As with the performance gap, an expert teacher offers more stability than one-hot supervision on tasks with higher BCP diversity.

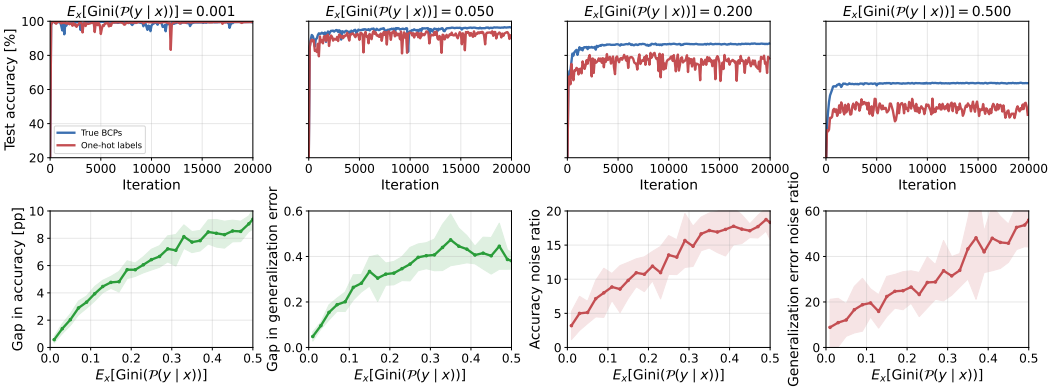

Figure 5: Students trained from perfectly expert teachers (true BCPs) compared to students trained from one-hot labels, at tasks with different expected BCP diversity. Top: test accuracy curves at four task difficulties. Bottom, as a function of $\mathbb{E}_{\boldsymbol{x}}[\mathrm{Gini}(\mathcal{P}(\boldsymbol{y}|\boldsymbol{x}))]$: the best-validation gap in accuracy, the best-validation gap in generalization error, and the two noise ratios, the standard deviation of the one-hot student's test curve divided by that of the true BCP student, for accuracy and for error.

Finally, we examine the effect of BCP diversity and teacher error simultaneously. Each cell in the heatmap (left) in Figure 1 reports the accuracy of a student trained from one-hot labels minus that of a student trained from a teacher. Each column represents a different setting, where task difficulty, and hence expected BCP diversity, grows as we go right. In each such setting we train one student from one-hot labels, and one student per row from a teacher whose expertise decreases as we go

up. The bottom row corresponds to a perfectly expert teacher, the top row to $s = 0.45$, an unreliable teacher. The heatmap shows that the teacher error at which teacher supervision stops being preferable to one-hot supervision itself depends on BCP diversity. Specifically, for tasks with low BCP diversity, one-hot supervision yields student performance comparable to that of even perfect teacher supervision, while for tasks with higher BCP diversity perfect teacher supervision is far more beneficial and the flipping point arrives at considerably more erroneous teachers.

C.3 Sample-wise Routing for PETs

So far, teacher expertise has been uniform across the input domain. We now consider PET settings, where teacher expertise varies across the input space: the teacher is an expert on some samples ($s = 0$) and unreliable on the rest ($s \to \infty$). The task is fixed throughout at $\nu = 0.4$. We compare four students, supervised by (i) one-hot labels alone ($\lambda = 0$), (ii) the teacher alone ($\lambda = 1$), (iii) a constant $\lambda \in (0, 1)$, and (iv) the sample-wise $\lambda(\boldsymbol{x})$ of Proposition 2, which sets $\lambda(\boldsymbol{x}) = 1$ where the teacher is an expert and $\lambda(\boldsymbol{x}) = 0$ where it is unreliable, so that one-hot labels are used there. The constant λ is not fixed in advance but is the best-performing value of a grid search over $[0, 1]$, selected on the validation set separately in each run, so this baseline is given the benefit of tuning.

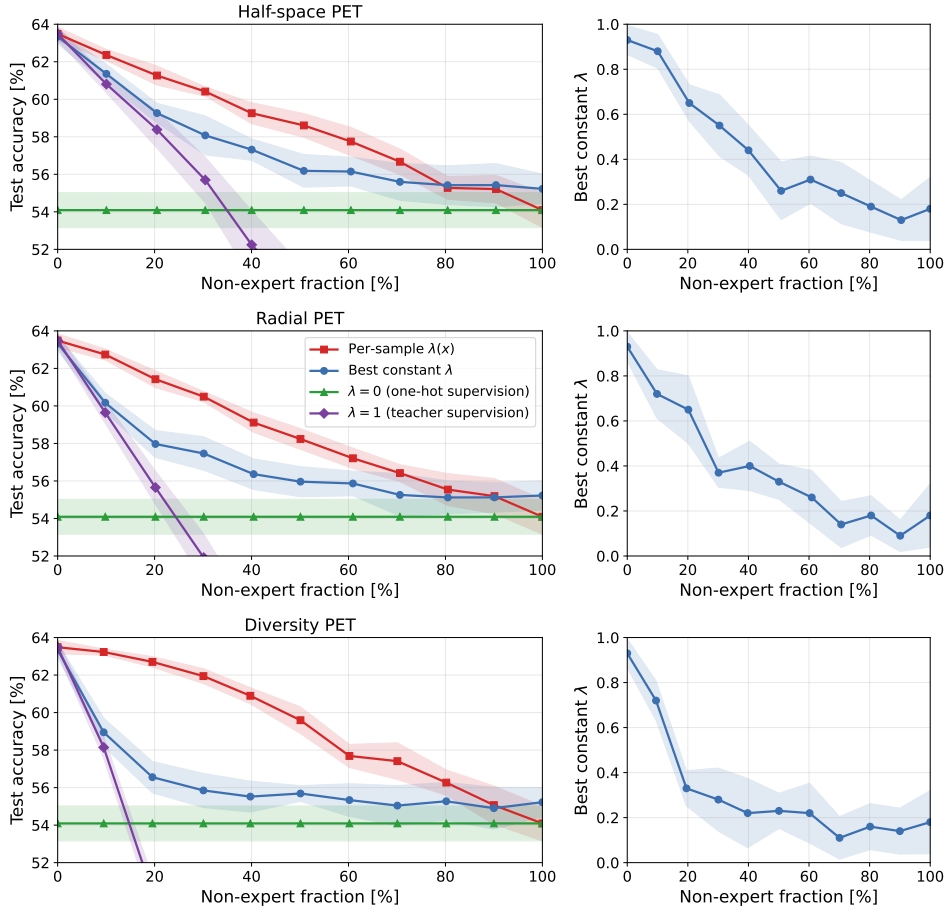

Figure 6: Sweeping the fraction of the domain on which the teacher is unreliable, one row per PET type. Left: test accuracy of the student routed by the per-sample rule, of the best constant λ, and of the two pure baselines $\lambda = 0$ and $\lambda = 1$. Right: the best performing constant λ at each fraction.

We consider three types of PETs, each defined by the region on which the teacher is unreliable. For the *half-space* PET, the region is one side of a hyperplane with a fixed random direction, corresponding to a teacher that was trained on samples from one part of the input space and has never seen the other part, while the student is required to generalize over the whole domain. This makes the expertise tied

to the geometry of the input domain, as is often the case in practice. For the *radial* PET, the teacher is unreliable on a ball around the origin and expert outside it, corresponding to a teacher whose training data covered the outer part of the domain but was sparse in the center of the domain. For the *diversity* PET, the teacher is unreliable on the inputs with lowest BCP diversity, so that its expertise is not tied to the geometry of the input domain. In each case, the region is grown or shrunk so that it covers a fraction f of the input distribution, the non-expert fraction, reported as a percentage in the figures. For each type we sweep f from 0 to 1, each value being a separate experiment, so that the routing is evaluated from a completely expert teacher up to a completely unreliable one, while the regime in between is where the routing is expected to be beneficial.

The results are presented in Figure 6, one row per type of PET. For each row, the left plot shows the accuracy of the four students at every f, and the right plot shows the best-performing constant λ there. All three types behave similarly. The one-hot student is flat in f, since it never uses the teacher. Pure distillation is the best of the four at $f = 0$, where the teacher is fully expert, and degrades steadily as f grows and more of its supervision becomes unreliable. The best-performing constant λ follows this degradation, starting near 1 at $f = 0$ and decreasing to about 0.2 at $f = 1$, and the resulting student remains competitive across the whole range. The routed student coincides with pure distillation at $f = 0$, where $\lambda(\boldsymbol{x}) = 1$ for all inputs, and from there on stays above the best constant λ over almost the entire range, with the largest gains at small to intermediate f, where both supervision sources are costly on part of the domain and only the routed student avoids the more expensive one on each sample. Figure 7 compares the routed student against every constant λ at one value of f per type, taken near the point where the two pure baselines cross. In all three, the routed student achieves substantial gains over every constant λ.

At f approaching 1, where the teacher is unreliable on almost the entire domain, a constant λ of about 0.2 slightly outperforms the routed student. Mixing a small amount of even an unreliable teacher into the one-hot labels acts as label smoothing (Yuan et al., 2020), which regularizes training. Our analysis models the supervision signal, not this regularization effect of soft targets, and the regime in which it appears is one where the teacher is not partially expert at all.

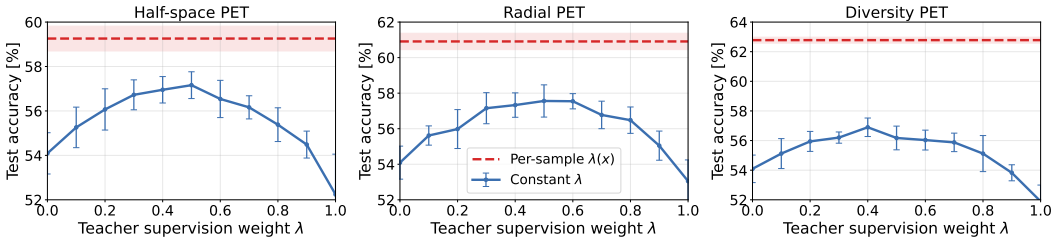

Figure 7: The per-sample rule against every constant λ, for the three types of PETs, each at a single non-expert fraction f (0.40, 0.25 and 0.17, respectively).

D Bayesian Deep Learning and Uncertainty Estimation

We present a brief review of Bayesian deep learning, deep ensembles, and methods for decomposing predictive uncertainty into AU and EU, with a particular focus on the techniques employed in this work. For a broader overview, we refer the reader to Gawlikowski et al. (2023); Jospin et al. (2022).

D.1 Bayesian Deep Learning

BNNs extend traditional (frequentist) NNs by modeling the network parameters $\boldsymbol{\theta}$ as random variables with a distribution over possible values. Unlike traditional NNs, which view learning from data $\mathcal{D}$ as seeking a single deterministic parameter vector, BNNs learn a posterior distribution over the parameter space $p(\boldsymbol{\theta}|\mathcal{D})$. For an input $\boldsymbol{x}$, the marginal predictive distribution over the output $\boldsymbol{y} \in \mathcal{Y}$ is obtained by marginalizing over the posterior:

$$p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D}) = \int p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta}. \tag{30}$$

The main challenge with Bayesian deep learning is that the posterior $p(\boldsymbol{\theta}|\mathcal{D})$ is analytically intractable for deep learning architectures. For this reason, the widespread approximation technique of *Monte*

Carlo (MC) Approximation is used. Following the law of large numbers, this technique approximates the expected prediction in equation (30) with the mean of M stochastic forward passes, each with an i.i.d. realization of $\boldsymbol{\theta}$ drawn from $p(\boldsymbol{\theta}|\mathcal{D})$. In the literature, BNNs are commonly binned into types based on how the posterior distribution is inferred. These methods include VI, sampling approaches, and LAs. We also consider deep ensembles, which are technically non-Bayesian but empirically mimic multimodal Bayesian model averaging (Wilson & Izmailov, 2020).

