Density-Ratio Losses for Post-Hoc Learning to Defer

Alexander Soen
KTH Royal Institute of Technology
asoen@kth.se

Ragnar Thobaben
KTH Royal Institute of Technology
ragnart@kth.se

Joakim Jaldén
KTH Royal Institute of Technology
jalden@kth.se

Richard Nock
Google Research
richardnock@google.com

Abstract

We study post-hoc Learning to Defer (L2D) through the lens of ideal distributions: divergence-regularized reweightings of the data distribution under which a model attains low task loss. We define deferral via the density-ratio between a model's and an expert's ideals. Using the reduction from density-ratio estimation to class-probability estimation, we derive the *DR CPE* losses for post-hoc L2D scorers. Deferral decisions are then made by thresholding the scorer. For KL-based ideal distributions, our deferral rule recovers Chow's rule under the original distribution and a connection to an expert-tilted Bayes posterior—which incorporates the expert's performance—depending on if the ideal distributions are marginal or joint distributions. This leads to two variants of the DR CPE losses, with the joint variant being computationally convenient in practice. We establish consistency properties of the joint DR CPE losses when considering the zero-one task loss and strongly proper composite CPE losses. Experimentally, our approach is competitive compared to common baselines and is more robust across dataset settings. More broadly, our results cast post-hoc L2D as density-ratio estimation between ideal distributions, bridging Chow-style rules, expert comparison, and elucidating connections to related learning settings including anomaly detection.

1 Introduction

Machine learning (ML) models are now embedded in many tasks we encounter every day. As these systems are deployed in increasingly high-stakes domains, including healthcare [Esteva et al., 2017, Kadampur and Al Riyaee, 2020, Zoabi et al., 2021, Ma et al., 2024, Singhal et al., 2025] and finance [Burrell, 2016, Yousefi et al., 2019, Li et al., 2023], it is crucial that they make predictions only when their outputs are reliable. In such high-stakes decision-making systems, ideally, a synthesis of both ML prediction and human expertise can be used [Topol, 2019]: ML models can defer to experts whenever their own predictions are deemed unreliable, and the burden on human experts can be reduced by leveraging ML models. Dual to the problem of predicting responsibly, the rise of huge ML models and edge deployment has created strong demand for systems that improve efficiency by routing only difficult inputs to a larger model while handling easier cases with a smaller model [Viola and Jones, 2001, Mamou et al., 2022, Varshney and Baral, 2022, Khalili et al., 2022, Dohan et al., 2022, Gupta et al., 2024, Narasimhan et al., 2025]. Learning to defer (L2D) proposes a solution to these problems by modifying the standard supervised learning setting to allow (smaller) models to defer their predictions to an upstream (potentially human) expert [Madras et al., 2018].

To allow models to abstain from predictions, learning to reject (L2R) was initially proposed [Chow, 1957, 1970, Cortes et al., 2016], where rejection is learned without access to the downstream expert used at inference. A useful categorization of these approaches are by how abstention is represented.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

In score-based approaches, abstention is derived from the predictor's own scores, encompassing classical reject-option rules, convex-surrogate formulations, and confidence-threshold methods [Chow, 1957, 1970, Herbei and Wegkamp, 2006, Bartlett and Wegkamp, 2008, Yuan and Wegkamp, 2010, Cao et al., 2022, El-Yaniv and Wiener, 2010, Geifman and El-Yaniv, 2017, 2019, Ni et al., 2019]. In predictor-rejector approaches, a separate rejector is learned [Cortes et al., 2016, Mao et al., 2024a]. Finally, reduction-based approaches cast rejection as a related supervised learning problem [Charoenphakdee et al., 2021, Ramaswamy et al., 2018]. L2D extends this picture by incorporating an expert into training [Madras et al., 2018]. Here too, one can distinguish score-based formulations, where deferral is encoded through shared scores or defer classes [Mozannar and Sontag, 2020, Verma and Nalisnick, 2022, Mao et al., 2024b], in contrast to predictor-rejector formulations with a separate deferral function [Okati et al., 2021, Mozannar et al., 2023, Mao et al., 2023]. A second, orthogonal distinction is between end-to-end methods and post-hoc (two-stage) methods.

The post-hoc regime is especially important since retraining modern pretrained models can be prohibitively expensive. We focus on post-hoc L2D [Kag et al., 2022, Jitkrittum et al., 2023, Soen et al., 2024] (also called two-stage L2D [Mao et al., 2023]), where the base model is fixed, and only the deferral rule is learned. In this setting, simple score-based heuristics can fail when downstream experts are specialized or under distribution shift [Jitkrittum et al., 2023], while many existing methods still amount to estimating pointwise expert-comparison targets, which aim to predict when the expert will outperform the model [Kag et al., 2022, Jitkrittum et al., 2023]. A desirable property in this regime is for deferral models to be adaptable to different rates of deferral without full retraining.

Recently, a distributional perspective on post-hoc L2R was proposed [Soen et al., 2024, Soen, 2026]. Its central object is an *ideal distribution*: a divergence-regularized reweighting of the data distribution under which a fixed predictor attains low loss. Rejection is then defined by thresholding a density-ratio, using either a marginal construction over $\mathcal{X}$ [Soen et al., 2024] or a joint construction over $\mathcal{X} \times \mathcal{Y}$ that is subsequently marginalized to $\mathcal{X}$ [Soen, 2026]. Extending this framework to post-hoc L2D is nontrivial: the resulting rules depend on the Bayes-optimal class-posterior and a direct adaption would additionally require pointwise access of the expert at test time.

In this paper, we make the following contributions (Fig. 1 summarizes our overall approach):

- *L2D via ideal distributions*: We introduce a distributional formulation of post-hoc L2D based on ideal distributions. Unlike L2R, we consider a pair of ideal distributions, one for the model h and one for the expert $h^{(e)}$. Deferral is then defined by thresholding the density-ratio between these ideal distributions (see Definitions 3.1 and 3.2).
- *Marginal versus joint ideals*: Our framework accommodates both marginal and joint ideal distributions. For KL-ideal distributions, we show that the marginal rule recovers Chow's rule w.r.t. the data distribution, whereas the joint rule corresponds to a Chow-style decision under an expert-tilted distribution. This tilt is practically meaningful as it augments the expert model's performance into the optimal deferral. We further relate the induced learning objectives (see Theorem 3.5 and Lemma 4.3).
- *The DR CPE losses*: Using the reduction from Density-Ratio Estimation (DRE) to Class-Probability Estimation (CPE), we derive a family of *DR CPE* post-hoc deferral losses. Consistency of the joint DR CPE loss objective is examined through a general regret bound with an approximation gap and an exact result for zero-one task loss (see Theorems 4.4 and 4.5). We further connect our method to expert-comparison deferral estimation [Kag et al., 2022, Jitkrittum et al., 2023] and anomaly detection [Steinwart et al., 2005] (see Definition 4.2).
- *Empirical performance*: Our DR CPE approach performs competitively against other L2D baselines and is also more robust against different dataset corruption combinations than these baselines. Furthermore, performance is maintained when utilizing different rates of deferral without retraining.

2 Settings

We consider a space of observations $\mathcal{X}$ and labels $\mathcal{Y} \stackrel{\text{def}}{=} [L] = \{1, 2, \ldots, L\}$, with a data distribution $\mathrm{P} \in \triangle(\mathcal{X} \times \mathcal{Y})$. The learning task we consider is to learn classifiers $f \colon \mathcal{X} \to \mathcal{Y}$ w.r.t. the data distribution P. Instead of directly learning a classifier f, we will consider scorers $h \colon \mathcal{X} \to \mathbb{R}^L$ which are related via $f(x) = \arg\min_{y \in \mathcal{Y}} h_y(x)$. To measure performance, we consider a task loss function $\ell \colon \mathcal{X} \times \mathcal{Y} \times \mathbb{R}^L$ to define the expected risk $\mathbb{E}_{\mathrm{P}}[\ell(\mathsf{X}, \mathsf{Y}, h(\mathsf{X}))]$. We shorthand a decomposition of the

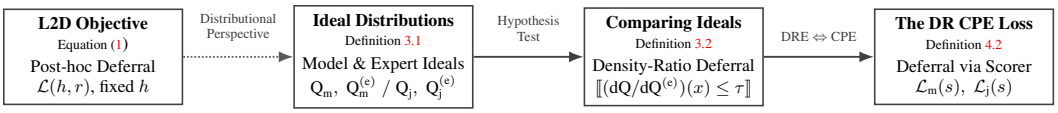

Figure 1: Post-hoc L2D to learnable density-ratio (DR) deferral. Starting from the post-hoc L2D objective in Eq. (1), we reformulate deferral through model and expert ideal distributions (Definition 3.1), compare them via a DR (Definition 3.2), and then derive learnable scorer objectives using the reduction from DR Estimation (DRE) to Class-Probability Estimation (CPE) (Definition 4.2).

data distribution by $\mathrm{P}(x, y) = \mathrm{P}_{\mathsf{x}}(x)\eta^{\star}_{y}(x)$, where $\mathrm{P}_{\mathsf{x}} \in \triangle(\mathcal{X})$ is the marginal over observations and $\eta^{\star}(x) \in \triangle(\mathcal{Y})$ corresponds to the Bayes class-posterior function, *i.e.*, $\eta^{\star}_{y}(x) = \mathrm{P}[\mathsf{Y} = y \mid \mathsf{X} = x]$.

Learning to Defer To move from the standard multi-class classification paradigm to learning to defer (L2D), alongside the typical optimization of classification model $h\colon \mathcal{X} \to \mathbb{R}^{L}$, we introduce an *expert* model $h^{(\mathsf{e})}\colon \mathcal{X} \to \mathbb{R}^{L}$. Using the expert, alongside the classifier h, we can also jointly learn a rejector $r\colon \mathcal{X} \to \{0, 1\}$ via an objective [Mozannar and Sontag, 2020]:

$$\mathcal{L}(h, r; c) \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{\mathsf{X},\mathsf{Y}\sim\mathrm{P}}\Big[(1 - r(\mathsf{X})) \cdot \ell(\mathsf{X}, \mathsf{Y}, h(\mathsf{X})) + r(\mathsf{X}) \cdot \ell(\mathsf{X}, \mathsf{Y}, h^{(\mathsf{e})}(\mathsf{X}))\Big] + c \cdot \mathop{\mathbb{P}}_{\mathsf{X}\sim\mathrm{P}_{\mathsf{x}}}[r(\mathsf{X}) = 1], \quad (1)$$

In Eq. (1), we explicitly separate the expert's performance $\ell(\mathsf{X}, \mathsf{Y}, h^{(\mathsf{e})}(\mathsf{X}))$ and the additional cost of calling the expert $c \geq 0$. If one removes the loss term corresponding to the expert $\ell(\mathsf{X}, \mathsf{Y}, h^{(\mathsf{e})}(\mathsf{X}))$, the objective simplifies to the learning to reject (L2R) objective where the performance of how deferred observations are dealt with are not explicitly modeled [Chow, 1970].

The L2D objective balances two tasks. First, it wants to decrease the loss of h by smartly deferring observations x to the expert $h^{(\mathsf{e})}$ for inputs it will predict poorly for. However, we do not want the rejector r to defer too much, and thus Eq. (1) balances the rate of deferral in its second term. Once one has a pair of classifier and rejector (h, r), one can derive a classifier which can defer to an expert $h^{(\mathsf{e})}$ through a combination $\bar{h}(x) = [\![r(x) = 0]\!] \cdot h(x) + [\![r(x) = 1]\!] \cdot h^{(\mathsf{e})}(x)$.

While Eq. (1) is posed as a joint optimization over (h, r), we can also imbue a fixed classifier h with a deferral mechanism, yielding the Bayes-optimal rejector in the post-hoc L2D setting.

Lemma 2.1 (Chow's Rule [Chow, 1970])**.** *For fixed classifier h, the Bayes-optimal deferral rule $r^{\star}$ of Eq. (1) is given by*

$$r^{\star}(x; c, \mathrm{P}) = \Big[\!\Big[\mathbb{E}_{\mathsf{Y}\sim\eta^{\star}(x)}\Big[\ell(x, \mathsf{Y}, h(x)) - \ell(x, \mathsf{Y}, h^{(\mathsf{e})}(x))\Big] \geq c\Big]\!\Big], \quad (2)$$

where $[\![\mathtt{pred}]\!]$ are Iverson brackets which evaluate to 1 when the predicate $\mathtt{pred}$ is true and is equal 0 when the predicate is false [Knuth, 1992].

Lemma 2.1 presents an intuitive mechanism for deciding which examples to defer to an expert: if the loss of h is expected to be much higher (w.r.t. threshold value c) than the expert $h^{(\mathsf{e})}$, then one should defer. In the sequel, we will solely consider the setting of post-hoc L2D.

3 Divergence-Based Deferral Rules

Eq. (1) presents a loss-based objective for learning rejectors. Alternatively, instead of directly learning mechanisms for deferral as a binary function r, one can first aim to learn *ideal distributions* and then define deferral decision corresponding to their density-ratios [Soen et al., 2024, Soen, 2026].

Ideal Distributions An ideal distribution is intuitively a distribution in which a classifier-loss pair (h, ℓ) performs best at within a specified (regularized) distance away from the true probability measure P. Distance is defined by a statistical divergence. In the following definition, we define two types of ideal distribution which differ in whether we are optimizing within a marginal or joint space.

Definition 3.1. *Given a fixed classifier-loss pair (h, ℓ), divergence* D*, and $\gamma > 0$:*

- *An* (D, γ)*-marginal ideal distribution* $\mathrm{Q_m} \in \triangle(\mathcal{X})$ *of* (h, ℓ) *is given by*
$$\mathrm{Q_m} \in \operatorname*{arg\,min}_{\mathrm{Q} \in \triangle(\mathcal{X})} \mathop{\mathbb{E}}_{\mathsf{X} \sim \mathrm{Q}} \mathop{\mathbb{E}}_{\mathsf{Y} \sim \eta^\star(\mathsf{X})} [\ell(\mathsf{X}, \mathsf{Y}, h(\mathsf{X}))] + \gamma \cdot \mathrm{D}(\mathrm{Q} \,\|\, \mathrm{P_x}). \tag{3}$$
- *An* (D, γ)*-joint ideal distribution* $\mathrm{Q_j} \in \triangle(\mathcal{X})$ *of* (h, ℓ) *is given by*
$$\mathrm{Q_j}(x; \gamma) = \int_{\mathcal{Y}} \bar{\mathrm{Q}}_\mathrm{j}(x, y)\, \mathrm{d}y, \qquad \bar{\mathrm{Q}}_\mathrm{j} \in \operatorname*{arg\,min}_{\bar{\mathrm{Q}} \in \triangle(\mathcal{X} \times \mathcal{Y})} \mathop{\mathbb{E}}_{(\mathsf{X}, \mathsf{Y}) \sim \bar{\mathrm{Q}}} [\ell(\mathsf{X}, \mathsf{Y}, h(\mathsf{X}))] + \gamma \cdot \mathrm{D}(\bar{\mathrm{Q}} \,\|\, \mathrm{P}). \tag{4}$$

Fitting our intuition, the ideal distributions defined in Definition 3.1 are found by explicitly minimizing the loss ℓ of the classifier h. The regularization weights γ control how far the ideal distributions are from the ground truth distribution P. We note that the definition abuses notation slightly, where the divergence D corresponds to both a divergence on the marginal space and also on the joint space.

Generalized Variational Inference (GVI) and Distributionally Robust Optimization (DRO): It has been previously noted in Soen et al. [2024] that Eqs. (3) and (4) resemble the optimization problems associated with GVI [Zellner, 1988, Knoblauch et al., 2022] and DRO [Scarf et al., 1957, Duchi et al., 2021, Zhang et al., 2021]. Indeed, $\mathrm{Q_i}$ corresponds to solutions to a GVI problem and is analogous to an adversarial distribution in DRO.

As the ideal distribution in Definition 3.1 corresponds to the distributions in which a classifier best performs at, we propose a deferral rule for L2D by comparing the ideal distributions of our classifier h versus the ideal distribution of our expert classifier $h^{(\mathrm{e})}$. More precisely, we propose considering the density-ratio between ideal distributions.

Definition 3.2. *Let* $\mathrm{i} \in \{\mathrm{m}, \mathrm{j}\}$. *Given* (D, γ)*-ideal distributions* $\mathrm{Q_i}$ *and* $\mathrm{Q_i^{(e)}}$ *of* (h, ℓ) *and* $(h^{(\mathrm{e})}, \ell)$*, respectively, the* D*-density-ratio deferral rule is given by*
$$r_\mathrm{i}(x; \tau, \gamma, \gamma^{(\mathrm{e})}) \stackrel{\text{def}}{=} [\![\rho_\mathrm{i}(x; \gamma, \gamma^{(\mathrm{e})}) \leq \tau]\!], \quad \textit{where } \rho_\mathrm{i}(x; \gamma, \gamma^{(\mathrm{e})}) \stackrel{\text{def}}{=} \frac{\mathrm{dQ_i}(x; \gamma)}{\mathrm{dQ_i^{(e)}}(x; \gamma^{(\mathrm{e})})}. \tag{5}$$

An advantage of making deferral decision through density-ratios is that the functional form of r_i becomes convenient once we make our choice of divergence D (*i.e.*, by considering φ-divergences).

Hypothesis Testing: The deferral decision of Definition 3.2 has a direct connection to hypothesis testing. Thresholding density-ratios corresponds exactly to optimal hypothesis testing in the Neyman-Pearson sense [Neyman and Pearson, 1933, Lehmann and Romano, 2005], *i.e.*, it is a likelihood ratio test. More concretely, $r_\mathrm{i}(x; \tau, \gamma, \gamma^{(\mathrm{e})})$ corresponds to the hypothesis test
$$H_0 \colon x \sim \mathrm{Q_i^{(e)}}(\cdot; \gamma^{(\mathrm{e})}) \quad \text{and} \quad H_1 \colon x \sim \mathrm{Q_i}(\cdot; \gamma). \tag{6}$$
The rule fails to reject the null hypothesis H_0, and hence defers to the expert, whenever the likelihood ratio $\mathrm{dQ_i}/\mathrm{dQ_i^{(e)}}$ is sufficiently small, *i.e.*, whenever $\rho_\mathrm{i}(x; \gamma, \gamma^{(\mathrm{e})}) \leq \tau$. Intuitively, this means that deferral occurs precisely when the observation x is better explained by the expert's ideal distribution than by the classifier's ideal distribution. Equivalently, the test asks which of the two classifiers performs better at x: if x is more likely under the expert's ideal distribution, then the expert's predictions are better supported there, whereas if x is more likely under the classifier's ideal distribution, then the evidence favors keeping the prediction with h rather than deferring to $h^{(\mathrm{e})}$.

Defining deferral decision via the density-ratios of ideal distributions was originally utilized for L2R [Soen et al., 2024]. In this setting, instead of considering a pair of classifiers h, $h^{(\mathrm{e})}$, L2R only considers a single classifier h. As a result, there is only a single ideal distribution, and the distribution we compare it against is just the data measure P. Although not stated in prior work, the hypothesis testing interpretation corresponding to Eq. (6) (replacing $\mathrm{Q^{(e)}}$ with P) also holds for the L2R setting.

KL-Divergence In prior work, it was shown that a canonical choice of divergence for constructing ideal distributions is the KL-divergence [Soen et al., 2024, Soen, 2026]. In this work, we will focus on this class of ideal distributions generated by this canonical choice. We first recall that the KL-divergence between probability measures $\mathrm{P} \ll \mathrm{Q}$ is given by [Amari and Nagaoka, 2000]: $\mathrm{KL}(\mathrm{P} \,\|\, \mathrm{Q}) \stackrel{\text{def}}{=} \mathbb{E}_{\mathsf{Z} \sim \mathrm{P}} [\log (\mathrm{dP}/\mathrm{dQ}(\mathsf{Z}))]$.

Utilizing the KL-divergence, both of the corresponding marginal and joint ideal distributions can be shown to be (exponential) multiplicative re-weightings of the true data distribution P.

Corollary 3.3. *Given a fixed classifier-loss pair (h, ℓ), the (KL, γ)-ideal distributions are multiplicative re-weightings $\mathrm{Q}_{\mathrm{i}}(x;\gamma) = w_{\mathrm{i}}(x;\gamma) \cdot \mathrm{P}_{\mathrm{x}}(x)$, with weights given by*

$$w_{\mathrm{i}}(x;\gamma) \stackrel{\text{def}}{=} \frac{1}{Z_{\mathrm{i},\gamma}} \tilde{w}_{\mathrm{i}}(x;\gamma), \quad \tilde{w}_{\mathrm{i}}(x;\gamma) \stackrel{\text{def}}{=} \begin{cases} \exp\left(-\frac{\mathbb{E}_{\mathsf{Y}\sim\eta^\star(x)}[\ell(x,\mathsf{Y},h(x))]}{\gamma}\right) & \text{if } \mathrm{i} = \mathrm{m} \\ \mathbb{E}_{\mathsf{Y}\sim\eta^\star(x)}\left[\exp\left(-\frac{\ell(x,\mathsf{Y},h(x))}{\gamma}\right)\right] & \text{if } \mathrm{i} = \mathrm{j}, \end{cases} \tag{7}$$

where $Z_{\mathrm{i},\gamma}$ are normalizing constants of the distributions, i.e., $Z_{\mathrm{i},\gamma} = \mathbb{E}_{\mathsf{X}\sim\mathrm{P}_{\mathrm{x}}}[\tilde{w}_{\mathrm{i}}(\mathsf{X};\gamma)]$.

Corollary 3.3 follows directly from a more general result that holds for φ-divergences (a.k.a., Csiszár f-divergences), as shown in Appendix C. In this more general case, the ideal distribution generated by φ-divergences are also multiplicative re-weightings of P. However, only the KL-divergence's *normalization* are determined by a multiplicative constant $1/Z_{\mathrm{i},\gamma}$. This latter property will be key in our construction of loss functions in Section 4.

From the functional form presented in Corollary 3.3, it becomes apparent why taking the threshold of density-ratios suits the functional form of KL-divergence (and φ-divergence) ideal distributions: dependence on the marginal data measure $\mathrm{P}_{\mathrm{x}}(x)$ will cancel out in the density-ratio. As a result, for optimal ideal distributions, the rejection decision follows from a ratio of weights defined in Corollary 3.3. Furthermore, in this case, the role of the regularizer γ becomes clear: γ corresponds to the *temperature* of the multiplicative reweighting.

The choice of utilizing the KL-divergence for ideal distributions is known to be canonical as it can be shown that in the L2R setting, rejection via the density-ratio mechanism and marginal ideal distributions recovers Chow's rule [Soen et al., 2024, Theorem 4.2]. Specific to the L2R setting, it was also shown that rejectors corresponding to joint ideal distributions are more lenient than their marginal ideal distribution counterparts [Soen, 2026, Lemma 1].

The shift from working with one ideal distribution to a pair of ideal distributions in L2D changes the relationship between marginal and joint ideal distributions that was previously established in L2R. The connection to Chow's rule for the marginal ideal distribution is retained; however, the analogous connection between the joint and marginal ideal distributions no longer holds for deferral mechanisms of the form in Definition 3.2. Instead, the joint deferral mechanism can be related to Chow's rule under a data distribution with a *tilted* Bayes class-posterior.

Definition 3.4. *For a distribution* P*, we denote its (ℓ, γ)-expert-tilted distribution as*

$$\tilde{\mathrm{P}}_\gamma(x,y) \stackrel{\text{def}}{=} \mathrm{P}_{\mathrm{x}}(x) \cdot \tilde{\eta}^\star_y(x;\gamma), \quad \tilde{\eta}^\star_y(x;\gamma) \stackrel{\text{def}}{=} t(y \mid x;\gamma) \cdot \eta^\star_y(x), \tag{8}$$

where

$$t(y \mid x;\gamma) \stackrel{\text{def}}{=} \frac{\exp\left(-\ell(x,y,h^{(\mathrm{e})}(x))/\gamma\right)}{\mathbb{E}_{\mathsf{Y}\sim\eta^\star(x)}\left[\exp\left(-\ell(x,\mathsf{Y},h^{(\mathrm{e})}(x))/\gamma\right)\right]}. \tag{9}$$

The informed reader may identify the transformation of the Bayes class-posterior as the exponential tilting of a distribution [Escher, 1932] or as the Boltzmann-Gibbs transformation [Moral, 2004].

Theorem 3.5. *For any cost c, there exists a $\tau > 0$ such that the KL-marginal density-ratio deferral rule satisfies $r_{\mathrm{m}}(x;\tau,\gamma,\gamma) = r^\star(x;c,\mathrm{P})$. Furthermore, there exists a $\tau > 0$ such that the KL-joint density-ratio deferral rule satisfies $r_{\mathrm{j}}(x;\tau,\gamma,\gamma) \leq r^\star(x;c,\tilde{\mathrm{P}}^{(\mathrm{e})}_\gamma)$.*

The second part of Theorem 3.5 shows that the joint density-ratio deferral rule is pointwise more conservative than Chow's rule under the expert-tilted distribution $\tilde{\mathrm{P}}_\gamma$. In particular, whenever the joint rule defers at an input x, Chow's rule under $\tilde{\mathrm{P}}_\gamma$ also defers at x. Equivalently, the rejection region of the joint rule is contained in the rejection region induced by Chow's rule under the tilted measure.

Beyond its theoretical interest, the second part of Theorem 3.5 also provides a useful practical interpretation of deferral. In many applications, the Bayes class-posterior $\eta^\star$ is not, on its own, the most relevant object for deciding whether to predict or defer. Indeed, $\eta^\star$ only captures the (aleatoric) label uncertainty under the data distribution, whereas the deferral decision should also reflect the downstream loss and the expert's comparative advantage. The tilted posterior $\tilde{\eta}^\star(x;\gamma)$ incorporates this additional information by reweighting labels according to the expert loss $y \mapsto \ell(x,y,h^{(\mathrm{e})}(x))$. Consequently, applying a Chow-type rule under $\tilde{\mathrm{P}}_\gamma$ means deferring not simply when the labels are ambiguous, but when they are ambiguous in a way that favors the expert. This can be preferable in

practice, since deferral is only useful in-so-far that the expert handles the prediction task better than the deferring model. From this perspective, Chow's rule under the tilted posterior ranks examples not only by uncertainty, but by uncertainty that is especially relevant to the expert.

4 Density-Ratio Estimation Losses from Ideal Distributions

So far we have seen that ideal distributions, together with thresholding their associated density-ratios, can yield analytical and even optimal deferral rules. However, directly applying the closed-form expressions in, *e.g.*, Corollary 3.3, raises two fundamental difficulties. First, both the marginal and joint ideal distributions (and hence the resulting density-ratio) depend on the Bayes class-posterior $\eta^\star$. Using these expressions as plug-in deferral rules would therefore require an approximation of $\eta^\star$ which can be suboptimal if not carefully calibrated. Second, the deferral rule also depends on the expert ideal distribution $\mathrm{Q}_{\mathrm{i}}^{(\mathrm{e})}$, which in turn requires pointwise evaluation of the expert model $h^{(\mathrm{e})}(x)$ at inference time. If the expert has already been queried on x, then (for many settings) one may as well use the expert's prediction directly rather than first using it only to decide whether to defer.

These two issues prevent us from directly utilizing the closed-form analytical solutions for L2D. However, instead of trying to use a plug-in estimator to evaluate ideal distribution density-ratios, we propose a method which aims to estimate the density-ratio $\mathrm{dQ_i}/\mathrm{dQ_i^{(e)}}$. To do so, we rely on links between *Density-Ratio Estimation* (DRE) [Sugiyama et al., 2012] and *Class-Probability Estimation* (CPE) [Buja et al., 2005, Reid and Williamson, 2011], where we reduce the problem of estimating the density-ratio to a binary classification problem. Unique to our setting, the binary classification problem will be to distinguish between two reweightings of the data distribution P with weights determined by the performance of the model h and expert $h^{(\mathrm{e})}$.

DRE to CPE Before specifying the CPE optimization problem for estimating the density-ratio between ideal distributions, we revise how general DRE can be expressed as a CPE problem. Suppose that we have a strictly proper composite loss $\lambda\colon \mathbb{R} \times \{\pm 1\} \to \mathbb{R}$ with invertible link function $\Phi\colon [0,1] \to \mathbb{R}$ [Reid and Williamson, 2010]. Recall that to estimate the density-ratio $\mathrm{d}\mu/\mathrm{d}\nu$ between probability measures μ and ν, we can first find a (binary) scoring function $s\colon \mathcal{X} \to \mathbb{R}$ that minimize the following CPE objective [Sugiyama et al., 2008, Menon and Ong, 2016, Nock et al., 2016]:

$$\operatorname*{arg\,min}_{s\colon \mathcal{X}\to\mathbb{R}} \quad \pi \cdot \mathbb{E}_\mu[(\lambda_{+1} \circ s)(\mathsf{X})] + (1-\pi) \cdot \mathbb{E}_\nu[(\lambda_{-1} \circ s)(\mathsf{X})], \tag{10}$$

where $\pi \in (0,1)$ and denote the partial losses $\lambda_{+1}(v) \stackrel{\text{def}}{=} \lambda(v,+1)$ and $\lambda_{-1}(v) \stackrel{\text{def}}{=} \lambda(v,-1)$.