Variational Inference In this approach, the posterior distribution is approximated with a parameterized family of tractable distributions, e.g., Bernoulli (as in MC Dropout) or Gaussian (Fortuin, 2022; Graves, 2011). The distribution used for the approximation is often called the *variational distribution*. The variational parameters (typically the mean and variance under a mean-field Gaussian assumption) are learned by maximizing the Evidence Lower Bound (ELBO), which minimizes the KL divergence between the variational distribution and the true posterior. To scale to deep networks, modern stochastic VI relies on the reparameterization trick with mini-batch backpropagation (Blundell et al., 2015). During inference, one can utilize the MC approximation by sampling realizations of $\boldsymbol{\theta}$ directly from the optimized variational distribution.

Laplace Approximation The LA offers a computationally efficient method for Bayesian inference, in which the posterior is approximated locally around the maximum a posteriori (MAP) estimate $\hat{\boldsymbol{\theta}}$ with a multivariate Gaussian distribution (Daxberger et al., 2021; Ritter et al., 2018). By computing a second-order Taylor series expansion of the log-posterior at $\hat{\boldsymbol{\theta}}$, the parameter covariance is formed by the inverted negative Hessian of the log-posterior. A key advantage of the LA is that it can be applied directly to standard, pretrained deterministic NNs. As with VI, during inference one can use the MC approximation by sampling realizations of $\boldsymbol{\theta}$ from this local Gaussian posterior.

Deep Ensembles In this approach, M deterministic NNs are independently trained, forming a deep ensemble, each behaving differently. Variety in behavior among members can be achieved through different initializations, data shuffling, bagging, boosting, or varying architectures. Although ensembles operate purely within a frequentist Empirical Risk Minimization (ERM) framework and lack explicit priors or normalized posteriors, their diversity allows the individual models to settle into distinct, isolated local minima across the non-convex loss landscape. This enables ensembles to capture multi-mode predictive disagreement, serving as an effective empirical surrogate for Bayesian model averaging (Wilson & Izmailov, 2020). During inference, the MC approximation in (30) is carried out by replacing stochastic draws from a parameter distribution with the fixed weight vectors of the M trained ensemble members and averaging their individual predictive distributions.

D.2 Uncertainty Estimation

In machine learning, predictive uncertainty is usually separated into AU and EU. AU represents irreducible errors stemming from the data itself, due to measurement noise or true ambiguity. On the other hand, EU stems from a lack of training data, errors in the learning procedure, or errors in the model structure. In classification tasks, standard deterministic NNs output a categorical probability vector for an input $\boldsymbol{x}$, which serves as a measure of predictive confidence. While this pointwise prediction reflects estimated AU, a fundamental shortcoming of standard deterministic NNs is that the predicted confidence conflates AU and EU, making it impossible to separate data noise from model ignorance. To overcome this limitation, Bayesian deep learning and deep ensembles provide a distribution over model predictions, allowing the total predictive uncertainty to be explicitly decomposed into its aleatoric and epistemic components. We now provide additional details on the information-theoretic formulation used in our method, as well as the alternative Gini-based decomposition.

Information-Theoretic Decomposition As introduced in (8), the total predictive entropy of the aggregated teacher prediction decomposes into the expected conditional entropy (AU) and the mutual information (EU) (Depeweg et al., 2018). Beyond the difference formulation in (9), the sample epistemic estimator $\widehat{\mathrm{EU}}(\boldsymbol{x})$ can be equivalently expressed as the average KL divergence from each individual teacher to the aggregated prediction:

$$\widehat{\mathrm{EU}}(\boldsymbol{x}) = \frac{1}{M} \sum_{m=1}^{M} D_{\mathrm{KL}}\big(\Phi_m(\boldsymbol{x}) \,\big\|\, \bar{\Phi}(\boldsymbol{x})\big)\,. \tag{31}$$

This formulation corresponds to the Jensen–Shannon divergence across the M realization draws, explicitly measuring the degree of disagreement or epistemic spread among plausible teacher parameterizations.

Gini-Based Decomposition An alternative sample-based formulation extends the classical law of total variance to categorical distributions using the Gini index, $\mathrm{G}(\mathbf{p}) = 1 - \sum_{k=1}^{K} p_k^2$ (Depeweg et al., 2018). Analogously to (9), total uncertainty (TU) and its components can be estimated across the M teacher predictions as:

$$\widehat{\mathrm{TU}}_{\text{Gini}}(\boldsymbol{x}) = \mathrm{G}\Big(\tfrac{1}{M}\sum_{m=1}^{M}\Phi_m(\boldsymbol{x})\Big), \quad \widehat{\mathrm{AU}}_{\text{Gini}}(\boldsymbol{x}) = \tfrac{1}{M}\sum_{m=1}^{M}\mathrm{G}\left(\Phi_m(\boldsymbol{x})\right),$$
$$\widehat{\mathrm{EU}}_{\text{Gini}}(\boldsymbol{x}) = \widehat{\mathrm{TU}}_{\text{Gini}}(\boldsymbol{x}) - \widehat{\mathrm{AU}}_{\text{Gini}}(\boldsymbol{x}). \tag{32}$$

Under this formulation, AU reflects the average impurity across individual teacher outputs, while EU simplifies exactly to the sum of sample variances for each class prediction, $\frac{1}{M}\sum_{m=1}^{M}\|\Phi_m(\boldsymbol{x}) - \bar{\Phi}(\boldsymbol{x})\|_2^2$. While the entropy-based estimators in (9) capture teacher disagreement in terms of relative information (KL divergence), the Gini decomposition provides an analogous variance-based measure of dispersion in probability space.

E Synthetic Experiments - Proxies

Our analysis in Section 3 identified the two quantities that govern, for each sample, whether distillation is preferable to one-hot supervision: the diversity of the true BCP $\mathcal{P}(\boldsymbol{y}|\boldsymbol{x})$, and the teacher's deviation from it. Both are unobservable in practice. Here we show that the AU of a Bayesian teacher tracks the first, and its EU tracks the second. This is a form of uncertainty quantification evaluation, which is usually carried out through downstream tasks such as separating correct from incorrect predictions or detecting out-of-distribution inputs (Ovadia et al., 2019). Our goal is different, which is the correspondence with the quantities identified by our analysis. This correspondence can only be shown in synthetic settings where $\mathcal{P}(\boldsymbol{y}|\boldsymbol{x})$ is known, unlike real-world settings in which $\mathcal{P}(\boldsymbol{y}|\boldsymbol{x})$ is generally unknown. The experiments are conducted for three settings and for three kinds of Bayesian models: VI, LA, and deep ensembles.

E.1 Setup

Setting Each experiment is a distribution over a bounded two-dimensional domain (for visualization) with $K = 3$ classes, specified by true BCPs $\mathcal{P}(\boldsymbol{y}|\boldsymbol{x})$ that we write down in closed form. We generate three distinct experiments, A, B, and C, which are visualized in the top row of Figure 8 and differ only in $\mathcal{P}(\boldsymbol{y}|\boldsymbol{x})$. Each input $\boldsymbol{x}$ is a location in the domain, and its label is drawn from $\mathcal{P}(\boldsymbol{y}|\boldsymbol{x})$. While inputs in some regions are compatible with a single class, inputs in other regions are compatible with multiple classes, making the task ambiguous. Areas where the task is easy appear as pure colors, while ambiguous regions where the task is difficult appear as mixed colors in the top row of Figure 8. Correspondingly, the bottom row displays the resulting BCP diversity across the input domain, highest where three regions meet. The training density decreases with distance from the origin, i.e., more samples in the center and fewer at the edges, as shown in Figure 9. This makes the teacher partially expert, i.e., well trained near the origin and progressively less well trained toward the edge of the domain. In these experiments, AU should track the true ambiguity, and EU should track the areas where the teacher's estimate of $\mathcal{P}(\boldsymbol{y}|\boldsymbol{x})$ degrades.

Experiment A A random field where the logits are three independent draws from a Gaussian process with an RBF kernel, one per class. They are then scaled by a factor that grows with the distance from the origin. This produces curved class regions, separated by narrow bands in which two classes are close to equally likely, with the strongest ambiguity where three classes meet.

Experiment B Cell centers are placed on a regular grid and then randomly perturbed. Every center is assigned a class uniformly at random. The logits for a class are the sum of a Gaussian kernel over the centers of that class, scaled by a factor that grows with the distance from the origin. This partitions the domain into cells. Neighboring cells of the same class merge into one region, and ambiguity is concentrated on the boundaries between cells of different classes.

Experiment C Inside a fixed radius, three class centers arranged on a ring produce heavily overlapping Gaussian blobs, so ambiguity is largest at the point where all three overlap. Outside that radius, the domain is divided into blocks by both radius and angle, and each block is assigned a class uniformly at random. The blocks are sharp in their interiors and vary smoothly across their boundaries, so ambiguity in the outer zone is confined to the boundaries between blocks. The two zones are blended over a transition band, so $\mathcal{P}(\boldsymbol{y}|\boldsymbol{x})$ varies smoothly across the boundary radius.

Teachers All teachers share the same architecture and the same training procedure. We use a network with two hidden layers of 256 units and ReLU activations, trained with Adam at a learning rate of 2×10^{-3} and a batch size of 256 for 250 epochs, with $N = 3 \times 10^4$ training samples. Next, we provide details related to the Bayesian method used. For VI and LA we use $M = 100$ draws from the posterior to compute AU and EU from Equation (9), and for the ensemble we use $M = 20$ members. AU and EU are evaluated on a uniform grid over the input domain to demonstrate the correspondence. For VI, every layer is Bayesian, with a mean-field Gaussian variational posterior trained by Bayes-by-Backprop. Each mini-batch draws a single weight sample, and the loss adds the KL divergence to a standard normal prior, scaled by $1/N$. The variational standard deviations are initialized at 1.8×10^{-2}. For LA, a deterministic model is first trained with weight decay of 10^{-4}. The LA is then fitted on the last layer with a prior precision of 1 and the exact GGN Hessian. For ensembles, $M = 20$ models are trained independently, differing in initialization and mini-batch drawing order.