The optimization of Eq. (10) is equivalent to a binary classification task on a dataset generated by class-conditionals defined by μ and ν—*i.e.*, $\mathbb{P}(\mathsf{X} = x \mid \mathsf{R} = +1) = \mu(x)$—and base label rate $\pi = \mathbb{P}(\mathsf{R} = +1)$. The label rate π is a free parameter which the user can set, where a typical choice is to set π based on the ratio of empirical data used to optimize Eq. (10). Key to this interpretation of the optimization problem is that we obtain class-probability estimates via the inverse link function $\hat{\mathbb{P}}(\mathsf{R} = +1 \mid \mathsf{X} = x) = (\Phi^{-1} \circ s)(x)$. As a result, the scorer s can then be converted to a density-ratio estimator between μ and ν via Bayes' rule [Bickel et al., 2009, Menon and Ong, 2016]

$$\hat{\rho}(x; s) \stackrel{\text{def}}{=} \frac{1-\pi}{\pi} \cdot \frac{(\Phi^{-1} \circ s)(x)}{1 - (\Phi^{-1} \circ s)(x)} \approx \frac{\mathrm{d}\mu}{\mathrm{d}\nu}. \tag{11}$$

If our pair of loss function and link (λ, Φ) is indeed strictly proper composite, then the minimizer of Eq. (10) will result in a scorer $s^\star$ corresponding to the Bayes-optimal class-posterior. As a result, the density-ratio learned will be exactly $\mathrm{d}\mu/\mathrm{d}\nu$. See Appendix D for an extended exposition of the DRE to CPE learning problem reduction.

Typically, the DRE to CPE reduction introduces the invertible link function to transform a scorer s into an approximate density-ratio. However, as we are interested in deferral, we only need to calculate the threshold of the approximated density-ratio. It turns out that the thresholding of density-ratios of the form Eq. (11) can be simplified to a thresholding of the scorer.

Lemma 4.1. *Let $s\colon \mathcal{X} \to \mathbb{R}$ be a scorer learned via the minimization of Eq. (10) with loss λ and strictly increasing, invertible link Φ. Then for $\hat{\rho}$ given by Eq. (11), we have for $\tau > 0$ that*

$$[\![\hat{\rho}(x) \leq \tau]\!] = \left[\!\!\left[s(x) \leq \Phi\left(\frac{\pi\tau}{1-\pi+\pi\tau}\right) \right]\!\!\right]. \tag{12}$$

Although it appears that the deferral decision is still dependent on the link function Φ on the RHS of Lemma 4.1, in practice one does not typically calculate an exact thresholding value. Instead, the scorer s can be evaluated over a validation dataset and the threshold value can be specified w.r.t. a target deferral percentage—a common approach for deferring via with scoring functions or confidence estimates [Pugnana and Ruggieri, 2023a,b]. One should note that even the "original" formulation of the L2D problem in Eq. (1) requires a calibration of the cost c—we remind that the deferral regularized $c \cdot \mathbb{P}_{\mathsf{P}_{\mathsf{X}}}[r(\mathsf{X}) = 1]$ can be interpreted as a Lagrange multiplier relaxation of a hard deferral rate constraint $\mathbb{P}_{\mathsf{P}_{\mathsf{X}}}[r(\mathsf{X}) = 1] \leq \texttt{budget}$ [Franc et al., 2023]. In the worst case, this may require training multiple deferral mechanisms r from scratch for each c in a cross validation scheme. As a result, Lemma 4.1 allows us to make deferral decision based on density-ratios by thresholding in the scorer's output space, rather than explicitly working in the space of density-ratio values.

The Ideal DR CPE Loss With the general setup of the DRE to CPE reduction, we now define a pair of density-ratio loss functions which correspond to the optimization problem of learning the density-ratio of ideal distributions. In particular, we exploit the fact that the introduced DRE framework of Eq. (10) allows for density-ratio estimates for any $\pi \in (0, 1)$.

Definition 4.2. *Given a DR loss λ, the corresponding $(\mathrm{i} \in \{\mathrm{m}, \mathrm{j}\})$-ideal DR CPE loss is given by*

$$\mathcal{L}_{\mathrm{i}}(s; \gamma, \gamma^{(\mathrm{e})}) \stackrel{\text{def}}{=} \mathbb{E}_{\mathsf{P}_{\mathsf{X}}}\left[\tilde{w}_{\mathrm{i}}(\mathsf{X}; \gamma) \cdot (\lambda_{+1} \circ s)(\mathsf{X}) + \tilde{w}_{\mathrm{i}}^{(\mathrm{e})}(\mathsf{X}; \gamma^{(\mathrm{e})}) \cdot (\lambda_{-1} \circ s)(\mathsf{X})\right]. \tag{13}$$

The i-ideal DR CPE loss in Definition 4.2 is simply derived by setting $\mu = \mathrm{Q}_{\mathrm{i}}(\cdot; \gamma)$, $\nu = \mathrm{Q}_{\mathrm{i}}^{(\mathrm{e})}(\cdot; \gamma^{(\mathrm{e})})$, and $\pi = Z_{\mathrm{i},\gamma}/(Z_{\mathrm{i},\gamma} + Z_{\mathrm{i},\gamma}^{(\mathrm{e})})$. The latter choice of setting the marginal label probabilities π conveniently eliminates the need to estimate the normalizers of the ideal distributions.

Although Definition 4.2 defines a pair of loss functions—one for the marginal ideal distributions, one for the joint ideal distributions—practically only the joint variant can be utilized. This is due to the dependence of $\mathbb{E}_{\eta^\star(x)}$ inside the exponential function of the weights $\tilde{w}_{\mathrm{m}}(x; \gamma) = \exp\left(-\mathbb{E}_{\eta^\star(x)}\left[\ell(x, \mathsf{Y}, h(x))\right]/\gamma\right)$. Although one could attempt to empirically estimate the weights (and the corresponding loss $\mathcal{L}_{\mathrm{m}}$), the resulting estimator would be biased. Furthermore, the estimated loss would typically have a high variance as for most high dimensional input spaces $\mathcal{X}$, only a single class label y is available for each observation x.

Despite the marginal variant of the DR CPE loss being difficult to estimate, the joint DR CPE loss $\mathcal{L}_{\mathrm{j}}(s; \gamma, \gamma^{(\mathrm{e})})$ has a convenient form:

$$\mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y})\sim\mathsf{P}}\left[\exp\left(-\frac{\ell(\mathsf{X}, \mathsf{Y}, h(\mathsf{X}))}{\gamma}\right) \cdot (\lambda_{+1} \circ s)(\mathsf{X}) + \exp\left(-\frac{\ell(\mathsf{X}, \mathsf{Y}, h^{(\mathrm{e})}(\mathsf{X}))}{\gamma^{(\mathrm{e})}}\right) \cdot (\lambda_{-1} \circ s)(\mathsf{X})\right]. \tag{14}$$

Indeed, as the expectation $\mathbb{E}_{\eta^\star(x)}$ is on the outside of the exponential function, the objective $\mathcal{L}_{\mathrm{j}}$ can be calculated as an expectation of a random variable w.r.t. the joint data measure P. From an estimation perspective, one can empirically estimate the expectation in Eq. (14) to yield an unbiased estimator.

A natural question is how are these two objectives related. As the marginal ideal distributions (and thus the corresponding loss function) relate to the typical notion of optimal deferral (via Theorem 3.5), the objective we would like to minimize is the marginal DR CPE loss $\mathcal{L}_{\mathrm{m}}$. Although $\mathcal{L}_{\mathrm{m}}$ is not easy to estimate, through Jensen's inequality, one can directly show that the joint DR CPE loss $\mathcal{L}_{\mathrm{j}}$ directly upper bounds the marginal DR CPE loss $\mathcal{L}_{\mathrm{m}}$.

Lemma 4.3. *Suppose $\ell, \lambda_{+1}, \lambda_{-1} \geq 0$. Then $\mathcal{L}_{\mathrm{m}}(s; \gamma, \gamma^{(\mathrm{e})}) \leq \mathcal{L}_{\mathrm{j}}(s; \gamma, \gamma^{(\mathrm{e})})$.*

From Lemma 4.3, one can expect that minimizing the joint DR CPE loss will reduce the corresponding marginal DR CPE loss via an upper bound. However, additional assumptions are needed to establish bounds which explore the consistency of the joint DR CPE loss w.r.t. minimization of the *original* L2D risk (Eq. (1)). In particular, we consider *strongly proper* loss functions, which have previously been used to derive consistency bounds [Agarwal, 2014]. The full details of the assumption, formal statements, and all proofs associated with consistency are deferred to Appendices G and H.

Given the appropriate assumptions, a general approximate consistency-style bound can be established for the shared temperature case $\gamma = \gamma^{(\mathrm{e})} > 0$.

Theorem 4.4 (Informal)**.** *Suppose $0 \leq \ell \leq B$, (λ, Φ) is β-strongly proper, and $\lambda_{+1}, \lambda_{-1} \geq 0$. Define $r_{c,\gamma}(x) = [\![(\Phi^{-1} \circ s)(x) \leq \tau_{c,\gamma}]\!]$. Then there exists some constant $K_{\gamma,B,c} > 0$ such that every scorer*

s and cost $c \geq 0$ yields,

$$\mathcal{L}(h, r_{c,\gamma}) - \mathcal{L}(h, r^\star) \leq K_{\gamma,B,c} \sqrt{\frac{\exp(B/\gamma)}{\beta} \left[\mathcal{E}_{\mathrm{j}}(s;\gamma,\gamma) + \Delta_{\mathrm{mj}}(\gamma,\gamma)\right]}, \tag{15}$$

where $\tau_{c,\gamma} \stackrel{\text{def}}{=} \sigma\left(-\frac{c}{\gamma}\right)$ *with* $\sigma(t) \stackrel{\text{def}}{=} (1+\exp(-t))^{-1}$ *denoting the sigmoid function and*

$$\mathcal{E}_{\mathrm{i}}(s;\gamma,\gamma^{(\mathrm{e})}) \stackrel{\text{def}}{=} \mathcal{L}_{\mathrm{i}}(s;\gamma,\gamma^{(\mathrm{e})}) - \inf_{s'} \mathcal{L}_{\mathrm{i}}(s';\gamma,\gamma^{(\mathrm{e})}) \tag{16}$$

$$\Delta_{\mathrm{mj}}(\gamma,\gamma^{(\mathrm{e})}) \stackrel{\text{def}}{=} \inf_{s} \mathcal{L}_{\mathrm{j}}(s;\gamma,\gamma^{(\mathrm{e})}) - \inf_{s'} \mathcal{L}_{\mathrm{m}}(s';\gamma,\gamma^{(\mathrm{e})}) \tag{17}$$

denoting the excess risk and marginal-joint ideal optimality gap, respectively.

Theorem 4.4 bounds the original L2D regret using the joint excess risk and the marginal-joint ideal optimality gap. The marginal-joint ideal optimality gap term $\Delta_{\mathrm{mj}}(\gamma,\gamma)$ can be bounded by the Jensen gap in Lemma 4.3. This can be further shown to be upper bounded by the (weighted conditional) task loss variance of the base and expert models (Eq. (64)). This can be shown to vanish when the task losses are (almost surely) constant conditioned on an input $\mathsf{X} = x$, *e.g.*, when the labels have no aleatoric uncertainty [Hüllermeier and Waegeman, 2021]. In this case, Theorem 4.4 yields consistency. Similar observation can be made by observing that another sufficient condition for $\Delta_{\mathrm{mj}}(\gamma,\gamma) = 0$ is that the Chow's rule and the "tilted" Chow's rule of Theorem 3.5 to be equal.

Generally $\Delta_{\mathrm{mj}}(\gamma,\gamma)$ is non-zero and thus Theorem 4.4 does not generally prove consistency. This naturally raises the question of whether there exists a setting which would yield a consistent joint objective. The following reports the affirmative by deviating from the common temperature setting.

Theorem 4.5 (Informal)**.** *Suppose that ℓ is the zero-one loss function and the composite loss (λ, Φ) is β-strongly proper with a strictly increasing link Φ. For $c \geq 0$ and $\gamma^{(\mathrm{e})} > 0$, there exists a $\gamma > 0$, $\tilde{K}_{c,\gamma,\gamma^{(\mathrm{e})}} > 0$, and $\tilde{\tau}_{c,\gamma,\gamma^{(\mathrm{e})}}$ such that the rejector $\bar{r}_{c,\gamma,\gamma^{(\mathrm{e})}}(x) = [\![(\Phi^{-1} \circ s)(x) \leq \tilde{\tau}_{c,\gamma,\gamma^{(\mathrm{e})}}]\!]$ satisfies*

$$\mathcal{L}^{\mathrm{zo}}(h, \bar{r}_{c,\gamma,\gamma^{(\mathrm{e})}}) - \mathcal{L}^{\mathrm{zo}}(h, r^\star) \leq \frac{\tilde{K}_{c,\gamma,\gamma^{(\mathrm{e})}}}{\sqrt{\beta}} \sqrt{\mathcal{E}^{\mathrm{zo}}_{\mathrm{j}}(s;\gamma,\gamma^{(\mathrm{e})})}, \tag{18}$$

*where the superscript "*zo*" is to indicate we are considering the zero-one loss task loss $\ell = \ell^{\mathrm{zo}}$.*

For non-zero deferral cost $c \neq 0$, the temperatures in Theorem 4.5 are required to be different $\gamma \neq \gamma^{(\mathrm{e})}$. The bound is an interesting counterpart to Theorem 4.4, where by specifying the "task loss" ℓ to be the zero-one loss function, the bound does not depend on the marginal-joint ideal optimality gap. Importantly, this bound yields an exact scenario where the joint DR CPE losses can provide exact consistency. It should be noted that consistency results typically only hold for a fixed cost Cao et al. [2022], *i.e.*, consistency at one cost c' does not imply simultaneously consistency for all possible costs $c \geq 0$. In our Theorems 4.4 and 4.5, the consistency of the scorer s is also tied to a specific threshold τ (which is depends on c).

Expert-Comparison Estimation: Previous work in the literature suggests learning deferral rules by regressing pointwise targets that compare model and expert performance [Kag et al., 2022, Jitkrittum et al., 2023]. In particular, one seeks to estimate a pointwise expert-comparison target that approximates how Chow's rule makes deferral decisions. For example, one considers the zero-one loss function to design a least-squares regression task on the random variable

$$\mathsf{Y}^{(\mathrm{d})} \stackrel{\text{def}}{=} [\![f(\mathsf{X}) = \mathsf{Y}]\!] - [\![f^{(\mathrm{e})}(\mathsf{X}) = \mathsf{Y}]\!]. \tag{19}$$

These target labels $\mathsf{Y}^{(\mathrm{d})}$ simply aim to identify whenever the expert $h^{(\mathrm{e})}$ would be correct but the model h would not. Interestingly, a direct connection can be made to the joint ideal DR CPE loss.