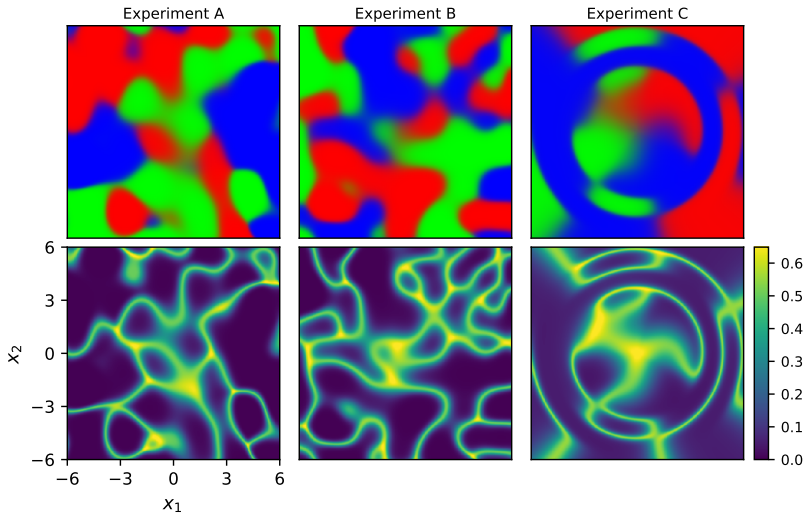

Figure 8: Each column is a different experiment. Top: the true BCPs $\mathcal{P}(\boldsymbol{y}|\boldsymbol{x})$, with red, green, and blue for the three classes, so a pure color means the input belongs to one class and a mixed color means it could belong to several. Bottom: the true BCP diversity $\mathrm{Gini}\big(\mathcal{P}(\boldsymbol{y}|\boldsymbol{x})\big)$, which is concentrated on the boundaries between classes and highest where three classes meet, where the task is hardest.

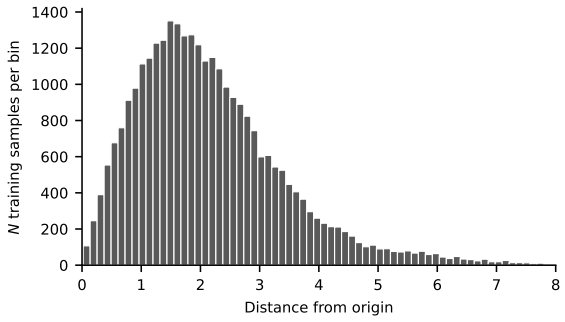

Figure 9: The training density, decreasing away from the origin, so the teacher's knowledge degrades toward the edges.

E.2 RESULTS

Figure 10 shows four quantities across the input domain for Experiment A, one column each and one row per teacher: the true diversity $\mathrm{Gini}\big(\mathcal{P}(\boldsymbol{y}|\boldsymbol{x})\big)$, the teacher's AU, the true teacher error $D_{KL}\big(\mathcal{P}(\boldsymbol{y}|\boldsymbol{x}) \,\|\, \Phi(\boldsymbol{x})\big)$, and the teacher's EU. Figures 11 and 12 show the same for Experiments B and C. Color scales are clipped at the 99th percentile so that a handful of extreme values do not wash out the rest of the map. BCP diversity is a property of the experiment alone, while the error depends on the teacher and therefore varies down the rows. Tables 2 and 3 report the correspondence numerically, AU against the diversity and EU against the error. As expected, in every experiment and for every teacher the error grows with the distance from the origin, where the training density is low, so each teacher is partially expert. Visually, the proxies recover the spatial structure of both BCP diversity and teacher error, appearing as smoothed versions of the ground truth. The smoothing is expected from the network's inductive bias and from averaging over the posterior, and where it widens a sharp boundary it lowers the pointwise correlations in Tables 2–3.

AU tracks BCP diversity where the teacher is expert, with the tracking degrading as the teacher becomes unreliable. This follows from the estimator: AU is the expected entropy of the individual predictions, which approach the BCPs where the teacher is expert, since the cross-entropy-optimal predictor is the true conditional (20). The two quantities differ in scale, since the one-hot cost of Proposition 3 uses the Gini index while (9) uses entropy, whose maximum values are $(K-1)/K$ and $\ln K$. Once rescaled, the two are nearly the same over most of the simplex. Deep ensembles achieve the highest correlation, followed closely by LA and VI.

EU tracks teacher error, staying low where the teacher is expert and growing where it is not. Unlike AU, however, the magnitude and shape of EU depend on the posterior approximation. Teacher error is the discrepancy between the mean prediction and the true BCPs, whereas EU is the spread of the posterior. The correspondence holds because the posterior spread grows where training data is sparse, which is where the teacher has not learned the underlying distribution and its prediction is far from the true BCPs.

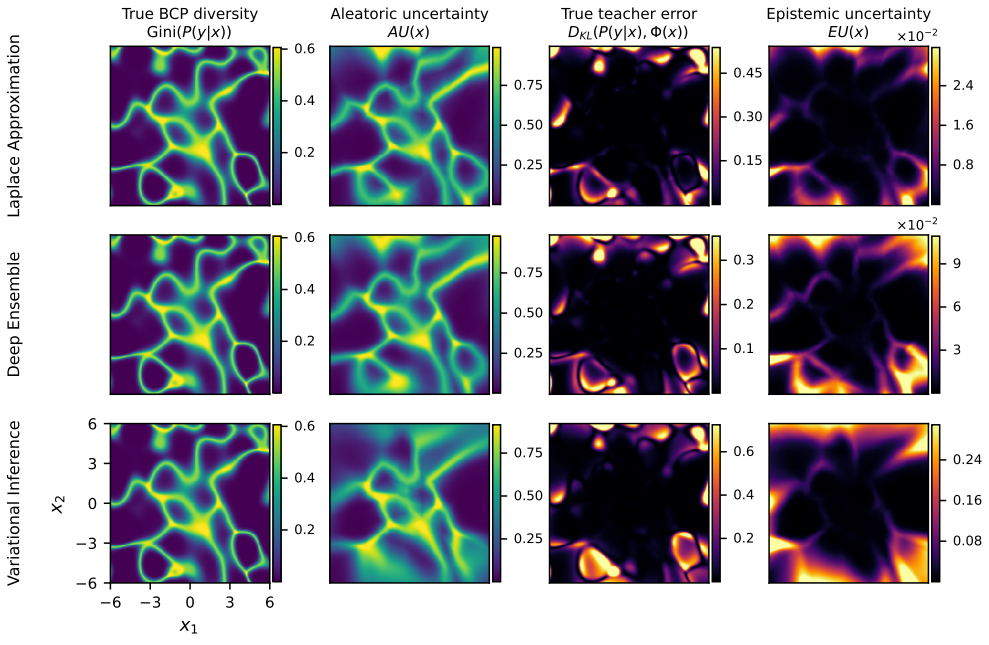

Figure 10: Experiment A. Rows are the three teachers (LA, deep ensemble, VI). Columns are, from left to right, the unobservable BCP diversity $\mathrm{Gini}\big(\mathcal{P}(\boldsymbol{y}|\boldsymbol{x})\big)$, the teacher's AU, the unobservable teacher error $D_{KL}\big(\mathcal{P}(\boldsymbol{y}|\boldsymbol{x}) \,\|\, \Phi(\boldsymbol{x})\big)$, and the teacher's EU. Uncertainty estimates are computed via (9).

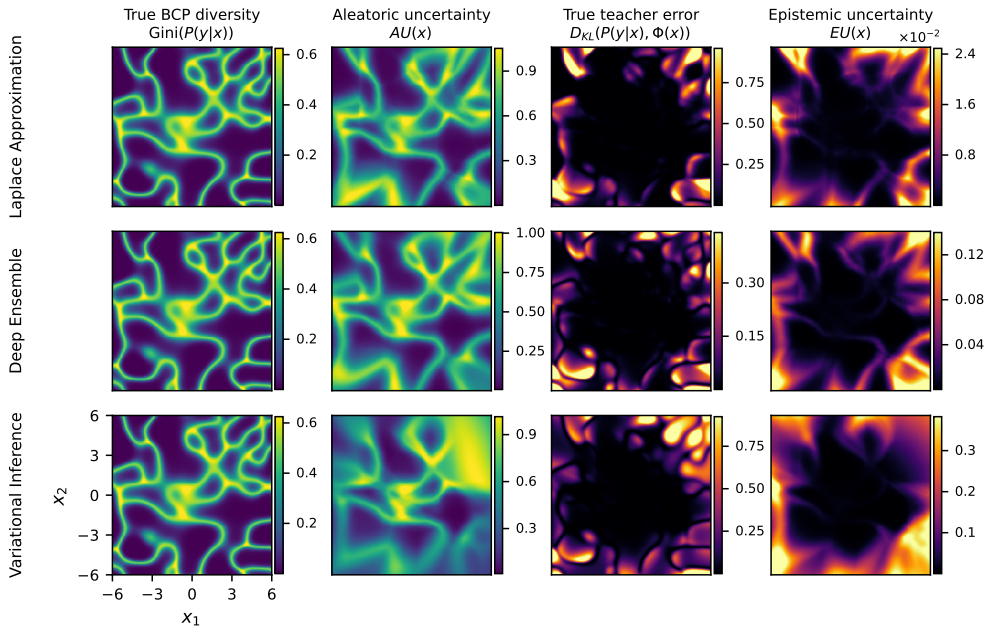

Figure 11: Experiment B. Rows are the three teachers (LA, deep ensemble, VI). Columns are, from left to right, the unobservable BCP diversity $\mathrm{Gini}\big(\mathcal{P}(\boldsymbol{y}|\boldsymbol{x})\big)$, the teacher's AU, the unobservable teacher error $D_{KL}\big(\mathcal{P}(\boldsymbol{y}|\boldsymbol{x}) \,\|\, \Phi(\boldsymbol{x})\big)$, and the teacher's EU. Uncertainty estimates are computed via (9).

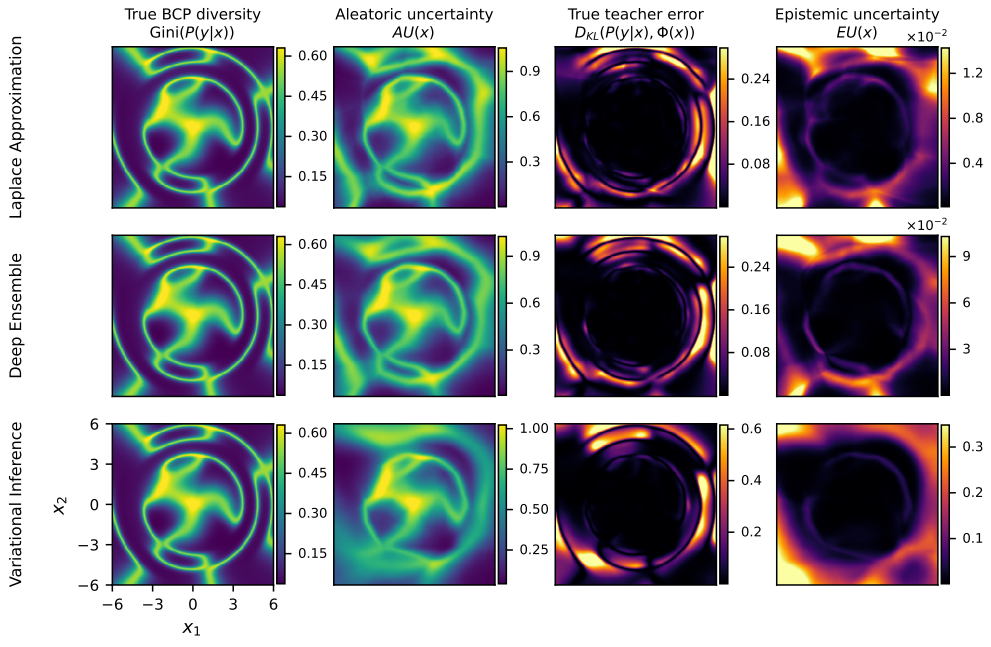

Figure 12: Experiment C. Rows are the three teachers (LA, deep ensemble, VI). Columns are, from left to right, the unobservable BCP diversity $\mathrm{Gini}\big(\mathcal{P}(\boldsymbol{y}|\boldsymbol{x})\big)$, the teacher's AU, the unobservable teacher error $D_{KL}\big(\mathcal{P}(\boldsymbol{y}|\boldsymbol{x}) \,\|\, \Phi(\boldsymbol{x})\big)$, and the teacher's EU. Uncertainty estimates are computed via (9).

Table 2: Correspondence between the teacher's AU and the unobservable BCP diversity $\mathrm{Gini}\big(\mathcal{P}(\boldsymbol{y}|\boldsymbol{x})\big)$, for experiments A,B, and C, and for VI, LA, and ens teachers, computed over the whole input region.

		Pearson correlation	Spearman correlation	Cosine similarity	Monotone fit R^2
Exp A	VI	0.702	0.706	0.842	0.509
	LA	0.797	0.800	0.889	0.648
	Ens	0.832	0.834	0.903	0.715
Exp B	VI	0.657	0.687	0.853	0.466
	LA	0.640	0.669	0.847	0.419
	Ens	0.798	0.815	0.904	0.654
Exp C	VI	0.699	0.706	0.893	0.525
	LA	0.769	0.770	0.916	0.624
	Ens	0.784	0.796	0.920	0.654

Table 3: Correspondence between the teacher's EU and the unobservable teacher error $D_{KL}\big(\mathcal{P}(\boldsymbol{y}|\boldsymbol{x}) \,\|\, \Phi(\boldsymbol{x})\big)$, for experiments A,B, and C, and for VI, LA, and ens teachers, computed over the whole input region.

		Pearson correlation	Spearman correlation	Cosine similarity	Monotone fit R^2
Exp A	VI	0.499	0.777	0.623	0.269
	LA	0.514	0.803	0.598	0.292
	Ens	0.530	0.783	0.627	0.296
Exp B	VI	0.518	0.777	0.698	0.316
	LA	0.473	0.824	0.607	0.281
	Ens	0.635	0.821	0.731	0.415
Exp C	VI	0.369	0.690	0.581	0.254
	LA	0.245	0.698	0.408	0.099
	Ens	0.340	0.730	0.502	0.225

E.3 Alternative Measures

The Gini index (as a measure of unobservable BCP diversity) and the forward KL divergence (as a measure of unobservable teacher error) are the specific forms that follow from the analysis and modeling choices in Section 3. Works that analyze KD through generalization bounds (globally) rather than through learning dynamics arrive at different measures, with the teacher error appearing as the squared L_2 distance $\|\mathcal{P}(\boldsymbol{y}|\boldsymbol{x}) - \Phi(\boldsymbol{x})\|_2^2$ rather than as $D_{KL}\big(\mathcal{P}(\boldsymbol{y}|\boldsymbol{x})\|\Phi(\boldsymbol{x})\big)$ (Menon et al., 2021; Dao et al., 2021). BCP diversity can also be measured differently, for example by its entropy. On the estimator side, AU and EU admit more than one decomposition, and the Gini-based alternative to (8) is given in Appendix D.2.

We therefore repeat the comparison across these alternatives. For the unobservable quantities we consider the Gini index and the entropy as measures of BCP diversity, and the forward KL divergence, the reverse KL divergence and the squared distance as measures of teacher error. For the proxies we consider the information-theoretic and the Gini-based decompositions of AU and EU. Results are shown in Figures 13 and 14 and Tables 4 and 5 for Experiment A with the LA teacher.

For BCP diversity, all combinations in Table 4 give similar correlations, since the Gini index and entropy behave similarly across the simplex and differ mainly in scale. For teacher error, by contrast, the correlation varies across combinations, as shown in Table 5. In terms of the error measure, all EU proxies track the squared L_2 distance best, which matches the L_2 teacher error that generalization-based analyses of KD link to performance (Menon et al., 2021; Dao et al., 2021). In terms of proxies, the mutual information of (9) correlates best with the forward KL of our analysis. We therefore adopt the information-theoretic decomposition in equation (9) throughout our work.

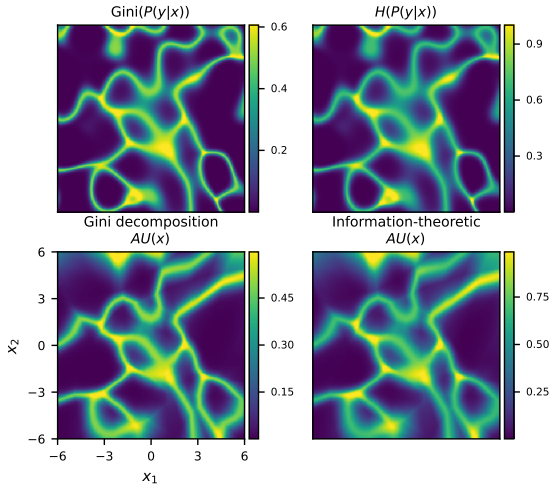

Figure 13: Alternative measures of BCP diversity, for Experiment A with the LA teacher. (Top) the two unobservable measures, the Gini index and the entropy of $\mathcal{P}(\boldsymbol{y}|\boldsymbol{x})$. (Bottom) the two AU proxies, from the Gini and the information-theoretic decompositions.

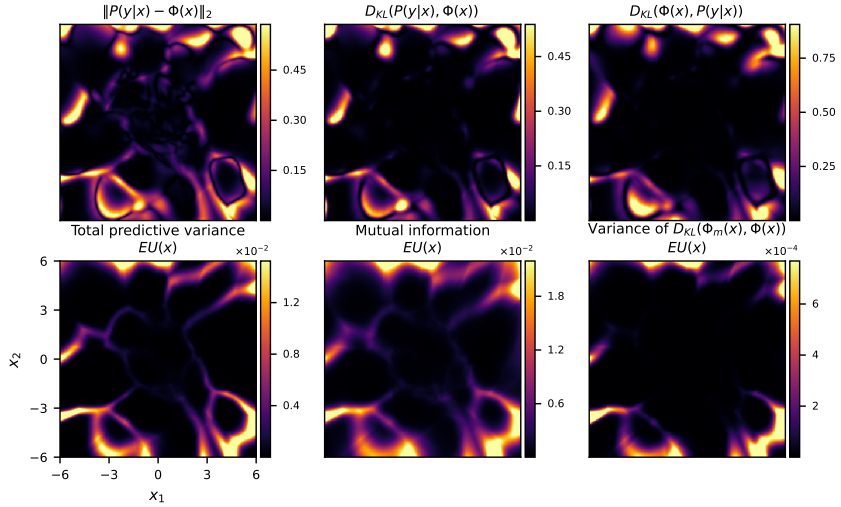

Figure 14: Alternative measures of teacher error, for Experiment A with the LA teacher. (Top) the three unobservable measures, the squared distance, the forward KL divergence and the reverse KL divergence between $\mathcal{P}(\boldsymbol{y}|\boldsymbol{x})$ and $\Phi(\boldsymbol{x})$. (Bottom) the three EU proxies, the total predictive variance, the mutual information, and the variance across posterior draws of $D_{KL}\big(\Phi_m(\boldsymbol{x}) \,\|\, \Phi(\boldsymbol{x})\big)$.

Table 4: Correspondence between each AU proxy and each measure of BCP diversity, for Experiment A with the LA teacher, computed over the whole input region.

AU proxy	Target	Pearson correlation	Spearman correlation	Cosine similarity	Monotone fit R^2
Information-theoretic decomp (9)	$\mathrm{Gini}\big(\mathcal{P}(\boldsymbol{y}\|\boldsymbol{x})\big)$	0.797	0.800	0.889	0.648
	$H\big(\mathcal{P}(\boldsymbol{y}\|\boldsymbol{x})\big)$	0.812	0.803	0.907	0.669
Gini-based decomp (32)	$\mathrm{Gini}\big(\mathcal{P}(\boldsymbol{y}\|\boldsymbol{x})\big)$	0.793	0.793	0.888	0.637
	$H\big(\mathcal{P}(\boldsymbol{y}\|\boldsymbol{x})\big)$	0.799	0.794	0.899	0.650

Table 5: Correspondence between each EU proxy and each measure of teacher error, for Experiment A with the LA teacher, computed over the whole input region.

EU proxy	Target	Pearson correlation	Spearman correlation	Cosine similarity	Monotone fit R^2
Information-theoretic decomp (9)	$D_{KL}\big(\mathcal{P}(\boldsymbol{y}\|\boldsymbol{x}) \,\|\, \Phi(\boldsymbol{x})\big)$	0.514	0.803	0.598	0.292
	$D_{KL}\big(\Phi(\boldsymbol{x}) \,\|\, \mathcal{P}(\boldsymbol{y}\|\boldsymbol{x})\big)$	0.693	0.786	0.759	0.512
	$\|\mathcal{P}(\boldsymbol{y}\|\boldsymbol{x}) - \Phi(\boldsymbol{x})\|_2$	0.700	0.839	0.788	0.573
Gini-based decomp (32)	$D_{KL}\big(\mathcal{P}(\boldsymbol{y}\|\boldsymbol{x}) \,\|\, \Phi(\boldsymbol{x})\big)$	0.483	0.718	0.563	0.280
	$D_{KL}\big(\Phi(\boldsymbol{x}) \,\|\, \mathcal{P}(\boldsymbol{y}\|\boldsymbol{x})\big)$	0.666	0.656	0.728	0.481
	$\|\mathcal{P}(\boldsymbol{y}\|\boldsymbol{x}) - \Phi(\boldsymbol{x})\|_2$	0.707	0.877	0.770	0.619
$Var(D_{KL})$ (31)	$D_{KL}\big(\mathcal{P}(\boldsymbol{y}\|\boldsymbol{x}) \,\|\, \Phi(\boldsymbol{x})\big)$	0.413	0.797	0.479	0.277
	$D_{KL}\big(\Phi(\boldsymbol{x}) \,\|\, \mathcal{P}(\boldsymbol{y}\|\boldsymbol{x})\big)$	0.592	0.794	0.642	0.491
	$\|\mathcal{P}(\boldsymbol{y}\|\boldsymbol{x}) - \Phi(\boldsymbol{x})\|_2$	0.548	0.808	0.607	0.551

E.4 Effect of the Number of MC Samples

The estimators in (9) are computed from M samples from the posterior (or M ensemble members), so both the uncertainty maps and their correspondence with the unobservable quantities depend on M. We now test how M affects this correspondence. Figure 15 reports the mean and standard deviation over repeated draws of the resulting Spearman correlation (AU on the left, EU on the right) as a function of M. For the ensemble, we use a maximum of $M = 20$ members. The results show that AU is largely unaffected by M, as averaging individual entropies stabilizes with very few draws. In contrast, EU is more sensitive to M, improving as M grows. Since EU measures the spread across draws, it needs more samples than an average does.