Proposition 4.6. *Suppose ℓ is the zero-one loss function and λ is the squared loss function. Let $\Sigma(x) \stackrel{\text{def}}{=} \mathbb{P}_{\mathrm{P}}[f(\mathsf{X}) = \mathsf{Y} \mid \mathsf{X} = x] + \mathbb{P}_{\mathrm{P}}[f^{(\mathrm{e})}(\mathsf{X}) = \mathsf{Y} \mid \mathsf{X} = x]$. Then, for P_{x}-almost every x with $\Sigma(x) > 0$, the unique pointwise minimizer $s^\star_0(x)$ of the limiting risk satisfies, for any $\tau \in \mathbb{R}$,*

$$[\![s^\star_0(x) \leq \tau]\!] = \left[\!\!\left[\mathbb{E}[\mathsf{Y}^{(\mathrm{d})} \mid \mathsf{X} = x] \leq \tau\, \Sigma(x)\right]\!\!\right]. \tag{20}$$

An interesting interpretation of Proposition 4.6 is that, in the low temperature limit $\gamma \to 0$, the optimal deferral mechanism induced by the joint ideal loss with squared loss is equivalent to the optimal

deferral mechanism learned by regressing on $\mathsf{Y}^{(\mathrm{d})}$. The only difference is that the threshold is no longer constant, but instead rescaled by the input-dependent factor $\Sigma(x)$. Thus, the rule preserves the same expert-comparison "signal" used in prior work, with an input-dependent threshold determined by the total correctness of the model and expert.

Anomaly Detection (AD): One definition of anomalous observations are examples which are not concentrated [Ripley, 2007, Chandola et al., 2009], *i.e.*, examples that are sampled from areas of low density w.r.t. P. To identify such examples, one approach in AD seeks to solve an objective similar to DRE Eq. (10). More concretely, they seek to learn a scorer s via optimizing [Steinwart et al., 2005]:

$$\mathcal{L}_{\text{anomaly}}(s) \stackrel{\text{def}}{=} \mathbb{E}_{\mathsf{P}}[(\lambda_{+1} \circ s)(\mathsf{X})] + \int (\lambda_{-1} \circ s)(x) \, \mathrm{d}\mu(x), \tag{21}$$

where μ corresponds to the reference measure of P. An observation x can then be classified as anomalous via the scorer $s(x)$, *i.e.*, via a threshold $[\![s(x) \leq \alpha_q]\!]$ with α_q chosen to correspond to the qth quantile of the scorer s. A calibrated variant of the AD task can also be considered, where instead of just wanting to identify anomalous observations, we want the learned scorer s (via a link function) to correspond to the cumulative distribution function of P [Menon and Williamson, 2018].

This form of AD mirrors that of learning deferral rules from ideal distributions—the only change is what is considered a typical versus anomalous distribution. Here, the typical distribution is considered as the data distribution itself P and anything which deviates from it is considered anomalous, *i.e.*, the reference measure μ. Instead, in L2D both the typical and anomalous distributions are reweightings of the data distribution. From an optimization perspective, our deferral task is easier. In Eq. (21), the integral term is difficult to estimate (for general μ) [Menon and Williamson, 2018].

5 Experiments

In the following section, we evaluate our ideal DR CPE method learning deferral decisions.[1] To do so, we consider a variant of the experimental model cascade setup presented in Jitkrittum et al. [2023] on corrupted CIFAR-100 [Krizhevsky, 2009] and MedMNIST [Yang et al., 2021, 2023] datasets. We consider label noise, long-tail, and specialization variants of these datasets: label noise randomly flips the labels of the first 10 classes; long-tail subsamples 500 examples from the first 25 classes and 50 examples from the rest; and specialization subsamples examples at a rate of 20% for non-specialist class labels. We consider a three-way partition of the dataset: a training set for h, $h^{(\mathrm{e})}$; a clean held-out training set for r; and a clean test set for evaluation. Label noise and long-tail subsampling affect both classifiers. In the specialist setting, h uses the full clean training partition and only $h^{(\mathrm{e})}$ subsamples non-specialist examples. When a validation split is available, we use it to train the deferral mechanism; otherwise, we use a partition of the training set. The model architectures are ResNet models of various sizes $\{8, 14, 32, 56\}$ ($h^{(\mathrm{e})}$ typically larger) and (when applicable) deferral approaches utilize a lightweight Multi-Layer Perceptron with features constructed from probability outputs from h [Jitkrittum et al., 2023]. Further details and additional experimental settings are presented in Appendices J and K. We further present ImageNet [Russakovsky et al., 2015] experiments and unequal temperatures experiments in Appendix L (for simplicity, we only consider the equal temperature settings $\gamma = \gamma^{(\mathrm{e})}$ here).

For comparison, we consider 3 types of baselines. First we consider a confidence-based approach (`Conf`), which simply defers based on thresholding h's perceived confidence (no deferral training) [Ramaswamy et al., 2018]. Second, we consider a two-stage surrogate loss method (`TwoStage-Exp`) [Mao et al., 2023]. Finally, we consider two expert-comparison estimations [Kag et al., 2022, Jitkrittum et al., 2023]: one which regresses on $\mathsf{Y}^{(d)}$ as per Eq. (19) (`Estimate-Diff01`); and one which uses the relative confidence between h and $h^{(\mathrm{e})}$ by estimating $h^{(\mathrm{e})}$'s maximum predicted probability (`Estimate-MaxProb`). For DR CPE, we set λ to be the squared loss and the temperatures to be $\gamma = \gamma^{(\mathrm{e})} = 0.5$; we consider ℓ to be either the generalized cross-entropy (GCE) [Zhang and Sabuncu, 2018] or a probabilistic version of the zero-one loss which consists of the L_2 distance between the one-hot label and the predicted probabilities (Prob01).

Fig. 2 summarizes our experiments, where we evaluate the accuracy-deferral trade-offs (upper bounds are better). At 0% deferral rate (left) the accuracy-deferral curve reports the base model h accuracy;

[1] Code is available at: https://github.com/alexandersoen/dr-losses-for-learning-to-defer.

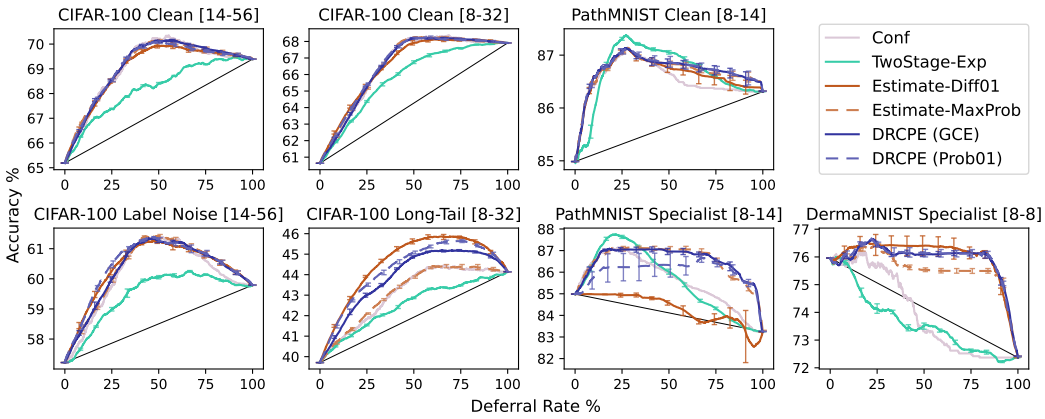

Figure 2: Accuracy-deferral trade-off curves over various experimental setups. The base model corresponds to 0% deferral rate and the expert model corresponds to 100% deferral rate. Curves which produce upper bounds are better. Titles detail the dataset and ResNet utilized, *i.e.*, "`[14-56]`" $\mapsto$ $h/h^{(\mathrm{e})}$ = ResNet14/56. Curves depict the mean and error bars depict ± 1 s.t.d range (over 11 runs).

and at 100% deferral rate (right) its reports the expert model $h^{(\mathrm{e})}$ accuracy. Accuracy-deferral curves evaluate the approaches beyond their consistency guarantees by evaluating performance across different (deferral cost) thresholds. Each curve sweeps the threshold of a fixed scorer; changing γ requires retraining. Although sweeping the threshold in this manner does not adhere to the fixed deferral cost (and threshold) setting that our consistency results require (Theorems 4.4 and 4.5), we find that DR CPE performs robustly across different deferral rates regardless. The approaches perform similarly under clean data. Over corrupted settings, DR CPE performs best to second best. When compared to the confidence-based and the two-stage approaches, DR CPE is consistently better. Both of these baselines perform extremely badly on the DermaMNIST specialist setting. The exception is that the two-stage approach performs well clean PathMNIST and corrupted PathMNIST for smaller deferral rates (less than $\sim 30\%$); however, for higher deferral rates it performs significantly worse than DR CPE (GCE). The most competitive baselines are the expert-comparison estimators. `Estimate-MaxProb` performs significantly worse than DR CPE in corrupted datasets, except the PathMNIST specialist setting where it performs similarly to DR CPE with GCE. `Estimate-Diff01` has strong performance, beating DR CPE for the CIFAR-100 long-tail and DermaMNIST specialist settings, but suffers from a larger variance in the latter. However, `Estimate-Diff01` critically fails in the PathMNIST specialist setting—which could be attributed to its tendency to overfit [Jitkrittum et al., 2023]. DR CPE with Prob01 also has a loss in performance (and a larger variance) in this setting despite outperforming its GCE counterpart in other settings.

6 Conclusion

We present a novel method for learning deferral mechanisms based on ideal distributions and the link between DRE and CPE. Practically, we find that the DR CPE loss function we propose can yield good accuracy-deferral trade-offs while also being robust in its performance over a variety of dataset and corruption combinations. One avenue of future work would be to extend the initial consistency results established in Theorems 4.4 and 4.5, *e.g.*, (exact) consistency for task losses ℓ beyond the zero-one loss. One could attempt to achieve this by deviating away from the KL-divergence based ideal distributions and instead leverage other φ-divergences [Soen et al., 2024]. Ideally, we want to seek a consistency result that could also hold for multiple deferral costs $c \geq$ simultaneously. So far, the narrative has been for L2D with a pair of models. Of course one could easily chain deferral mechanism, as per model cascades [Wang et al., 2022, Jitkrittum et al., 2023]; however, ideally we would want to directly deal with multi-expert deferral, where there may be multiple base models and/or experts [Verma et al., 2023, Mao et al., 2023, 2024b, 2025, Liu et al., 2026]. An obvious connection here could be to exploit the connection between multi-distributional DRE and CPE in the multi-class setting [Yu et al., 2021, Nock et al., 2016]. Another important direction is to modify our ideal distribution framework to allow additional constraints for, *e.g.*, fairness or limiting expert calls in training [Narasimhan et al., 2024, Charusaie and Samadi, 2024, DeSalvo et al., 2025].

Acknowledgments

The authors would like to acknowledge the NeurIPS 2026 reviews for providing helpful advice, particularly on the consistency of our approach. The authors would like to thank Aditya Krishna Menon, Wittawat Jitkrittum, and Harikrishna Narasimhan for various fruitful discussions.

Computational resources were provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), funded by the Swedish Research Council. This work was funded, in part, by the Swedish Research Council under contracts 2023-05195 (Alexander Soen and Joakim Jaldén) and 2021-05266 (Ragnar Thobaben). This work was also funded by KTH Digital Futures (Alexander Soen).

References

Shivani Agarwal. Surrogate regret bounds for bipartite ranking via strongly proper losses. *The Journal of Machine Learning Research*, 15(1):1653–1674, 2014.

Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*, volume 191. American Mathematical Soc., 2000.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of machine learning research*, 3(Nov):463–482, 2002.

Peter L Bartlett and Marten H Wegkamp. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9(8), 2008.

Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10(9), 2009.

Andreas Buja, Werner Stuetzle, and Yi Shen. Loss functions for binary class probability estimation and classification: Structure and applications. *Working draft, November*, 3:13, 2005.

Jenna Burrell. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big data & society*, 3(1):2053951715622512, 2016.

Yuzhou Cao, Tianchi Cai, Lei Feng, Lihong Gu, Jinjie Gu, Bo An, Gang Niu, and Masashi Sugiyama. Generalizing consistent multi-class classification with rejection to be compatible with arbitrary losses. *Advances in neural information processing systems*, 35:521–534, 2022.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.

Nontawat Charoenphakdee, Zhenghang Cui, Yivan Zhang, and Masashi Sugiyama. Classification with rejection based on cost-sensitive classification. In *International Conference on Machine Learning*, pages 1507–1517. PMLR, 2021.

Mohammad-Amin Charusaie and Samira Samadi. A unifying post-processing framework for multi-objective learn-to-defer problems. *Advances in Neural Information Processing Systems*, 37: 23705–23755, 2024.

C. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16:41–46, 1970. doi: 10.1109/TIT.1970.1054406.

Chi-Keung Chow. An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers*, EC-6(4):247–254, 1957. doi: 10.1109/TEC.1957.5222035.

Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Learning with rejection. In *International conference on algorithmic learning theory*, pages 67–82. Springer, 2016.

Giulia DeSalvo, Clara Mohri, Mehryar Mohri, and Yutao Zhong. Budgeted multiple-expert deferral. *arXiv preprint arXiv:2510.26706*, 2025.

David Dohan, Winnie Xu, Aitor Lewkowycz, Jacob Austin, David Bieber, Raphael Gontijo Lopes, Yuhuai Wu, Henryk Michalewski, Rif A. Saurous, Jascha Sohl-dickstein, Kevin Murphy, and Charles Sutton. Language model cascades. *arXiv preprint arXiv:2207.10342*, 2022.

John C Duchi, Peter W Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 46(3):946–969, 2021.

Krishnamurthy Dj Dvijotham, Jamie Hayes, Borja Balle, Zico Kolter, Chongli Qin, Andras Gyorgy, Kai Xiao, Sven Gowal, and Pushmeet Kohli. A framework for robustness certification of smoothed classifiers using f-divergences. In *International Conference on Learning Representations*, 2020.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53), 2010.

F Escher. On the probability function in the collective theory of risk. *Skand. Aktuarie Tidskr.*, 15: 175–195, 1932.

Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

Vojtech Franc, Daniel Prusa, and Vaclav Voracek. Optimal strategies for reject option classifiers. *Journal of Machine Learning Research*, 24(11):1–49, 2023.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *International conference on machine learning*, pages 2151–2159. PMLR, 2019.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *Conference on learning theory*, pages 297–299. PMLR, 2018.

Neha Gupta, Harikrishna Narasimhan, Wittawat Jitkrittum, Ankit Singh Rawat, Aditya Krishna Menon, and Sanjiv Kumar. Language model cascades: Token-level uncertainty and beyond. In *The Twelfth International Conference on Learning Representations*, 2024.

Radu Herbei and Marten H Wegkamp. Classification with reject option. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, pages 709–721, 2006.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506, 2021.

Wittawat Jitkrittum, Neha Gupta, Aditya K Menon, Harikrishna Narasimhan, Ankit Rawat, and Sanjiv Kumar. When does confidence-based cascade deferral suffice? *Advances in Neural Information Processing Systems*, 36:9891–9906, 2023.

Mohammad Ali Kadampur and Sulaiman Al Riyaee. Skin cancer detection: Applying a deep learning based model driven architecture in the cloud for classifying dermal cell images. *Informatics in Medicine Unlocked*, 18:100282, 2020.

Anil Kag, Igor Fedorov, Aditya Gangrade, Paul Whatmough, and Venkatesh Saligrama. Efficient edge inference by selective query. In *The Eleventh International Conference on Learning Representations*, 2022.

Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *The Journal of Machine Learning Research*, 10:1391–1445, 2009.

Leila Khalili, Yao You, and John Bohannon. BabyBear: Cheap inference triage for expensive language models. *arXiv preprint arXiv:2205.11747*, 2022.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Jeremias Knoblauch, Jack Jewson, and Theodoros Damoulas. An optimization-centric view on bayes' rule: Reviewing and generalizing variational inference. *Journal of Machine Learning Research*, 23 (132):1–109, 2022.

Donald E Knuth. Two notes on notation. *The American Mathematical Monthly*, 99:403–422, 1992.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

E. L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer International Publishing, 2005.

Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. Large language models in finance: A survey. In *Proceedings of the fourth ACM international conference on AI in finance*, pages 374–382, 2023.

Shuang Liu and Kamalika Chaudhuri. The inductive bias of restricted f-gans. *arXiv preprint arXiv:1809.04542*, 2018.

Shuqi Liu, Yuzhou Cao, Lei Feng, Bo An, and Luke Ong. When more experts hurt: Underfitting in multi-expert learning to defer. In *Forty-third International Conference on Machine Learning*, 2026.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature communications*, 15(1):654, 2024.

David Madras, Toni Pitassi, and Richard Zemel. Predict responsibly: improving fairness and accuracy by learning to defer. *Advances in neural information processing systems*, 31, 2018.

Jonathan Mamou, Oren Pereg, Moshe Wasserblat, and Roy Schwartz. TangoBERT: Reducing inference cost by using cascaded architecture. *arXiv preprint arXiv:2204.06271*, 2022.

Anqi Mao, Christopher Mohri, Mehryar Mohri, and Yutao Zhong. Two-stage learning to defer with multiple experts. *Advances in Neural Information Processing Systems*, 36:3578–3606, 2023.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Predictor-rejector multi-class abstention: Theoretical analysis and algorithms. In *International Conference on Algorithmic Learning Theory*, pages 822–867. PMLR, 2024a.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Principled approaches for learning to defer with multiple experts. In *International Workshop on Combinatorial Image Analysis*, pages 107–135. Springer, 2024b.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Mastering multiple-expert routing: Realizable H-consistency and strong guarantees for learning to defer. In *Forty-second International Conference on Machine Learning*, 2025.

Aditya Krishna Menon and Cheng Soon Ong. Linking losses for density ratio and class-probability estimation. In *International Conference on Machine Learning*, pages 304–313, 2016.

Aditya Krishna Menon and Robert C Williamson. A loss framework for calibrated anomaly detection. In *Proceedings of the 32nd international conference on neural information processing systems*, pages 1494–1504, 2018.

Pierre Del Moral. *Feynman-Kac Formulae*. Springer, 2004.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *International conference on machine learning*, pages 7076–7087. PMLR, 2020.

Hussein Mozannar, Hunter Lang, Dennis Wei, Prasanna Sattigeri, Subhro Das, and David Sontag. Who should predict? exact algorithms for learning to defer to humans. In *International conference on artificial intelligence and statistics*, pages 10520–10545. PMLR, 2023.

Harikrishna Narasimhan, Aditya Krishna Menon, Wittawat Jitkrittum, Neha Gupta, and Sanjiv Kumar. Learning to reject meets long-tail learning. In *The Twelfth International Conference on Learning Representations*, 2024.

Harikrishna Narasimhan, Wittawat Jitkrittum, Ankit Singh Rawat, Seungyeon Kim, Neha Gupta, Aditya Krishna Menon, and Sanjiv Kumar. Faster cascades via speculative decoding. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jerzy Neyman and Egon S. Pearson. Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London Series A Containing Papers of a Mathematical or Physical Character*, 231:289–337, 1933. doi: 10.1098/rsta.1933.0009.

Chenri Ni, Nontawat Charoenphakdee, Junya Honda, and Masashi Sugiyama. On the calibration of multiclass classification with rejection. *Advances in neural information processing systems*, 32, 2019.

Richard Nock, Aditya Menon, and Cheng Soon Ong. A scaled bregman theorem with applications. *Advances in Neural Information Processing Systems*, 29, 2016.

Nastaran Okati, Abir De, and Manuel Gomez-Rodriguez. Differentiable learning under triage. *Advances in Neural Information Processing Systems*, 34:9140–9151, 2021.

Antoine Picard-Weibel and Benjamin Guedj. On change of measure inequalities for f-divergences. *arXiv preprint arXiv:2202.05568*, 2022.

Andrea Pugnana and Salvatore Ruggieri. AUC-based selective classification. In *International conference on artificial intelligence and statistics*, pages 2494–2514. PMLR, 2023a.

Andrea Pugnana and Salvatore Ruggieri. A model-agnostic heuristics for selective classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 9461–9469, 2023b.

Harish G Ramaswamy, Ambuj Tewari, and Shivani Agarwal. Consistent algorithms for multiclass classification with an abstain option. *Electronic Journal of Statistics*, 12:530–554, 2018.

Mark D. Reid and Robert C. Williamson. Composite binary losses. *Journal of Machine Learning Research*, 11:2387–2422, 2010.

Mark D Reid and Robert C Williamson. Information, divergence and risk for binary experiments. *Journal of Machine Learning Research*, 12:731–817, 2011.

Brian D Ripley. *Pattern recognition and neural networks*. Cambridge university press, 2007.

R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

Vincent Roulet, Tianlin Liu, Nino Vieillard, Michael Eli Sander, and Mathieu Blondel. Loss functions and operators generated by f-divergences. In *Forty-second International Conference on Machine Learning*, 2025.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3): 211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted residuals and linear bottlenecks. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pages 4510–4520. Ieee, 2018.

Herbert E Scarf, KJ Arrow, and S Karlin. A min-max solution of an inventory problem. Technical report, Rand Corporation Santa Monica, 1957.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature medicine*, 31(3):943–950, 2025.

Alexander Soen. A connection between learning to reject and bhattacharyya divergences. In *Geometric Science of Information*, pages 369–377. Springer Nature Switzerland, 2026. doi: 10.1007/978-3-032-03918-7_38.

Alexander Soen, Hisham Husain, Philip Schulz, and Vu Nguyen. Rejection via learning density ratios. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Ingo Steinwart, Don Hush, and Clint Scovel. A classification framework for anomaly detection. *Journal of Machine Learning Research*, 6(2), 2005.

Masashi Sugiyama, Taiji Suzuki, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60:699–746, 2008. doi: 10.1007/s10463-008-0197-x.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PmLR, 2019.

Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56, 2019.

Neeraj Varshney and Chitta Baral. Model cascading: Towards jointly improving efficiency and accuracy of NLP systems. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11007–11021, 2022.

Rajeev Verma and Eric Nalisnick. Calibrated learning to defer with one-vs-all classifiers. In *International Conference on Machine Learning*, pages 22184–22202. PMLR, 2022.

Rajeev Verma, Daniel Barrejón, and Eric Nalisnick. Learning to defer to multiple experts: Consistent surrogate losses, confidence calibration, and conformal ensembles. In *International Conference on Artificial Intelligence and Statistics*, pages 11415–11434. PMLR, 2023.

P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I, 2001. doi: 10.1109/CVPR.2001.990517.

Xiaofang Wang, Dan Kondratyuk, Eric Christiansen, Kris M. Kitani, Yair Movshovitz-Attias, and Elad Eban. Wisdom of committees: An overlooked approach to faster and more accurate models. In *International Conference on Learning Representations*, 2022.

Jiancheng Yang, Rui Shi, and Bingbing Ni. Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis. In *IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 191–195, 2021.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. MedMNIST v2 - A large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

Niloofar Yousefi, Marie Alaghband, and Ivan Garibay. A comprehensive survey on machine learning techniques and user authentication approaches for credit card fraud detection. *arXiv preprint arXiv:1912.02629*, 2019.

Lantao Yu, Yujia Jin, and Stefano Ermon. A unified framework for multi-distribution density ratio estimation. *arXiv preprint arXiv:2112.03440*, 2021.

Ming Yuan and Marten Wegkamp. Classification methods with reject option based on convex risk minimization. *Journal of Machine Learning Research*, 11(1), 2010.

Arnold Zellner. Optimal information processing and bayes's theorem. *The American Statistician*, 42 (4):278–280, 1988.

Jingzhao Zhang, Aditya Krishna Menon, Andreas Veit, Srinadh Bhojanapalli, Sanjiv Kumar, and Suvrit Sra. Coping with label shift via distributionally robust optimisation. In *International Conference on Learning Representations*, 2021.

Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018.

Yazeed Zoabi, Shira Deri-Rozov, and Noam Shomron. Machine learning-based prediction of covid-19 diagnosis based on symptoms. *npj digital medicine*, 4(1):3, 2021.

Appendix Contents

A Limitations **19**

B Broader Impact **19**

C Ideal Distribution with φ-Divergences **19**

C.1 The Result 19

C.2 Helper Lemmas 21

C.3 Proof of Result 22

D DRE to CPE Reduction **24**

D.1 Reduction 24

D.2 Example 25

E Connection to Chow's Rule **25**

E.1 Proof of Theorem 3.5 25

F Joint Ideal DR CPE Loss Properties **26**

F.1 Proof of Lemma 4.3 26

F.2 Variance upper bound of Jensen's gap 26

G Approximate Consistency of DR CPE Losses **27**

G.1 Assumptions and Definitions 27

G.2 Helper Lemma 29

G.3 Proof of Theorem 4.4 (Approximate Consistency Bound) 29

G.4 Variance and Finite-Sample Bounds 31

H Exact Consistency of DR CPE Losses **33**

H.1 Choice of loss functions 33

H.2 Mapping between temperature and deferral costs 33

H.3 Proof of Theorem 4.5 (Exact Consistency Bound) 35

I Connections to Expert-Comparison Estimation **36**

I.1 Proof of Proposition 4.6 38

I.2 Ratio Expert Comparison 38

J Experiment Settings **40**

J.1 Base Datasets 40

J.2 Corruptions 40

J.3 Parameterized Deferral Models 41

J.4 Optimizers 41

J.5 Evaluation 42

J.6 Baselines 42

K Extended Experiments **43**

L ImageNet Experiments **48**

L.1 Unequal Temperatures 48

L.2 Non-Dog Subgroup Evaluation 51

A Limitations

The general bound in Theorem 4.4 only provides "approximate" consistency: the upper bound still depends on a potential non-zero marginal-joint optimality gap. To achieve an "exact" consistency counterpart, one further needs to assume that the task loss ℓ is the zero-one loss function. The consistency results of our paper Theorems 4.4 and 4.5 only hold for a specific deferral cost $c \geq 0$ (and scorer threshold τ). Despite this our practical results show promising results when one does not "commit" to a deferral costs c (and τ) at train time. Nevertheless, one should be careful about adapting our scorer to different thresholds at inference time. In our experiments, we have also assumed that the deferral mechanism r (and its scorer s) are trained using data which is i.i.d. to the test set at inference time. Different data settings which has distribution shift could cause a non-trivial drop in performance.

B Broader Impact

L2D inherently has impact in various downstream tasks, *e.g.*, in health. This can cause potential issues in scenarios of *overly conservative deferral*, *i.e.*, the overall system does not defer to a human expert when the base model is unreliable. As such, in practice, one should ensure that the deferral rates τ selected are validated to ensure that unreliable base model predictions are not utilized at a high rate. On the flip side, a well-calibrated L2D model has the potential of reducing the burden on a human expert. Other impacts our and other L2D approaches could have is on the fairness and privacy of the overarching deferring system (when compared to just the base model). Although, L2D has been previously motivated to improve fairness [Madras et al., 2018], our approach does not have any explicit constraints on any fairness metrics. Furthermore, deferral could impact the privacy of users, *e.g.*, in a loan prediction system, patterns of deferral could leak certain demographic information. These aspects should be considered when utilizing our method.

C Ideal Distribution with φ-Divergences

In the following section, we present a generalization of the optimal ideal distributions considered in Corollary 3.3. Instead of considering the KL-divergence, we will pick the family of φ-divergences (a.k.a., Csiszár f-divergences) as our choice of dissimilarity D in Definition 3.1. We first recall the definition of an φ-divergence.

Definition C.1 (φ-Divergence)**.** *Given a convex lower semi-continuous generator $\varphi\colon \mathbb{R} \to (-\infty, \infty]$ that satisfies $\varphi(1) = 0$, the corresponding φ-divergence between non-negative measures $\mathrm{P} \ll \mathrm{Q}$ is defined by*

$$\mathrm{D}_\varphi(\mathrm{P} \,\|\, \mathrm{Q}) = \int_{\mathcal{Z}} \varphi\left(\frac{\mathrm{dP}}{\mathrm{dQ}}(z)\right) \mathrm{dQ}(z). \tag{22}$$

We will be looking at the generic problem of solving:

$$\begin{aligned} \min_{\rho\colon \mathcal{Z} \to \mathbb{R}_{\geq 0}} \quad & \mathbb{E}_{\mathrm{P}}\left[\rho(\mathrm{Z}) L(\mathrm{Z})\right] + \gamma \cdot \mathrm{D}_\varphi(\rho \cdot \mathrm{P} \,\|\, \mathrm{P}), \\ s.t. \quad & \mathbb{E}_{\mathrm{P}}[\rho(\mathrm{Z})] = 1. \end{aligned} \tag{23}$$

where $L\colon \mathbb{Z} \to \mathbb{R}$ and $\gamma > 0$.

Eq. (23) accounts for both marginal and joint variants of the optimization problems in Definition 3.1. The only difference is that instead of optimizing for a distribution, we optimize for the density-ratio directly (similar to Dvijotham et al. [2020]). The corresponding ideal distribution can be recovered via $\mathrm{Q}(z) = \rho(z)\mathrm{P}(z)$, *i.e.*, ρ is the reweighting. Solutions to such an equation under different assumptions on φ are well-known and have been explored in various literatures [Knoblauch et al., 2022, Picard-Weibel and Guedj, 2022, Soen et al., 2024, Roulet et al., 2025]. In what follows, we present a novel (to the best of our knowledge) solution to Eq. (23). This will immediately give the proof of Corollary 3.3 via an instantiation of $\mathrm{D}_\varphi = \mathrm{KL}$. The approach we utilize is heavily inspired by Picard-Weibel and Guedj [2022].

C.1 The Result

We first present the result and the necessary definitions required to state it.