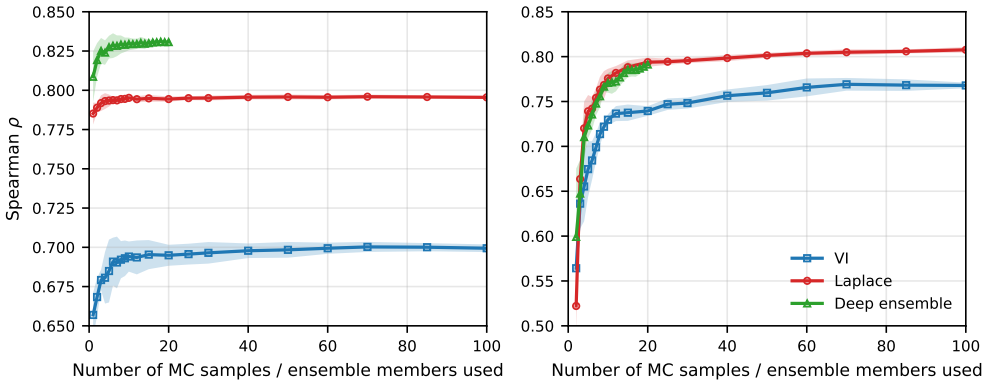

Figure 15: Spearman correlation as a function of the number of MC-samples used (or number of ensemble members) in experiment A. (Left) AU against the true BCP diversity. (Right) EU against the true teacher error. Shaded bands are one standard deviation over 5 repeated draws.

F Experiment Settings

In the following experiments, the student needs to generalize to deployment conditions composed of two distinct regions. We call the condition on which the teacher is expert region A, and the other condition region B. The teacher sees data from region A only during training, so it is a PET. The student's training and test data are composed of a fraction $1 - f$ from region A and a fraction f from region B, which we call the non-expert fraction. Unless stated otherwise, experiments are performed at $f = 0.3$. The teacher's and the student's data do not overlap. For a given setting, all students are trained from the same teacher model using SGD. Apart from Appendix G.4, which reports results with deep ensembles and VI, all teachers are trained as deterministic NNs and made Bayesian post hoc with the LA, using 30 posterior samples during student training. The implementation for LA provided in `https://github.com/aleximmer/Laplace` allows loading the pretrained teacher parameters and applying the approximation.

F.1 BASELINES

All methods are trained under identical settings and differ only in the KD method itself. For a fair comparison, every baseline with tunable hyperparameters is evaluated over a grid that includes its published setting, and we report its best configuration in each experiment. Likewise, the methods are run with temperature $T = 1$ and $T = 4$, as is standard in distillation. As the original papers evaluate different architectures, datasets, and, in particular, teachers that are reliable on the student's data, the accuracies reported here reflect the PET setting rather than how these methods perform in their own settings.

- **Ours.** The selective routing of Algorithm 1, with the two-component mixture fitted using scikit-learn's EM implementation. As is common when combining losses, the two terms are rescaled per batch so that their gradients have comparable magnitude.
- **One-hot training.** The student is trained on the one-hot labels alone ($\lambda = 0$).
- **KD** Hinton et al. (2015). The student matches teacher predictions ($\lambda = 1$).
- **Best constant λ.** A grid search over $\lambda \in \{0, 0.2, 0.4, 0.5, 0.6, 0.8, 1\}$ with the original KD implementation (Hinton et al., 2015). The best-performing constant λ is reported.
- **DKD** Zhao et al. (2022). The KD term is split into a binary target-versus-rest part and a part over the non-target classes. In vanilla KD the latter carries an implicit per-sample weight, one minus the teacher's target-class probability, which DKD replaces by a constant β, weighting the two parts separately. We follow their released implementation with $\alpha = 1$ and select β from $\{0.5, 1, 2, 4, 8\}$.
- **ABKD** Wang et al. (2025). The KL divergence is replaced by a two-parameter α-β divergence with forward and reverse KL at its corners, which controls how strongly mismatches with the teacher are penalized and how much the student's own confidence weights them. We follow their released code and select the pair from $\{(1, 0), (0, 1), (0.9, 0.2), (0.8, 0.3), (0.7, 0.4)\}$ and the distillation weight from $\{2, 8, 32\}$.
- **LSD** Sun et al. (2024). Both models' logits are standardized before the softmax, so the student reproduces the teacher's ranking without having to reproduce its logit magnitude. We follow their released code and select the distillation and CE weights from $\{0.9, 3, 9\}$ and $\{0.1, 1\}$.
- **Beta-KD** Sun et al. (2026). The teacher's supervision is treated as a Gibbs prior over the student's activations and its weight becomes a learnable precision, so the weight adapts during training. We implement the task-level variant of their objective.
- **LUD** Guo et al. (2024). We use its category uncertainty weighting, in which the teacher's maximum predicted probability gates distillation per sample: samples above the median confidence of the training set are distilled and the rest are trained on the label.
- **UBKD** Hemmatian et al. (2024). Both T and λ are made per-sample from a Bayesian teacher, with the mixing weight decreasing in EU. We run the configuration of their own ablation with the adaptive mixing weight, and we deterministically rescale their slope to the uncertainty scale of each experiment, over four possible rescaling points.
- **WTTM** Zheng & Yang (2024). The teacher's prediction is tempered while the student's is not, and each sample's distillation term is scaled by a monotone function of the teacher's Rényi entropy, so the student is pulled harder toward the teacher where its prediction is smoother. We implement their weighted form and select the anchor from $\{0.5, 0.7, 0.9\}$, setting β from it by their Eq. 19.
- **MCD** Mishra et al. (2026). Each batch is split on whether the teacher's most likely class matches the observed label, amplifying the distillation term where they agree and the CE term where they do not. We implement their objective (their Eq. 4) at their published $\beta = 2$, and select α from $\{8, 12\}$.
- **WLS** Zhou & Song (2021). The soft-label term of each sample is weighted by a monotone function of the ratio between the student's and the teacher's CE, discounting the teacher wherever the student predicts that sample better. We select α from $\{1, 2.25, 2.5\}$, which contains their published values.
- **TGeo-KD** Hu et al. (2024). A small network predicts a per-sample fusion ratio from the geometry relating the student's prediction, the teacher's prediction, the label, and a class-average teacher prediction, trained in a bilevel fashion against a validation objective. We implement the method as described in the paper, taking the controller architecture and edge features from their released code, and approximate the bilevel step by a single unrolled student update.

- **RW-KD** Lu et al. (2021). The per-sample weights on the two loss terms are read off a one-step lookahead, each being the alignment between that term's gradient for the sample and the gradient of a validation mini-batch, so the weight derives from what currently helps the student rather than from any notion of teacher quality. We implement their algorithm with their stated $\delta = 10^{-8}$.
- **Fusions with our method.** The routing rule is independent of the distillation loss it routes, so some of the methods above can take the place of the teacher term: samples our rule routes to the teacher are trained with that method's objective, the rest on the label alone. Hyperparameters are kept from the standalone row, so the two rows differ only in the gate.

F.2 CIFAR-10

The CIFAR-10 dataset contains 60,000 color images of size 32×32 across 10 classes, split into training and test sets with a 5:1 ratio. For this dataset, we consider different image acquisition conditions, i.e., regions A and B contain images with corruptions from Hendrycks & Dietterich (2019). For the main experiments, and unless stated otherwise, region A corresponds to Gaussian blur while region B corresponds to glass blur, both with severity 3. Of the original training set, 80% are allocated to the teacher and the remaining 20% to the student. We use standard models such as ConvNet (based on VGG (Simonyan & Zisserman, 2014)), ResNet (He et al., 2016), ShuffleNet (Zhang et al., 2018), and WideResNet (Zagoruyko & Komodakis, 2016). We use standard CIFAR-10 normalization (mean $[0.4914, 0.4822, 0.4465]$ and standard deviation $[0.2470, 0.2435, 0.2616]$) and standard augmentations including pad-4 random crop and horizontal flip. The teachers are trained with SGD with momentum 0.9, an initial learning rate of 0.1 with a cosine annealing scheduler, and a batch size of 100 for 200 epochs. For ResNet teachers, we use a weight decay of 10^{-4}, while for the other teachers, we use 5×10^{-4} and dropout 0.3. For the LA, we use a last-layer K-FAC Hessian approximation with a prior precision fitted by marginal likelihood. Students are all trained with SGD and momentum of 0.9, an initial learning rate of 0.05 with a cosine annealing scheduler, weight decay of 5×10^{-4}, and a batch size of 64 for 600 epochs, over 3 seeds.

F.3 URBANSOUND8K

For audio classification we use UrbanSound8K, 8,732 clips of up to 4 seconds excerpted from 1,297 field recordings, labeled with 10 urban sound classes. We split the recordings 60:20:20 between teacher, student and test, so no recording is shared between them. Every clip is decoded to mono at 22.05 kHz and placed in a fixed 4-second window, and training and test examples are rendered from the three pools: 20,000 mixtures for the teacher, 1,500 for the student and 4,000 for testing. To reflect the overlapping sound sources of real urban recordings, each example is a mixture of two clips: a clip of a dominant class is picked along with a clip of its fixed partner class, both are RMS-normalized and mixed as $\rho \cdot s_{label} + (1 - \rho) \cdot s_{partner}$ with $\rho = 0.8$, and the mixture is RMS-normalized again. Each example is converted to a 64-band log-mel spectrogram ($n_{\text{fft}} = 1024$, hop 512), giving a 64×173 input to the models. We define region A as these standard, unmasked spectrograms, while region B introduces a SpecAugment-style frequency mask (dropping a contiguous band of 16 mel bins) to simulate a band-limited or partly occluded microphone. The models are small spectrogram classifiers from two families. The CNNs stack 3×3 conv–BN–ReLU blocks with max-pooling: CNN-8 with two convolutions per block and CNN-4 with one, plus half-width variants CNN-8-N and CNN-4-N. The ResNets (He et al., 2016) use basic residual blocks, ResNet-14 and ResNet-8. The teachers are trained with Adam, an initial learning rate of 10^{-3} with a cosine annealing scheduler, weight decay of 10^{-4}, dropout 0.3, label smoothing of 10^{-2}, and a batch size of 128 for 20 epochs. For the LA, we use a last-layer K-FAC Hessian approximation with prior precision 1. Students are all trained with SGD and momentum of 0.9, an initial learning rate of 0.01 with a cosine annealing scheduler, weight decay of 10^{-4}, dropout 0.1 in the head, and a batch size of 128 for 100 epochs, over 10 seeds.

F.4 JETCLASS

For high-energy physics, we consider jet tagging on JetClass (Qu et al., 2022), a simulated dataset of ten jet types, where the task is to identify the particle that initiated a collimated spray of outgoing particles. Jets can be represented either as the set of their constituent particles or by jet-level summary features, and we use the latter: 24 features, of which eight are provided with the dataset and sixteen are aggregated over its constituents. These include the charged multiplicity and two shape variables, the energy fractions carried by charged hadrons, neutral hadrons, photons, and leptons, the lepton count, and eight summaries of the charged particles' trajectory displacements, which carry the heavy-flavor information. Accuracies here are therefore not comparable to those of taggers trained on the full dataset with particle-level inputs. Region A contains jets with transverse momentum between 500 and 600 GeV and region B jets between 850 and 1000 GeV. This corresponds to a teacher trained on jets collected under a selection that covers only lower-momentum jets, which is then used for a student that must also handle higher-momentum jets. The teacher is an MLP with two hidden layers of 256 units and ReLU activations, and the student has the same architecture with 64 units. The teacher is trained on 100,000 jets, the student on 1,500, and both are tested on 10,000. The teacher is trained with SGD, a learning rate of 0.01, and a batch size of 1024 for 250,000 iterations. For the LA, we use a last-layer exact GGN with prior precision 0.1. Students are all trained with SGD, a learning rate of 0.03, weight decay of 10^{-3}, and a batch size of 128 for 5,000 iterations, over 10 seeds.