First, we will only consider a subset of possible generators $\varphi\colon \mathbb{R} \to (-\infty, +\infty]$. Formally, we consider the set

$$\Phi \stackrel{\text{def}}{=} \{\varphi\colon \mathbb{R} \to (-\infty, +\infty] \mid \varphi \text{ is proper convex, lower semi-continuous}, \varphi(1) = 0, \varphi(\mathbb{R}_{\geq 0}) \subset \mathbb{R}\}. \tag{24}$$

Φ defines the class of usual φ-generators with an additional constraint that $\varphi(\mathbb{R}_{\geq 0}) \subset \mathbb{R}$. This ensures that φ is subdifferentiable on $\mathbb{R}_{>0}$ and that the right derivative $\varphi'_+(0)$ is well-defined in $[-\infty, +\infty)$.

Definition C.2 (Right Derivative, Rockafellar [1970])**.** *Let $\varphi\colon \mathbb{R} \to [-\infty, +\infty]$ be any function. Let x be a point where φ is finite. Then given that the following limit exists, the right derivative is defined as,*

$$\varphi'_+(x) \stackrel{\text{def}}{=} \lim_{h \downarrow 0} \frac{\varphi(x+h) - \varphi(x)}{h}. \tag{25}$$

We also consider the standard convex conjugate.

Definition C.3 (Convex Conjugate, Rockafellar [1970])**.** *Let $\varphi\colon \mathbb{R} \to [-\infty, +\infty]$ be any function. The convex-conjugate (Fenchel conjugate) is defined as,*

$$\varphi^*(p) \stackrel{\text{def}}{=} \sup_{x \in \mathbb{R}} \{xp - \varphi(x)\}. \tag{26}$$

We also consider a refinement of the generator class Φ:

$$\Phi_\uparrow \stackrel{\text{def}}{=} \{\varphi\colon \mathbb{R} \to (-\infty, +\infty] \mid \varphi \in \Phi, \varphi(\mathbb{R}_{<0}) = \{+\infty\}\}. \tag{27}$$

Intuitively, $\Phi_\uparrow$ is just the class of generators where negative inputs are set to ∞. Indeed, we can define a transform from any generator $\varphi \in \Phi$ to $\Phi_\uparrow$ via

$$\varphi_\uparrow(x) \stackrel{\text{def}}{=} \begin{cases} \varphi(x) & \text{if } x \geq 0 \\ +\infty & \text{otherwise.} \end{cases} \tag{28}$$

It immediately follows that as φ only takes the Radon-Nikodym derivative of non-negative measures in φ-divergences, the resulting divergence is invariant from the choice of generator φ or $\varphi_\uparrow$, *i.e.*, $\mathrm{D}_\varphi = \mathrm{D}_{\varphi_\uparrow}$. Despite this, in what follows, we will see that the properties of $\varphi_\uparrow$ are much easier to work with. For instance, the convex-conjugate of $\varphi_\uparrow$ conveniently becomes

$$\varphi^*_\uparrow(p) = \sup_{x \geq 0} \{xp - \varphi(x)\}. \tag{29}$$

With these definitions, we can now state our result.

Theorem C.4. *Let $\varphi \in \Phi$. Suppose that there exists a constant $b \in \mathbb{R}$ such that*

$$\mathbb{E}_{\mathsf{P}}\left[(\gamma\varphi_\uparrow)^*(b - L(\mathsf{Z}))\right] < +\infty, \tag{30}$$

the right derivative $(\varphi^*)'_+$ exists P*-almost surely at $\frac{b - L(\mathsf{Z})}{\gamma}$, the quantity*

$$\mathbb{E}_{\mathsf{P}}\left[\left|(\gamma\varphi_\uparrow)^*(b - L(\mathsf{Z})) - \max\left\{0, (\varphi^*)'_+\left(\frac{b - L(\mathsf{Z})}{\gamma}\right)\right\} \cdot (b - L(\mathsf{Z}))\right|\right] < +\infty \tag{31}$$

is finite, and

$$\mathbb{E}_{\mathsf{P}}\left[\max\left\{0, (\varphi^*)'_+\left(\frac{b - L(z)}{\gamma}\right)\right\}\right] = 1 \tag{32}$$

Then optimal ratio $\rho^\star$ of Eq. (23) is given by

$$\rho^\star(z) = \max\left\{0, (\varphi^*)'_+\left(\frac{b - L(z)}{\gamma}\right)\right\}. \tag{33}$$

It follows that Corollary 3.3 follows immediately from Theorem C.4 by consider the generator of the KL-divergence.

In theory, Theorem C.4 provides a general bridge to connect an φ-divergence to different density-ratios which would general deviate from those primarily considered in this paper (*i.e.*, those which

correspond to the KL-divergence). This naturally leads to a different set of marginal and joint ideal weights, and thus, DR loss functions (as defined in Section 4).

However, there is a practical catch: we can no longer exploit the functional form of the KL-divergence ideal weights to select nice marginal label probabilities π to utilize un-normalized ideal weights. Indeed, for the KL-divergence, the normalizer $Z_{\mathrm{i},\gamma}$ is multiplicative (Corollary 3.3). However, in the general case, the normalizer b in Eq. (33) cannot be factored out. This is a problem as we do not immediately get an unbiased objective as per Definition 4.2 and Eq. (14). The best we could do is provide a two step approximation, where we first approximate the normalizer b (for each model h and $h^{(\mathrm{e})}$), and then create a version of Definition 4.2 using Eq. (33).

C.2 Helper Lemmas

To prove Theorem C.4, we first prove a number of helper lemmas.

We first consider a simplification one can yield for the convex conjugate of $\varphi_\uparrow$ by examining its right derivative.

Lemma C.5. *Let $\varphi\colon \mathbb{R} \to (-\infty, +\infty]$ be a proper convex, lower semi-continuous function such that $\varphi(0) < +\infty$ and $\varphi'_+(0) \in [-\infty, +\infty)$.*

$$\varphi^*_\uparrow(p) = \begin{cases} -\varphi(0) & \text{if } p \le \varphi'_+(0) \\ \varphi^*(p) & \text{if } p > \varphi'_+(0). \end{cases} \tag{34}$$

Proof. Fix $p \in \mathbb{R}$ and define $g_p(x) = xp - \varphi(x)$. Then g_p is concave and

$$\varphi^*_\uparrow(p) = \sup_{x \ge 0} g_p(x).$$

Suppose first that $\varphi'_+(0) = -\infty$. We claim that then $\varphi(x) = +\infty$ for all $x < 0$. Indeed, if there existed $x_0 < 0$ with $\varphi(x_0) < +\infty$, then for any $y > 0$ convexity would imply

$$\frac{\varphi(y) - \varphi(0)}{y} \ge \frac{\varphi(0) - \varphi(x_0)}{-x_0} > -\infty,$$

and sending $y \downarrow 0$ would yield $\varphi'_+(0) > -\infty$ (note, Definition C.2), a contradiction. Hence, in this case $\varphi(x) = +\infty$ for all $x < 0$, and thus $\varphi = \varphi_\uparrow$, so $\varphi^*_\uparrow(p) = \varphi^*(p)$ for all $p \in \mathbb{R}$.

We now proceed further with two cases with $\varphi'_+(0) > -\infty$.

① Suppose that $p \le \varphi'_+(0)$. For any $x > 0$, convexity implies that

$$\frac{\varphi(x) - \varphi(0)}{x} \ge \varphi'_+(0) \ge p.$$

Rearranging gives $g_p(x) = xp - \varphi(x) \le -\varphi(0) = g_p(0)$, and so $x = 0$ is a maximizer of g_p over $\mathbb{R}_{\ge 0}$. Therefore, $\varphi^*_\uparrow(p) = -\varphi(0)$.

② Now suppose that $p > \varphi'_+(0)$. By the definition of the right derivative, there exists $\bar{x} > 0$ such that

$$\frac{\varphi(\bar{x}) - \varphi(0)}{\bar{x}} < p,$$

which implies $g_p(\bar{x}) > -\varphi(0) = g_p(0)$. Hence, $x = 0$ is not a maximizer over $\mathbb{R}$.

For $x < 0$ and any $y > 0$, convexity gives

$$\frac{\varphi(0) - \varphi(x)}{-x} \le \frac{\varphi(y) - \varphi(0)}{y}.$$

Sending $y \downarrow 0$ yields $\varphi(x) \ge \varphi(0) + \varphi'_+(0)x$, and thus

$$g_p(x) = xp - \varphi(x) \le -\varphi(0) + (p - \varphi'_+(0))x < -\varphi(0).$$

Since $g_p(\bar{x}) > -\varphi(0)$ for some $\bar{x} > 0$, no negative x can maximize g_p over $\mathbb{R}$. Therefore,

$$\sup_{x \in \mathbb{R}} g_p(x) = \sup_{x \ge 0} g_p(x),$$

which yields $\varphi^*_\uparrow(p) = \varphi^*(p)$. □

Lemma C.6. *Let $\varphi\colon \mathbb{R} \to (-\infty, +\infty]$ be a proper convex, lower semi-continuous function such that $\varphi(0) < +\infty$. Then, at every $p \in \mathbb{R}$ where the right derivatives are defined,*

$$(\varphi_\uparrow^*)'_+(p) = \max\{0, (\varphi^*)'_+(p)\}. \tag{35}$$

Proof. We utilize Lemma C.5 and consider a few cases. First consider when $\varphi'_+(0) = -\infty$. From the proof of Lemma C.5, this implies that $\varphi(x) = +\infty$ for all $x < 0$. Hence, $\varphi = \varphi_\uparrow$. Therefore,

$$\varphi^*(p) = \sup_{x \geq 0}\{xp - \varphi(x)\}.$$

Since each affine function $p \mapsto xp - \varphi(x)$ has non-negative slope $x \geq 0$, the supremum φ^* is non-decreasing. Thus, at every point where $(\varphi^*)'_+$ is defined, we have $(\varphi^*)'_+(p) \geq 0$, and therefore

$$(\varphi_\uparrow^*)'_+(p) = (\varphi^*)'_+(p) = \max\{0, (\varphi^*)'_+(p)\}.$$

This proves the claim in the case $\varphi'_+(0) = -\infty$.

We further proceed with the assumption that $\varphi'_+(0) > -\infty$.

We note that $\varphi'_+(0) \in \partial\varphi(0)$. Indeed, for $x > 0$, convexity implies

$$\frac{\varphi(x) - \varphi(0)}{x} \geq \varphi'_+(0),$$

and hence $\varphi(x) \geq \varphi(0) + \varphi'_+(0)x$. For $x < 0$ and any $y > 0$, convexity gives

$$\frac{\varphi(0) - \varphi(x)}{-x} \leq \frac{\varphi(y) - \varphi(0)}{y}.$$

Sending $y \downarrow 0$ yields $\varphi(x) \geq \varphi(0) + \varphi'_+(0)x$. Thus, $\varphi'_+(0) \in \partial\varphi(0)$, and Fenchel-Young duality gives $0 \in \partial\varphi^*(\varphi'_+(0))$. Hence, $\varphi'_+(0)$ is a minimizer of the convex function φ^*, and therefore

$$(\varphi^*)'_+(p) \leq 0 \quad \text{for all } p < \varphi'_+(0), \qquad (\varphi^*)'_+(p) \geq 0 \quad \text{for all } p \geq \varphi'_+(0),$$

at every point where the right derivative is defined.

We now further proceed via considering three cases.

① Suppose first that $p < \varphi'_+(0)$. Then $\varphi_\uparrow^*$ is constant on the interval $[p, \varphi'_+(0)]$, and hence $(\varphi_\uparrow^*)'_+(p) = 0$. Therefore, $(\varphi_\uparrow^*)'_+(p) = 0 = \max\{0, (\varphi^*)'_+(p)\}$.

② Now suppose that $p > \varphi'_+(0)$. Then $\varphi_\uparrow^*(q) = \varphi^*(q)$ for all q in a right-neighborhood of p, and therefore $(\varphi_\uparrow^*)'_+(p) = (\varphi^*)'_+(p)$. Since $p > \varphi'_+(0)$, we also have $(\varphi^*)'_+(p) \geq 0$. Thus, $(\varphi_\uparrow^*)'_+(p) = \max\{0, (\varphi^*)'_+(p)\}$.

③ Finally, suppose that $p = \varphi'_+(0)$. For any $h > 0$, we have $\varphi_\uparrow^*(p+h) = \varphi^*(p+h)$, while Lemma C.5 gives $\varphi_\uparrow^*(p) = -\varphi(0)$. Since the argument above shows $p = \varphi'_+(0) \in \partial\varphi(0)$, Fenchel-Young duality yields $\varphi^*(p) = -\varphi(0) = \varphi_\uparrow^*(p)$. Therefore,

$$(\varphi_\uparrow^*)'_+(p) = \lim_{h \downarrow 0} \frac{\varphi_\uparrow^*(p+h) - \varphi_\uparrow^*(p)}{h} = \lim_{h \downarrow 0} \frac{\varphi^*(p+h) - \varphi^*(p)}{h} = (\varphi^*)'_+(p).$$

Since p minimizes φ^*, we have $(\varphi^*)'_+(p) \geq 0$, and hence $(\varphi_\uparrow^*)'_+(p) = \max\{0, (\varphi^*)'_+(p)\}$. □

C.3 Proof of Result

We now prove Theorem C.4.

Proof of Theorem C.4. In the following proof, we will show that $\rho^\star$ (as per Eq. (33)) obtains an upper bound, and hence will be optimal for the specified objective Eq. (23).

First notice that from Lemma C.6, we can rewrite $\rho^\star$ as

$$\rho^\star(z) = \max\left\{0, (\varphi^*)'_+\left(\frac{b - L(z)}{\gamma}\right)\right\} = (\varphi_\uparrow^*)'_+\left(\frac{b - L(z)}{\gamma}\right) = \left((\gamma\varphi_\uparrow)^*\right)'_+(b - L(z)), \tag{36}$$

where the middle equality follows from Lemmas C.5 and C.6: if $\frac{b-L(z)}{\gamma} < \varphi'_+(0)$ then $\varphi^*_\uparrow$ is locally constant, if $\frac{b-L(z)}{\gamma} > \varphi'_+(0)$ then $\varphi^*_\uparrow = \varphi^*$ locally, and if $\frac{b-L(z)}{\gamma} = \varphi'_+(0)$ then the right derivative is computed from the right and again agrees with $(\varphi^*)'_+$. Hence $(\varphi^*_\uparrow)'_+$ exists P-almost surely at $\frac{b-L(z)}{\gamma}$. The last equality follows from the scaling rule of convex-conjugates and the chain rule of differentiation.

We will now reframe the optimization objective in Eq. (23) as a maximization problem. In particular, we consider

$$\begin{aligned} \max_{\rho\colon \mathcal{Z}\to\mathbb{R}_{\geq 0}} \quad & \mathbb{E}_{\mathrm{P}}\left[\rho(\mathsf{Z})(-L(\mathsf{Z}))\right] - \gamma \cdot \mathrm{D}_\varphi(\rho\cdot \mathrm{P} \,\|\, \mathrm{P}), \\ s.t. \quad & \mathbb{E}_{\mathrm{P}}[\rho(\mathsf{Z})] = 1. \end{aligned}$$

This gives us the following:

$$\begin{aligned} & \max_{\rho\colon \mathcal{Z}\to\mathbb{R}_{\geq 0}, \mathbb{E}_{\mathrm{P}}[\rho(\mathsf{Z})]=1} \quad \mathbb{E}_{\mathrm{P}}\left[\rho(\mathsf{Z})(-L(\mathsf{Z}))\right] - \gamma \cdot \mathrm{D}_\varphi(\rho\cdot \mathrm{P} \,\|\, \mathrm{P}) \\ &= \max_{\rho\colon \mathcal{Z}\to\mathbb{R}_{\geq 0}, \mathbb{E}_{\mathrm{P}}[\rho(\mathsf{Z})]=1} \quad \mathbb{E}_{\mathrm{P}}\left[\rho(\mathsf{Z})(-L(\mathsf{Z}))\right] - \gamma \cdot \mathrm{D}_{\varphi_\uparrow}(\rho\cdot \mathrm{P} \,\|\, \mathrm{P}) \\ &= \max_{\mathrm{Q}\in\triangle(\mathcal{Z})} \quad \mathbb{E}_{\mathrm{Q}}\left[-L(\mathsf{Z})\right] - \gamma \cdot \mathrm{D}_{\varphi_\uparrow}(\mathrm{Q} \,\|\, \mathrm{P}) \\ \text{(a)} &= \inf_{c\in\mathbb{R}} \quad \gamma\mathbb{E}_{\mathrm{P}}\left[\varphi^*_\uparrow\left(\frac{c-L(\mathsf{Z})}{\gamma}\right)\right] - c \\ \text{(b)} &= \inf_{c\in\mathbb{R}} \quad \mathbb{E}_{\mathrm{P}}\left[(\gamma\varphi_\uparrow)^*\left(c-L(\mathsf{Z})\right)\right] - c \\ \text{(c)} &\leq \mathbb{E}_{\mathrm{P}}\left[(\gamma\varphi_\uparrow)^*\left(b-L(\mathsf{Z})\right)\right] - b, \end{aligned}$$

where (a) follows from a standard convex analysis result, *e.g.*, see Liu and Chaudhuri [2018, Equation (22)]; (b) follows from the scaling rule of convex-conjugates; and (c) follows from taking a specific choice for $c \in \mathbb{R}$, in particular, our b as per our assumption Eq. (32). The finiteness assumption Eq. (30) ensures that the right-hand side is well-defined at $c = b$. This yields our upper bound.

We now consider a corresponding lower bound. By Eq. (32), $\rho^\star$ is feasible for Eq. (23). Moreover, by Eq. (31) and the Fenchel-Young identity established below, all expressions evaluated at $\rho^\star$ are integrable, so the objective value at $\rho^\star$ is well-defined. In particular, we consider

$$\begin{aligned} & \max_{\rho\colon \mathcal{Z}\to\mathbb{R}_{\geq 0}, \mathbb{E}_{\mathrm{P}}[\rho(\mathsf{Z})]=1} \quad \mathbb{E}_{\mathrm{P}}\left[\rho(\mathsf{Z})(-L(\mathsf{Z}))\right] - \gamma \cdot \mathrm{D}_{\varphi_\uparrow}(\rho\cdot \mathrm{P} \,\|\, \mathrm{P}) \\ &= \max_{\rho\colon \mathcal{Z}\to\mathbb{R}_{\geq 0}, \mathbb{E}_{\mathrm{P}}[\rho(\mathsf{Z})]=1} \quad \mathbb{E}_{\mathrm{P}}\left[\rho(\mathsf{Z})(-L(\mathsf{Z})) - \gamma\cdot(\varphi_\uparrow\circ\rho)(\mathsf{Z})\right] \\ &= \max_{\rho\colon \mathcal{Z}\to\mathbb{R}_{\geq 0}, \mathbb{E}_{\mathrm{P}}[\rho(\mathsf{Z})]=1} \quad \mathbb{E}_{\mathrm{P}}\left[\rho(\mathsf{Z})(-L(\mathsf{Z})) - ((\gamma\varphi_\uparrow)\circ\rho)(\mathsf{Z})\right] \\ &\geq \mathbb{E}_{\mathrm{P}}\left[\rho^\star(\mathsf{Z})(-L(\mathsf{Z})) - ((\gamma\varphi_\uparrow)\circ\rho^\star)(\mathsf{Z})\right] \\ &= \mathbb{E}_{\mathrm{P}}\left[\rho^\star(\mathsf{Z})(b-L(\mathsf{Z})) - ((\gamma\varphi_\uparrow)\circ\rho^\star)(\mathsf{Z})\right] - \mathbb{E}_{\mathrm{P}}\left[\rho^\star(\mathsf{Z})b\right] \\ &= \mathbb{E}_{\mathrm{P}}\left[\rho^\star(\mathsf{Z})(b-L(\mathsf{Z})) - ((\gamma\varphi_\uparrow)\circ\rho^\star)(\mathsf{Z})\right] - b\cdot\mathbb{E}_{\mathrm{P}}\left[\rho^\star(\mathsf{Z})\right] \\ &= \mathbb{E}_{\mathrm{P}}\left[\rho^\star(\mathsf{Z})(b-L(\mathsf{Z})) - ((\gamma\varphi_\uparrow)\circ\rho^\star)(\mathsf{Z})\right] - b, \end{aligned}$$

where the last equality, we utilize the assumption on b, Eq. (32).

From our assumptions, it follows that $p \mapsto (\gamma\varphi_\uparrow)^*(p)$ is proper, convex, and lower semi-continuous on $\mathbb{R}$. By Eq. (36), $(\varphi^*_\uparrow)'_+$ exists at $\frac{b-L(z)}{\gamma}$ for P-almost every z. Therefore, the scaling rule for convex conjugates implies that $((\gamma\varphi_\uparrow)^*)'_+$ exists at $b-L(z)$ for P-almost every z. Moreover, Eq. (30) implies $(\gamma\varphi_\uparrow)^*(b-L(z)) < +\infty$ P-almost surely. For any proper convex lower semi-continuous function on $\mathbb{R}$, whenever the right derivative exists at a point in its effective domain, that right derivative belongs to the subdifferential at that point. Hence, leveraging Eq. (36), via Fenchel-Young duality, we have for P-almost every z

$$\rho^\star(z) \in \partial(\gamma\varphi_\uparrow)^*(b-L(z)) \quad \iff \quad b - L(z) \in \partial(\gamma\varphi_\uparrow)(\rho^\star(z)).$$

Furthermore, Fenchel-Young equality gives for P-almost every z

$$(\gamma\varphi_\uparrow)^*(b-L(z)) + (\gamma\varphi_\uparrow)(\rho^\star(z)) = (\gamma\varphi_\uparrow)^*(b-L(z)) + ((\gamma\varphi_\uparrow)\circ\rho^\star)(z) = (b-L(z))\cdot(\rho^\star(z)).$$

Since Eq. (31) ensures the right-hand side below is integrable, taking expectations and substituting this identity into our lower bound yields,

$$\begin{aligned}
&\max_{\rho\colon \mathcal{Z}\to\mathbb{R}_{\geq 0}, \mathbb{E}_{\mathsf{P}}[\rho(\mathsf{Z})]=1} \mathbb{E}_{\mathsf{P}}\left[\rho(\mathsf{Z})(-L(\mathsf{Z}))\right] - \gamma \cdot \mathrm{D}_{\varphi_\uparrow}(\rho\cdot \mathsf{P}\,\|\,\mathsf{P}) \\
&\geq \mathbb{E}_{\mathsf{P}}\left[\rho^\star(\mathsf{Z})(b - L(\mathsf{Z})) - ((\gamma\varphi_\uparrow)\circ\rho^\star)(\mathsf{Z})\right] - b \\
&= \mathbb{E}_{\mathsf{P}}\left[(\gamma\varphi_\uparrow)^*(b - L(\mathsf{Z}))\right] - b,
\end{aligned}$$

which exactly matches our upper bound. That is, $\rho^\star$ is optimal as per Eq. (23). □

D DRE to CPE Reduction

The following provides a self-contained explanation of the DRE to CPE reduction. For a full exposition, see Menon and Ong [2016].

D.1 Reduction

We first consider the problem where we wish to estimate a density-ratio. Let $\mu, \nu \in \triangle(\mathcal{X})$. Our task is to estimate

$$r^\star = \frac{\mathrm{d}\mu}{\mathrm{d}\nu}. \tag{37}$$

We now design a binary decision problem as per the following. Let $\pi \in (0,1)$ be the label probability. We define the following sampling scheme:

1. Sample a label R via
$$\mathsf{R} = \begin{cases} +1 & \text{with probability } \pi; \\ -1 & \text{with probability } 1-\pi. \end{cases} \tag{38}$$
2. Sample an example X via
$$\mathsf{X} \sim \begin{cases} \mu & \text{if } \mathsf{R} = +1; \\ \nu & \text{if } \mathsf{R} = -1. \end{cases} \tag{39}$$

Denote the joint law of this sampling scheme (X, R) as E, *i.e.*, $(\mathsf{X}, \mathsf{R}) \sim \mathsf{E}$. We can now define a classification problem for this binary decision problem. Let (λ, Φ) correspond to a strictly proper composite loss function. Then for a scorer s, we minimize

$$\mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{R})\sim\mathsf{E}}[\lambda(s(\mathsf{X}), \mathsf{R})] = \pi \cdot \mathbb{E}_\mu[(\lambda_{+1}\circ s)(\mathsf{X})] + (1-\pi)\cdot\mathbb{E}_\nu[(\lambda_{-1}\circ s)(\mathsf{X})], \tag{40}$$

where we have factorized the law E into its marginal label probability Eq. (38) and its conditional Eq. (39). This motivates the objective in Eq. (10).

As (λ, Φ) is strictly proper composite [Reid and Williamson, 2010], the population minimizer $s^\star$ will satisfy the following Bayes-optimal class-probabilities

$$(\Phi^{-1}\circ s^\star)(x) = \mathbb{P}[\mathsf{R} = +1 \mid \mathsf{X} = x]. \tag{41}$$

Finally, by the construction of our binary decision problem (Eqs. (38) and (39)), one can verify via Bayes rule that (symmetrically)

$$\begin{aligned}
\mathbb{P}[\mathsf{R} = +1 \mid \mathsf{X} = x] &= \frac{\pi\mu(x)}{\pi\mu(x) + (1-\pi)\nu(x)} \\
\mathbb{P}[\mathsf{R} = -1 \mid \mathsf{X} = x] &= \frac{(1-\pi)\nu(x)}{\pi\mu(x) + (1-\pi)\nu(x)}.
\end{aligned} \tag{42}$$

Dividing the two class-probabilities in Eq. (42) yields

$$\frac{\mathbb{P}[\mathsf{R} = +1 \mid \mathsf{X} = x]}{\mathbb{P}[\mathsf{R} = -1 \mid \mathsf{X} = x]} = \frac{\pi}{1-\pi}\frac{\mu(x)}{\nu(x)}.$$

Substituting in Eq. (41) (and taking an approximate s) yields Eq. (11).

D.2 Example

We further provide a simple worked example on a two point distribution. Suppose we have $\mu = [0.3, 0.7]$, $\nu = [0.5, 0.5]$, and $\pi = 0.5$ for simplicity. Writing the two points as x_1 ,x_2, the Bayes-optimal class-posterior is

$$\mathbb{P}[R = +1 \mid X = x_1] = \frac{0.3}{0.3 + 0.5} = \frac{3}{8}, \qquad \mathbb{P}[R = +1 \mid X = x_2] = \frac{0.7}{0.7 + 0.5} = \frac{7}{12}.$$

Since $\pi = 0.5$, the prior-odds correction $\pi/(1-\pi)$ is one. Thus, a Bayes-optimal CPE estimator yields

$$\left[\frac{3/8}{1 - 3/8}, \frac{7/12}{1 - 7/12}\right] = \left[\frac{3}{5}, \frac{7}{5}\right] = \left[\frac{0.3}{0.5}, \frac{0.7}{0.5}\right] = \frac{\mathrm{d}\mu}{\mathrm{d}\nu}.$$

E Connection to Chow's Rule

E.1 Proof of Theorem 3.5

Proof of Theorem 3.5. Fix $\gamma > 0$. From Corollary 3.3 we identify that the density-ratio ρ_{i} corresponds to the ratio of normalized weight function $w_{\mathrm{i}}^{\cdot}$. Thus, for a $\tau > 0$, we have rejectors

$$r_{\mathrm{i}}(x; \tau, \gamma, \gamma) = \left[\!\!\left[\frac{w_{\mathrm{i}}(x;\gamma)}{w_{\mathrm{i}}^{(\mathrm{e})}(x;\gamma)} \leq \tau \right]\!\!\right] = \left[\!\!\left[\frac{\tilde{w}_{\mathrm{i}}(x;\gamma)}{\tilde{w}_{\mathrm{i}}^{(\mathrm{e})}(x;\gamma)} \leq \tau' \right]\!\!\right],$$

where we are just setting $\tau' = \tau Z_{\mathrm{i},\gamma}/Z_{\mathrm{i},\gamma}^{(\mathrm{e})}$ (a positive scaling of the original threshold value).

To show that the KL-marginal density-ratio deferral rule is equivalent to Chow's Rule w.r.t. P, we consider the following simplification:

$$\begin{aligned}
\frac{\tilde{w}_{\mathrm{m}}(x;\gamma)}{\tilde{w}_{\mathrm{m}}^{(\mathrm{e})}(x;\gamma)} \leq \tau' &\iff \frac{\exp\left(-\frac{\mathbb{E}_{\eta^\star(x)}[\ell(x,\mathsf{Y},h(x))]}{\gamma}\right)}{\exp\left(-\frac{\mathbb{E}_{\eta^\star(x)}\left[\ell(x,\mathsf{Y},h^{(\mathrm{e})}(x))\right]}{\gamma}\right)} \leq \tau' \\
&\iff \exp\left(\frac{\mathbb{E}_{\eta^\star(x)}\left[\ell(x,\mathsf{Y},h^{(\mathrm{e})}(x))\right] - \mathbb{E}_{\eta^\star(x)}\left[\ell(x,\mathsf{Y},h(x))\right]}{\gamma}\right) \leq \tau' \\
&\iff \mathbb{E}_{\eta^\star(x)}\left[\ell(x,\mathsf{Y},h^{(\mathrm{e})}(x)) - \ell(x,\mathsf{Y},h(x))\right] \leq \gamma \log \tau' \\
&\iff \mathbb{E}_{\eta^\star(x)}\left[\ell(x,\mathsf{Y},h(x)) - \ell(x,\mathsf{Y},h^{(\mathrm{e})}(x))\right] \geq -\gamma \log \tau'.
\end{aligned}$$

Notice that this is exactly the condition for $r^\star(x; -\gamma \log \tau', \mathrm{P}) = 1$. For any prescribed cost $c \in \mathbb{R}$, set $\tau' = \exp(-c/\gamma) > 0$. Since $\tau' = \tau Z_{\mathrm{m},\gamma}/Z_{\mathrm{m},\gamma}^{(\mathrm{e})}$, choosing

$$\tau = \exp(-c/\gamma)\frac{Z_{\mathrm{m},\gamma}^{(\mathrm{e})}}{Z_{\mathrm{m},\gamma}} > 0$$

yields $-\gamma \log \tau' = c$ and hence $r_{\mathrm{m}}(x; \tau, \gamma, \gamma) = r^\star(x; c, \mathrm{P})$.