G ADDITIONAL RESULTS AND ABLATIONS

G.1 ADDITIONAL ARCHITECTURES

In the main body, Table 1 and Figure 3 compare our method to the baselines on teacher-student pairs from the same architecture family. We now display additional results with eight pairs whose teacher and student architectures differ. The results are presented in Table 6 and Figure 16. On almost all new pairs, our method is the highest-scoring standalone method. On ResNet-14 $\rightarrow$ CNN-4 and ResNet-14 $\rightarrow$ CNN-4-N, UBKD is higher by 1.2 and 0.7 points, less than one standard deviation of the run-to-run spread. These two pairs are also where one-hot training is strongest and plain KD is most harmful, and no standalone method gains more than 3.2 points over one-hot training on them. Fusing the router with another KD objective gives the highest accuracy on every pair, including the two on which UBKD exceeds the router alone. Specifically, fusing DKD with our method improves on our method alone on all eight pairs, by 0.3 to 3.0 points (Table 6). Moreover, as seen in Figure 16, fusing our method with complementary distillation objectives gives the same picture as in the main body: many improve upon our standalone method, and all match or exceed the same objective applied without our method.

Table 6: Accuracy (%) of the student and the PET on the student's test set, for teacher-student pairs with different architectures. All methods within a column use the same PET. Results are mean (std) over 3 runs on CIFAR-10 and 10 on UrbanSound8K. Best results are in **bold** and second best are underlined.

Modality	CIFAR-10				UrbanSound8K			
Partially expert teacher **Student**	WRN-40-2 ShuffleNetV1	ResNet-20 WRN-16-1	WRN-16-2 ConvNet-3	ConvNet-6 WRN-16-1	CNN-8 ResNet-8	ResNet-14 CNN-4	ResNet-14 CNN-4-N	CNN-8-N ResNet-8
Partially expert teacher	69.72 (0.41)	67.44 (0.70)	68.08 (0.58)	65.16 (0.17)	60.58 (2.01)	60.61 (2.33)	60.61 (2.33)	61.77 (2.46)
Standard one-hot training	78.74 (0.63)	78.79 (0.24)	69.15 (0.59)	78.62 (0.58)	51.30 (1.37)	60.10 (2.43)	57.13 (2.76)	51.39 (1.43)
KD Hinton et al. (2015)	64.02 (0.34)	65.12 (0.33)	58.25 (0.26)	63.49 (0.09)	53.32 (0.79)	52.12 (0.71)	50.42 (0.88)	54.39 (0.86)
Best constant λ	81.02 (0.77)	78.79 (0.24)	69.15 (0.59)	78.62 (0.58)	55.67 (1.68)	61.59 (1.40)	58.44 (1.47)	56.19 (1.40)
DKD Zhao et al. (2022)	78.96 (0.93)	71.02 (0.19)	63.46 (0.39)	76.04 (0.22)	55.89 (1.22)	56.90 (0.88)	55.34 (1.96)	56.80 (0.89)
ABKD Wang et al. (2025)	80.27 (0.04)	77.92 (0.06)	68.73 (0.12)	79.59 (0.51)	56.38 (1.65)	57.99 (0.69)	56.51 (1.44)	56.70 (0.75)
LSD Sun et al. (2024)	79.78 (0.64)	78.90 (0.07)	69.71 (0.13)	78.60 (0.30)	55.34 (1.56)	61.26 (1.21)	58.15 (2.13)	55.68 (0.87)
Beta-KD Sun et al. (2026)	63.43 (0.41)	63.94 (0.14)	57.48 (0.15)	62.81 (0.18)	49.04 (1.29)	50.24 (1.46)	48.97 (1.74)	51.03 (1.03)
LUD Guo et al. (2024)	79.47 (0.21)	79.32 (0.38)	67.39 (0.23)	81.10 (0.06)	53.39 (1.46)	60.42 (1.36)	57.17 (2.34)	52.06 (1.02)
UBKD Hemmatian et al. (2024)	78.91 (0.16)	79.04 (0.42)	68.54 (0.21)	79.54 (0.49)	56.49 (1.36)	62.73 (1.03)	60.30 (2.20)	56.04 (1.61)
WTTM Zheng & Yang (2024)	68.68 (0.82)	67.38 (0.25)	60.48 (0.35)	70.72 (0.32)	55.27 (1.51)	56.18 (0.62)	55.29 (1.43)	56.41 (1.14)
MCD Mishra et al. (2026)	76.76 (0.96)	76.52 (0.37)	66.15 (0.47)	78.20 (0.10)	54.21 (1.12)	59.41 (2.18)	57.21 (2.16)	55.30 (1.16)
WLS Zhou & Song (2021)	78.05 (0.57)	78.09 (0.38)	67.69 (0.24)	81.59 (0.16)	55.74 (1.22)	60.60 (1.02)	58.49 (2.45)	55.66 (1.08)
TGeo-KD Hu et al. (2024)	79.20 (0.44)	76.03 (0.37)	69.10 (0.32)	77.14 (0.32)	57.03 (1.34)	58.58 (1.07)	56.29 (2.15)	56.95 (0.98)
RW-KD Lu et al. (2021)	76.72 (0.39)	73.91 (0.23)	63.01 (0.35)	76.57 (0.31)	55.29 (1.69)	61.26 (0.85)	58.16 (1.80)	55.91 (1.08)
Ours	81.04 (0.96)	82.41 (0.34)	72.12 (0.32)	81.86 (0.14)	59.60 (1.26)	61.54 (0.65)	59.56 (1.96)	58.90 (1.21)
Ours + DKD	**82.23** (0.16)	**82.75** (0.47)	**72.61** (0.19)	**82.73** (0.14)	**60.30** (1.11)	**64.39** (1.06)	**62.57** (1.86)	**60.53** (1.08)

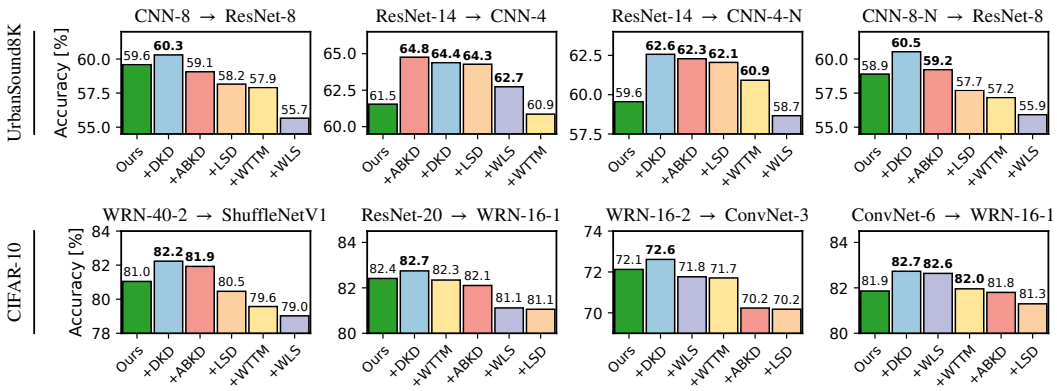

Figure 16: Test accuracy (%) of the student trained by our router combined with existing KD objectives on the eight cross-architecture teacher-student pairs of Table 6, averaged over 10 runs for UrbanSound8K and 3 runs for CIFAR-10.

G.2 Varying Expertise Type

We now demonstrate that the results generalize to additional types of expertise and non-expertise. Specifically, we consider three additional pairs of corruptions from Hendrycks & Dietterich (2019) as the region A → region B conditions: Gaussian blur → frost, defocus blur → impulse noise, and spatter → Gaussian noise, all with severity 3. We repeat the comparison from Table 1 and compare our method to the baselines and fusions for different teacher-student pairs. The results for the 8 pairs in total are given in Table 7. Across all types of expertise and for all pairs, our method, alone or fused with a published KD objective, achieves the highest accuracy. Alone, it outperforms all baselines on seven of the eight pairs and is comparable to WLS on the remaining one, defocus blur → impulse noise with ConvNet-6 → ConvNet-3, where it trails by 0.55 points, within one standard deviation of either method. For five of the eight pairs, fusing our method with a published KD objective yields further gains.

Table 7: Accuracy (%) of the student and the PET on the student's test set, on CIFAR-10 for three additional pairs of acquisition conditions, region A → region B. All methods within a column are trained from the same PET. Results are mean (standard deviation) over 3 runs. Best results are in **bold** and second best are underlined.

Region A → B	Gaussian blur → Frost			Defocus blur → Impulse noise			Spatter → Gaussian noise	
Teacher	WRN-16-2	ResNet-20	ConvNet-6	WRN-16-2	ResNet-20	ConvNet-6	WRN-16-2	ConvNet-6
Student	ConvNet-3	ResNet-20	ConvNet-3	ResNet-20	WRN-16-1	ConvNet-3	WRN-16-1	ConvNet-3
Partially expert teacher	69.42 (0.26)	69.73 (1.36)	67.40 (0.34)	70.54 (1.22)	71.93 (1.17)	67.55 (0.58)	68.06 (0.27)	64.24 (0.24)
Standard one-hot training	72.33 (0.31)	83.31 (0.09)	72.56 (0.35)	84.26 (0.03)	83.89 (0.11)	71.99 (0.41)	78.35 (0.09)	67.21 (0.25)
KD Hinton et al. (2015)	55.87 (0.21)	68.28 (0.06)	59.83 (0.16)	67.02 (0.07)	70.13 (0.17)	60.16 (0.29)	64.62 (0.00)	57.29 (0.32)
Best constant λ	72.33 (0.31)	83.31 (0.09)	72.56 (0.35)	86.56 (0.03)	83.89 (0.11)	72.92 (0.06)	78.64 (0.50)	68.87 (0.18)
DKD Zhao et al. (2022)	68.14 (0.60)	79.18 (0.28)	69.16 (0.11)	84.80 (0.42)	80.45 (0.35)	70.97 (0.31)	75.13 (1.18)	67.45 (0.39)
ABKD Wang et al. (2025)	71.67 (0.18)	83.19 (0.39)	73.16 (0.23)	86.96 (0.22)	83.80 (0.33)	73.54 (0.15)	80.17 (0.36)	69.25 (0.28)
LSD Sun et al. (2024)	72.59 (0.36)	83.50 (0.15)	72.92 (0.33)	85.20 (0.07)	83.67 (0.02)	72.67 (0.29)	79.53 (0.21)	68.68 (0.20)
Beta-KD Sun et al. (2026)	58.13 (0.11)	68.50 (0.12)	60.42 (0.03)	65.57 (0.13)	69.00 (0.29)	60.13 (0.16)	63.02 (0.34)	57.41 (0.18)
LUD Guo et al. (2024)	70.87 (0.38)	83.83 (0.14)	68.24 (0.64)	85.73 (0.50)	84.27 (0.20)	68.97 (0.52)	80.91 (0.29)	69.11 (0.23)
UBKD Hemmatian et al. (2024)	72.44 (0.38)	83.47 (0.19)	72.03 (0.50)	84.67 (0.18)	83.83 (0.33)	71.18 (0.31)	79.13 (0.13)	67.99 (0.10)
WTTM Zheng & Yang (2024)	65.94 (0.61)	77.83 (0.26)	66.45 (0.38)	79.65 (0.49)	78.79 (0.57)	66.35 (0.08)	72.74 (0.30)	63.85 (0.16)
MCD Mishra et al. (2026)	68.47 (0.36)	81.58 (0.37)	69.29 (1.98)	82.96 (0.09)	81.45 (0.32)	70.20 (0.44)	76.49 (0.50)	67.64 (1.56)
WLS Zhou & Song (2021)	72.19 (0.10)	84.48 (0.29)	73.94 (0.32)	86.90 (0.27)	84.20 (0.35)	75.19 (0.63)	82.63 (0.05)	71.63 (0.32)
TGeo-KD Hu et al. (2024)	70.83 (0.23)	80.47 (0.18)	71.57 (1.02)	85.09 (0.44)	79.81 (0.33)	72.41 (0.17)	78.23 (0.65)	69.10 (0.43)
RW-KD Lu et al. (2021)	64.93 (1.21)	80.43 (0.55)	65.32 (0.45)	81.26 (0.16)	80.65 (0.10)	66.48 (0.27)	74.41 (0.32)	64.44 (0.21)
Ours	73.66 (0.37)	85.53 (0.20)	76.57 (0.16)	**87.84** (0.19)	**85.70** (0.33)	74.64 (0.65)	**84.46** (0.13)	72.21 (0.37)
Ours + DKD	**74.30** (0.13)	85.80 (0.30)	76.80 (0.42)	87.60 (0.34)	85.67 (0.28)	75.79 (0.46)	82.59 (0.16)	73.05 (0.31)
Ours + ABKD	73.05 (0.40)	**85.99** (0.06)	74.38 (0.14)	87.54 (0.13)	85.64 (0.23)	73.62 (0.47)	82.76 (0.33)	69.91 (0.20)
Ours + LSD	73.55 (0.15)	85.47 (0.06)	73.63 (0.30)	87.10 (0.24)	85.14 (0.26)	73.51 (0.01)	82.07 (0.15)	70.29 (0.19)
Ours + WLS	73.64 (0.40)	85.03 (0.35)	**77.02** (0.24)	87.37 (0.22)	84.80 (0.27)	**77.12** (0.22)	83.19 (0.54)	**73.60** (0.48)
Ours + WTTM	73.15 (0.54)	84.87 (0.32)	74.30 (0.28)	86.40 (0.10)	84.50 (0.40)	73.36 (0.55)	81.32 (0.07)	69.85 (0.24)

G.3 Teacher Expertise Fraction

Our method decides, per sample, whether teacher supervision is preferable to one-hot supervision, and this decision has content when the student's data contains samples of both kinds. We now sweep the non-expert fraction f. Each f is a different task, so points are compared only within the same f. At $f = 0$ the teacher's and the student's generalization demands coincide, the teacher is an expert on every sample, and the rule of Proposition 2 assigns every sample to the teacher. At $f = 1$ the teacher is unreliable on every sample, and the rule assigns every sample to the label. Both edges are therefore covered by a fixed rule, and the gain from selective routing lies at intermediate f, where the teacher is partially expert.

We consider four teacher-student pairs and sweep f from 0 to 1 in steps of 0.1. For each f we report the student's accuracy under (a) our method, (b) one-hot training ($\lambda = 0$), (c) teacher supervision ($\lambda = 1$), and (d) the best constant λ, selected by a grid search separately for each f. The results are shown in Figure 17. Selective routing gains over every constant λ from $f = 0.1$ up to $f = 0.7$ or 0.8, depending on the pair, by up to 5 percentage points, with the largest gains at intermediate f. At $f = 0$ it is at most 1.5 percentage points below teacher supervision. At the upper edge, $f \geq 0.8$ for CNN-8 $\rightarrow$ CNN-4, $f \geq 0.9$ for ResNet-14 $\rightarrow$ ResNet-8 and CIFAR-10, and $f = 1$ for JetClass, it falls below one-hot training: the inputs form a single population, the mixture step splits it in two regardless, and the samples assigned to the teacher receive unreliable supervision. A guard that falls back to one-hot training when the mixture step finds one component instead of two would remove this behavior at the upper edge.

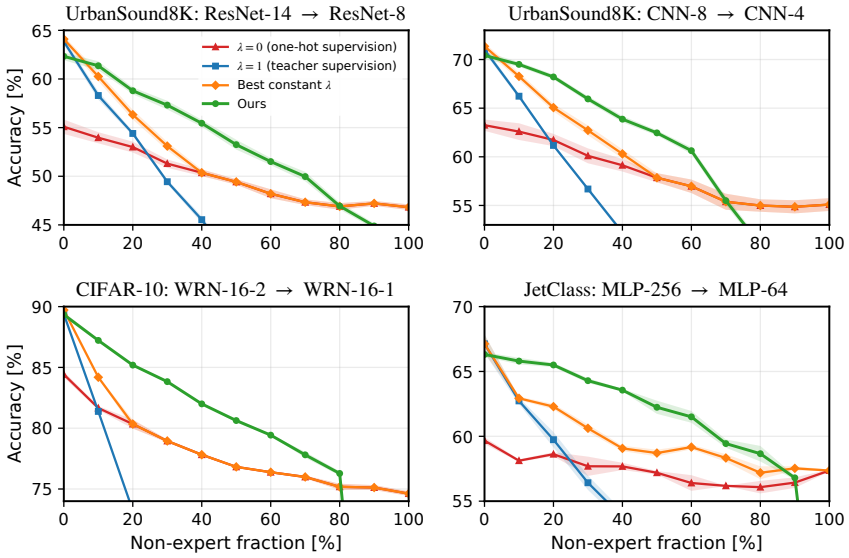

Figure 17: Test accuracy (%) of the student as the non-expert fraction f of its data is swept. Each f is a different task, so points should only be compared to other points at the same f. Shaded bands are one standard deviation over 3 runs on CIFAR-10 and 10 otherwise.

G.4 Bayesian Teachers

Our method uses the Bayesian teacher only through its AU and EU estimates, so any posterior approximation that provides them can be used. In the main body we use a post-hoc LA, and we now repeat the comparison of Table 1 with deep ensembles and VI for different teacher-student pairs. More details on these Bayesian methods are in Appendix D. The teacher architecture, the training data and all methods are unchanged, and only the Bayesian posterior approximation differs. Every method within a column uses the same teacher. The results are given in Table 8. Similarly to the main body, our method outperforms the other baselines when the teacher is a deep ensemble or a VI BNN. The one exception is the VI teacher on ResNet-20 $\rightarrow$ ResNet-20, where WLS is the most accurate standalone method, but fusing our router with it gives the best performance. As in the main body, fusing with DKD improves on our method alone in six of the nine columns.

Table 8: Accuracy (%) of the student and the PET on the student's test set, with three types of Bayesian teachers: LA, a five-member deep ensemble (DE) and VI, mean and standard deviation over 3 runs for CIFAR-10, and 10 runs otherwise. Every baseline within a column uses the same teacher predictions. For DE, the PET row is the mean (standard deviation) over the five ensemble members. Best results are in **bold** and second best are underlined.

Modality	JetClass			CIFAR-10			CIFAR-10		
Teacher → Student	MLP-256 → MLP-64			WRN-16-2 → WRN-16-1			ResNet-20 → ResNet-20		
Posterior	LA	DE	VI	LA	DE	VI	LA	DE	VI
Partially expert teacher	58.67 (1.09)	58.67 (1.09)	60.26 (0.97)	68.08 (0.58)	68.08 (0.58)	68.08 (0.64)	67.44 (0.70)	67.44 (0.70)	67.20 (0.26)
Standard one-hot training	58.14 (0.78)	58.13 (0.78)	58.13 (0.78)	78.66 (0.16)	78.84 (0.47)	78.72 (0.50)	80.30 (0.62)	79.99 (0.36)	80.46 (0.26)
KD Hinton et al. (2015)	56.35 (0.90)	57.26 (0.68)	56.88 (0.89)	65.27 (0.12)	66.16 (0.11)	64.28 (0.20)	65.90 (0.19)	66.02 (0.07)	66.70 (0.10)
Best constant λ	60.56 (0.53)	61.33 (0.50)	61.11 (0.65)	78.66 (0.16)	81.48 (0.06)	78.72 (0.50)	80.30 (0.62)	79.99 (0.36)	80.46 (0.26)
DKD Zhao et al. (2022)	60.76 (0.88)	61.68 (0.53)	59.48 (0.65)	72.52 (0.19)	80.30 (0.42)	67.20 (0.42)	72.17 (1.38)	69.66 (1.00)	73.65 (0.20)
ABKD Wang et al. (2025)	61.14 (0.55)	61.31 (0.72)	59.24 (0.61)	78.72 (0.36)	81.38 (0.32)	75.87 (0.17)	79.88 (0.35)	79.71 (0.30)	80.54 (0.19)
LSD Sun et al. (2024)	60.44 (0.41)	60.61 (0.55)	60.99 (0.49)	78.79 (0.30)	78.49 (0.39)	79.23 (0.37)	80.42 (0.41)	80.43 (0.44)	79.99 (0.33)
Beta-KD Sun et al. (2026)	54.41 (1.23)	56.80 (1.48)	56.85 (1.02)	63.76 (0.11)	64.73 (0.20)	62.30 (0.22)	64.69 (0.25)	65.09 (0.05)	65.91 (0.15)
LUD Guo et al. (2024)	57.69 (0.71)	59.20 (0.42)	62.29 (0.33)	79.25 (0.36)	81.05 (0.11)	77.28 (0.11)	80.83 (0.11)	81.81 (0.36)	79.69 (0.15)
UBKD Hemmatian et al. (2024)	61.39 (0.45)	61.29 (0.41)	61.95 (0.36)	78.23 (0.37)	79.01 (0.25)	77.78 (0.09)	80.42 (0.30)	80.22 (0.42)	79.39 (0.25)
WTTM Zheng & Yang (2024)	60.55 (0.68)	60.58 (0.46)	58.79 (0.57)	68.16 (0.19)	74.47 (0.29)	65.75 (0.14)	67.83 (0.18)	66.66 (0.29)	70.51 (0.37)
MCD Mishra et al. (2026)	59.93 (1.01)	60.51 (0.68)	60.38 (0.57)	76.18 (0.40)	79.62 (0.34)	74.94 (0.37)	78.13 (0.26)	78.41 (0.19)	78.80 (0.26)
WLS Zhou & Song (2021)	61.98 (0.54)	62.50 (0.42)	59.76 (1.03)	80.65 (0.25)	82.18 (0.62)	78.73 (0.53)	81.49 (0.11)	78.49 (0.10)	83.31 (0.25)
TGeo-KD Hu et al. (2024)	61.21 (0.39)	60.89 (0.58)	59.25 (0.71)	75.63 (0.21)	80.86 (0.27)	75.62 (0.32)	76.52 (0.05)	76.43 (0.71)	76.79 (0.25)
RW-KD Lu et al. (2021)	60.14 (0.41)	60.22 (0.42)	61.00 (0.61)	72.86 (0.30)	78.52 (0.20)	72.04 (0.30)	75.67 (0.46)	76.68 (0.44)	77.16 (0.24)
Ours	**64.43** (0.35)	**65.05** (0.34)	**64.78** (0.37)	83.58 (0.35)	84.10 (0.22)	80.94 (0.24)	83.26 (0.44)	82.92 (0.22)	80.75 (0.34)
Ours + DKD	63.85 (0.50)	64.47 (0.51)	64.41 (0.71)	**83.81** (0.38)	**84.21** (0.12)	**81.82** (0.24)	**83.97** (0.06)	**84.33** (0.25)	82.65 (0.10)
Ours + ABKD	63.79 (0.51)	64.04 (0.24)	63.78 (0.61)	82.76 (0.22)	82.37 (0.15)	81.32 (0.12)	83.56 (0.30)	83.58 (0.32)	82.95 (0.10)
Ours + LSD	63.48 (0.30)	64.33 (0.30)	64.02 (0.26)	81.50 (0.33)	81.44 (0.06)	81.36 (0.22)	83.21 (0.31)	82.89 (0.15)	82.07 (0.03)
Ours + WLS	62.56 (0.25)	62.30 (0.82)	62.35 (0.60)	82.78 (0.21)	82.58 (0.14)	79.87 (0.47)	83.05 (0.27)	82.86 (0.51)	**83.68** (0.19)
Ours + WTTM	63.52 (0.34)	64.13 (0.37)	63.77 (0.47)	82.62 (0.34)	82.75 (0.30)	81.67 (0.20)	83.71 (0.37)	82.51 (0.34)	81.46 (0.11)

We now test how the number of MC samples drawn from the teacher posterior during student training affects performance. The number of samples enters in two places. First, the AU and EU estimates are MC averages, so more samples make them more accurate (as analyzed in Figure 15), which affects the decision of which samples our method routes to one-hot supervision and which to teacher supervision. Second, the teacher's prediction used for distillation is itself an MC average, and with more samples it is a better approximation of the posterior predictive, improving the teacher's prediction. Figure 18 shows the test accuracy of students trained by our method for several teacher-student pairs across the three modalities, with the LA teacher of the main body. For some pairs the effect is large at a small number of MC samples, and performance climbs steadily as more samples are used, while for others the effect is small. Nevertheless, for all pairs the effect plateaus at about 15 samples, which is enough for our method to reach its full performance.

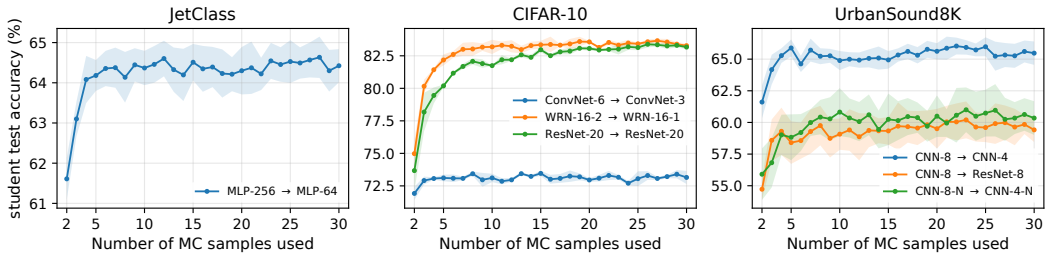

Figure 18: Test accuracy (%) of our method as a function of the number of MC samples used for computing teacher predictions, AU, and EU. Shaded bands are one standard deviation over 3 repeated draws for CIFAR-10, and 10 repeated draws otherwise.

G.5 Ablation Study

We now perform two types of ablations. The first is on the signal used for routing, i.e., on how $\lambda(\boldsymbol{x})$ is computed per sample from the AU and EU estimates, with everything else in our method unchanged. Instead of the EU fraction of (10), which uses both AU and EU, we route on (a) total entropy, the overall predictive uncertainty of the teacher, (b) AU alone, the aleatoric part representing BCP diversity, (c) EU alone, the epistemic part representing the teacher's lack of knowledge, (d) max probability, the teacher's confidence in its top prediction, and (e) agreement, which uses teacher supervision if the teacher's top prediction agrees with the label and one-hot supervision otherwise. The results are in Table 9. Our method gives the best student in five of the six pairs and is within one standard deviation of the best in the remaining pair. How well the other signals perform is connected to which of the two costs each captures. In our method, the EU fraction compares the teacher's error, estimated by EU, against the BCP diversity, estimated by AU, and is the only signal that accounts for both the cost of teacher supervision and the cost of one-hot supervision. The next best signals are EU alone and agreement, which both capture teacher reliability, and hence the cost of teacher supervision, but not the cost of one-hot supervision. EU alone is not comparable across samples, as ambiguous samples carry more EU than clean ones even when the teacher knows them equally well. Agreement reads any disagreement between the label and the teacher's top prediction as a teacher error, so on samples with diverse BCP it withholds teacher supervision exactly where it is most valuable. The remaining signals do not capture teacher reliability at all: AU alone cannot tell a confidently wrong teacher from a confidently right one, and total entropy and max probability mix the two kinds of uncertainty without separating them.

Table 9: An ablation on the signal used for routing. Each row replaces the EU fraction $r(\boldsymbol{x})$ (10) in our method with a different signal and keeps everything else in our method unchanged. Each cell reports the test accuracy of a student trained by replacing the signal minus that of a student trained with our method (pp), paired per run, mean (standard deviation) over 3 runs for CIFAR-10, and 10 runs otherwise, so a negative value means the replacement is worse. In each column the best replacement is in **bold** and the second best is underlined.

Modality	JetClass	CIFAR-10	CIFAR-10	UrbanSound8K	UrbanSound8K	UrbanSound8K
Teacher	MLP-256	WRN-16-2	ResNet-20	CNN-8	ResNet-14	CNN-8
Student	MLP-64	WRN-16-1	ResNet-20	CNN-4	ResNet-8	ResNet-8
EU alone	**-0.03** (0.38)	-1.93 (0.19)	-1.64 (0.18)	-2.52 (1.08)	**+0.81** (1.45)	-3.09 (1.42)
Agreement	-2.88 (0.30)	**-0.56** (0.40)	**-0.13** (0.14)	**-1.74** (1.02)	-3.59 (1.78)	**-3.01** (1.32)
Total entropy	-5.17 (0.79)	-1.52 (0.10)	-0.80 (0.19)	-3.25 (2.64)	-1.68 (3.57)	-6.14 (2.67)
Max probability	-5.47 (0.60)	-1.86 (0.54)	-2.72 (0.14)	-2.92 (2.74)	-0.63 (1.91)	-5.98 (3.02)
AU alone	-5.30 (0.75)	-1.33 (0.36)	-0.86 (0.25)	-3.46 (2.38)	-3.45 (1.27)	-7.31 (3.07)

Table 10: An ablation on how the routing decision is formed from the AU and EU estimates, with everything else in our method unchanged. Each row keeps our signals, AU and EU, and changes only how they are turned into $\lambda(\boldsymbol{x})$. Each cell reports the test accuracy of a student trained with the replaced decision rule minus that of a student trained with our method (pp), paired per run, mean (standard deviation) over 3 runs for CIFAR-10, and 10 runs otherwise, so a negative value means the replacement is worse. In each column the best replacement is in **bold** and the second best is underlined.

Modality	JetClass	CIFAR-10	CIFAR-10	UrbanSound8K	UrbanSound8K	UrbanSound8K
Teacher	MLP-256	WRN-16-2	ResNet-20	CNN-8	ResNet-14	CNN-8
Student	MLP-64	WRN-16-1	ResNet-20	CNN-4	ResNet-8	ResNet-8
Median gate on r	-1.37 (0.22)	**+0.66** (0.07)	**+0.38** (0.14)	+0.50 (1.24)	+0.21 (1.22)	-0.54 (1.19)
Two-means split on r	+0.08 (0.33)	-6.73 (0.22)	-3.53 (0.19)	-0.71 (0.80)	+0.08 (0.64)	-0.17 (1.03)
Soft mixture on r	-0.06 (0.28)	+0.09 (0.11)	-0.12 (0.37)	+0.55 (0.56)	+0.12 (0.79)	+0.88 (1.35)
Sigmoid at the decision point on r	-1.33 (0.26)	-3.17 (0.28)	-3.15 (0.68)	-1.22 (0.52)	-1.27 (0.62)	-0.41 (0.57)
Mixture on AU and EU	**+0.26** (0.61)	-1.88 (0.25)	-2.08 (0.35)	-0.52 (1.14)	+0.00 (1.75)	-1.23 (1.62)
Soft mixture on AU and EU	+0.18 (0.63)	-1.93 (0.05)	-2.24 (0.07)	**+0.75** (1.12)	**+2.40** (1.09)	**+1.03** (0.91)
Two-means split on AU and EU	-4.53 (0.65)	-1.11 (0.22)	+0.10 (0.14)	-2.73 (1.02)	-0.61 (2.19)	-3.03 (1.09)

The second ablation is on how the routing decision is formed from the AU and EU estimates, with everything else in our method unchanged. Our method fits a two-component Gaussian mixture to $r(\boldsymbol{x})$ over the student's data and assigns each sample to its more likely component, as described in Algorithm 1. We compare against (a) a median gate, which routes half of the data to one-hot supervision, (b) a two-means split, (c) a soft version of our mixture, where instead of a hard assignment each sample gets $\lambda(\boldsymbol{x})$ equal to the mixture's probability that it belongs to the familiar component, (d) a sigmoid around our decision point, which smooths the hard threshold over a width set by the spread of the two components, and (e) hard, soft, and two-means variants fitted on the two-dimensional AU and EU pair instead of on their combination into the EU fraction $r(\boldsymbol{x})$. The results are in Table 10. No alternative improves on our decision rule consistently. The soft version of the same mixture is within one standard deviation of it on every pair, so the hard assignment of Proposition 2 is not a restriction in practice. Every other variant, whether it changes how the threshold is fitted (median, two-means, sigmoid) or fits on the AU and EU pair instead of on $r(\boldsymbol{x})$, falls below our rule by more than one standard deviation on at least one pair, and those that are above it on some pairs are below it on others. In these experiments, how the decision is made matters less than what it is based on: the gains come from using both indicators, of BCP diversity and of teacher error, as the first ablation shows.

H LIMITATIONS AND FUTURE WORK

• Our theoretical results in Section 3 rely on several assumptions commonly used in optimization literature, such as strong quasi-convexity or smoothness. These assumptions, while standard, can of course be questioned.

• Our proposed method advocates the use of BNN teachers, possibly obtained by converting a pretrained deterministic teacher into a BNN with the LA. Still, one has to account for the additional computational complexity of BNNs. Nevertheless, while posterior sampling introduces additional computational overhead during distillation, the final deployed student model retains its standard architecture and incurs zero inference overhead.

• Our proposed method improves upon existing KD methods across modalities and architectures in PET settings, yet the fidelity of the AU and EU proxies depends on the posterior approximation, where residual entanglement (Mucsányi et al., 2024) or overconfident predictions can arise. This motivates exploring alternative uncertainty representations, such as credal sets, and deriving formal bounds connecting Bayesian proxies to BCP diversity and teacher error. This would enable the costs of Proposition 2 to be evaluated directly from the proxies themselves, thereby strengthening the theoretical coupling between the routing rule and its underlying analysis.