To show the connection that KL-joint density-ratio deferral rule has between Chow's Rule on the expert-tilted distribution, we consider the following:

$$\begin{aligned}
\frac{\tilde{w}_{\mathrm{j}}(x;\gamma)}{\tilde{w}_{\mathrm{j}}^{(\mathrm{e})}(x;\gamma)} \leq \tau' &\iff \frac{\mathbb{E}_{\eta^\star(x)}\left[\exp\left(-\frac{\ell(x,\mathsf{Y},h(x))}{\gamma}\right)\right]}{\mathbb{E}_{\eta^\star(x)}\left[\exp\left(-\frac{\ell(x,\mathsf{Y},h^{(\mathrm{e})}(x))}{\gamma}\right)\right]} \leq \tau' \\
&\iff \mathbb{E}_{\eta^\star(x)}\left[t(\mathsf{Y} \mid x; \gamma, \mathrm{P})\frac{\exp\left(-\frac{\ell(x,\mathsf{Y},h(x))}{\gamma}\right)}{\exp\left(-\frac{\ell(x,\mathsf{Y},h^{(\mathrm{e})}(x))}{\gamma}\right)}\right] \leq \tau' \\
&\iff \log \mathbb{E}_{\tilde{\eta}^\star(x;\gamma)}\left[\exp\left(\frac{\ell(x,\mathsf{Y},h^{(\mathrm{e})}(x)) - \ell(x,\mathsf{Y},h(x))}{\gamma}\right)\right] \leq \log \tau'.
\end{aligned}$$

Using Jensen's inequality, we have

$$\frac{\tilde{w}_{\mathrm{j}}(x;\gamma)}{\tilde{w}_{\mathrm{j}}^{(\mathrm{e})}(x;\gamma)} \leq \tau' \implies \mathbb{E}_{\tilde{\eta}^\star(x;\gamma)}\left[\ell(x,\mathsf{Y},h(x)) - \ell(x,\mathsf{Y},h^{(\mathrm{e})}(x))\right] \geq -\gamma \log \tau'.$$

The RHS condition is exactly equivalent to $r^\star(x; -\gamma \log \tau', \tilde{\mathrm{P}}_\gamma) = 1$ for the posterior twisted distribution $\tilde{\mathrm{P}}_\gamma$. For any prescribed cost $c \in \mathbb{R}$, set $\tau' = \exp(-c/\gamma) > 0$, and then choose

$$\tau = \exp(-c/\gamma)\frac{Z_{\mathrm{j},\gamma}^{(\mathrm{e})}}{Z_{\mathrm{j},\gamma}} > 0.$$

This gives $-\gamma \log \tau' = c$, hence $r_{\mathrm{j}}(x;\tau,\gamma,\gamma) \leq r^\star(x;c,\tilde{\mathrm{P}}_\gamma)$. □

F Joint Ideal DR CPE Loss Properties

In the following section, we provide a simple proof of the Jensen's ordering and also give a result to show that the gap between the marginal and joint ideal DR loss objectives is controlled by conditional task loss variances (weighted by the DR partial losses).

F.1 Proof of Lemma 4.3

Proof of Lemma 4.3. The result immediately follows from Jensen's inequality. □

F.2 Variance upper bound of Jensen's gap

Lemma F.1. *Let* Z *be a bounded random variable in* $[m, M]$ *for* $m < M$*. Furthermore, let* $\mathbb{E}[\mathsf{Z}] = 0$*. Then for any* $\gamma > 0$*, we have that*

$$\exp\left(-\frac{M}{\gamma}\right)\frac{1}{2\gamma^2}\mathbb{V}[\mathsf{Z}] \leq \left|\mathbb{E}\left[\exp\left(-\frac{\mathsf{Z}}{\gamma}\right)\right] - 1\right| \leq \exp\left(-\frac{m}{\gamma}\right)\frac{1}{2\gamma^2}\mathbb{V}[\mathsf{Z}]. \tag{43}$$

Proof. We consider the 2nd order Taylor expansion of the map $z \mapsto \exp(-z/\gamma)$ around $z = 0$ with an explicit Lagrange remainder. Given the maps sufficient smoothness, we have

$$\begin{aligned}\exp(-z/\gamma) &= \exp(0) - \frac{\exp(0)}{\gamma}z + \frac{\exp(-(\theta z)/\gamma)}{2\gamma^2}z^2 \\ &= 1 - \frac{z}{\gamma} + \frac{z^2}{2\gamma^2}\exp(-(\theta z)/\gamma),\end{aligned}$$

where $\theta \in [0,1]$.

Re-arranging, setting $z = \mathsf{Z}$, and using the positivity of $z^2/(2\gamma^2)$, we get the bound

$$-\frac{\mathsf{Z}}{\gamma} + \frac{\mathsf{Z}^2}{2\gamma^2}\exp\left(-\frac{M}{\gamma}\right) \leq \exp\left(-\frac{\mathsf{Z}}{\gamma}\right) - 1 \leq -\frac{\mathsf{Z}}{\gamma} + \frac{\mathsf{Z}^2}{2\gamma^2}\exp\left(-\frac{m}{\gamma}\right).$$

Finally, taking an expectation and noting that Z is a centered random variable yields the desired result. □

Theorem F.2. *Suppose* $0 \leq \ell \leq B$ *and that the partial losses satisfy* $\lambda_{+1} \geq 0$ *and* $\lambda_{-1} \geq 0$*. Then*

$$\begin{aligned}&\frac{\exp(-B/\gamma)}{2\gamma^2}\mathbb{E}_{\mathrm{P}_{\mathsf{x}}}\left[(\lambda_{+1}\circ s)(\mathsf{X})\mathbb{V}_{\eta^\star(\mathsf{X})}[\ell(\mathsf{X},\mathsf{Y},h(\mathsf{X}))]\right] \\ &\qquad + \frac{\exp(-B/\gamma^{(\mathrm{e})})}{2(\gamma^{(\mathrm{e})})^2}\mathbb{E}_{\mathrm{P}_{\mathsf{x}}}\left[(\lambda_{-1}\circ s)(\mathsf{X})\mathbb{V}_{\eta^\star(\mathsf{X})}[\ell(\mathsf{X},\mathsf{Y},h^{(\mathrm{e})}(\mathsf{X}))]\right] \\ &\leq \mathcal{L}_{\mathrm{j}}(s;\gamma,\gamma^{(\mathrm{e})}) - \mathcal{L}_{\mathrm{m}}(s;\gamma,\gamma^{(\mathrm{e})}) \\ &\leq \frac{1}{2\gamma^2}\mathbb{E}_{\mathrm{P}_{\mathsf{x}}}\left[(\lambda_{+1}\circ s)(\mathsf{X})\mathbb{V}_{\eta^\star(\mathsf{X})}[\ell(\mathsf{X},\mathsf{Y},h(\mathsf{X}))]\right] \\ &\qquad + \frac{1}{2(\gamma^{(\mathrm{e})})^2}\mathbb{E}_{\mathrm{P}_{\mathsf{x}}}\left[(\lambda_{-1}\circ s)(\mathsf{X})\mathbb{V}_{\eta^\star(\mathsf{X})}[\ell(\mathsf{X},\mathsf{Y},h^{(\mathrm{e})}(\mathsf{X}))]\right].\end{aligned}$$

Proof of Theorem F.2. From Lemma 4.3 and the non-negativity of the partial losses, we have that

$$\mathcal{L}_{\mathrm{j}}(s;\gamma,\gamma^{(\mathrm{e})}) - \mathcal{L}_{\mathrm{m}}(s;\gamma,\gamma^{(\mathrm{e})}) = A + B,$$

where

$$A = \mathbb{E}_{\mathsf{X}\sim\mathrm{P}_{\mathrm{x}}}\left[(\lambda_{+1}\circ s)(\mathsf{X})\mathbb{E}_{\mathsf{Y}\sim\eta^\star(\mathsf{X})}\left[\exp\left(-\frac{\ell(\mathsf{X},\mathsf{Y},h(\mathsf{X}))}{\gamma}\right) - \exp\left(-\frac{\mathbb{E}_{\mathsf{Y}\sim\eta^\star(\mathsf{X})}[\ell(\mathsf{X},\mathsf{Y},h(\mathsf{X}))]}{\gamma}\right)\right]\right]$$
$$B = \mathbb{E}_{\mathsf{X}\sim\mathrm{P}_{\mathrm{x}}}\left[(\lambda_{-1}\circ s)(\mathsf{X})\mathbb{E}_{\mathsf{Y}\sim\eta^\star(\mathsf{X})}\left[\exp\left(-\frac{\ell(\mathsf{X},\mathsf{Y},h^{(\mathrm{e})}(\mathsf{X}))}{\gamma^{(\mathrm{e})}}\right) - \exp\left(-\frac{\mathbb{E}_{\mathsf{Y}\sim\eta^\star(\mathsf{X})}[\ell(\mathsf{X},\mathsf{Y},h^{(\mathrm{e})}(\mathsf{X}))]}{\gamma^{(\mathrm{e})}}\right)\right]\right].$$

We will bound each of these terms identically, noting that the only difference between them is that γ is replaced by $\gamma^{(\mathrm{e})}$ and h is replaced by $h^{(\mathrm{e})}$. In particular, we examine the inner expectation of A. W.l.o.g., we have

$$\begin{aligned}
&\mathbb{E}_{\mathsf{Y}\sim\eta^\star(\mathsf{X})}\left[\exp\left(-\frac{\ell(\mathsf{X},\mathsf{Y},h(\mathsf{X}))}{\gamma}\right) - \exp\left(-\frac{\mathbb{E}_{\mathsf{Y}\sim\eta^\star(\mathsf{X})}[\ell(\mathsf{X},\mathsf{Y},h(\mathsf{X}))]}{\gamma}\right)\right] \\
&= \exp\left(-\frac{\mathbb{E}_{\mathsf{Y}\sim\eta^\star(\mathsf{X})}[\ell(\mathsf{X},\mathsf{Y},h(\mathsf{X}))]}{\gamma}\right)\mathbb{E}_{\mathsf{Y}\sim\eta^\star(\mathsf{X})}\left[\exp\left(-\frac{\ell(\mathsf{X},\mathsf{Y},h(\mathsf{X}))}{\gamma} + \frac{\mathbb{E}_{\mathsf{Y}\sim\eta^\star(\mathsf{X})}[\ell(\mathsf{X},\mathsf{Y},h(\mathsf{X}))]}{\gamma}\right) - 1\right] \\
&= \exp\left(-\frac{\mathbb{E}_{\mathsf{Y}\sim\eta^\star(\mathsf{X})}[\ell(\mathsf{X},\mathsf{Y},h(\mathsf{X}))]}{\gamma}\right)\mathbb{E}_{\mathsf{Y}\sim\eta^\star(\mathsf{X})}\left[\exp\left(-\frac{\ell(\mathsf{X},\mathsf{Y},h(\mathsf{X})) - \mathbb{E}_{\mathsf{Y}\sim\eta^\star(\mathsf{X})}[\ell(\mathsf{X},\mathsf{Y},h(\mathsf{X}))]}{\gamma}\right) - 1\right].
\end{aligned}$$

Noting that for fixed $\mathsf{X} = x$, we can define a centered random variable

$$\mathsf{Z} = \ell(\mathsf{X},\mathsf{Y},h(\mathsf{X})) - \langle\ell\rangle,$$

where we shorthand $\langle\ell\rangle = \mathbb{E}_{\mathsf{Y}\sim\eta^\star(\mathsf{X})}[\ell(\mathsf{X},\mathsf{Y},h(\mathsf{X}))]$.

Since $0 \le \ell \le B$, we have $\langle\ell\rangle \in [0,B]$ and therefore $\mathsf{Z} \in [-\langle\ell\rangle, B - \langle\ell\rangle]$. Thus, via Lemma F.1, we have that

$$\begin{aligned}
&\mathbb{E}_{\mathsf{Y}\sim\eta^\star(\mathsf{X})}\left[\exp\left(-\frac{\ell(\mathsf{X},\mathsf{Y},h(\mathsf{X}))}{\gamma}\right) - \exp\left(-\frac{\mathbb{E}_{\mathsf{Y}\sim\eta^\star(\mathsf{X})}[\ell(\mathsf{X},\mathsf{Y},h(\mathsf{X}))]}{\gamma}\right)\right] \\
&\in \frac{1}{2\gamma^2}\mathbb{V}[\ell(\mathsf{X},\mathsf{Y},h(\mathsf{X})) \mid \mathsf{X} = x]\left[\exp\left(-\frac{B}{\gamma}\right), 1\right].
\end{aligned}$$

Multiplying by the non-negative factor $(\lambda_{+1}\circ s)(x)$ and taking an outer expectation gives

$$\begin{aligned}
&\frac{\exp(-B/\gamma)}{2\gamma^2}\mathbb{E}_{\mathrm{P}_{\mathrm{x}}}\left[(\lambda_{+1}\circ s)(\mathsf{X})\mathbb{V}[\ell(\mathsf{X},\mathsf{Y},h(\mathsf{X})) \mid \mathsf{X}]\right] \\
&\qquad \le A \le \\
&\frac{1}{2\gamma^2}\mathbb{E}_{\mathrm{P}_{\mathrm{x}}}\left[(\lambda_{+1}\circ s)(\mathsf{X})\mathbb{V}[\ell(\mathsf{X},\mathsf{Y},h(\mathsf{X})) \mid \mathsf{X}]\right].
\end{aligned}$$

Applying the same argument to B with $(\gamma^{(\mathrm{e})}, h^{(\mathrm{e})})$ in place of (γ, h) and summing the two bounds yields the claim. □

G Approximate Consistency of DR CPE Losses

In the following section, we provide formal proofs and additional result regarding the approximate consistency of the joint ideal DR CPE losses. We formally define the additional assumptions required to prove various consistency-style results. For this section, we focus on the setting where we have a common temperature $\gamma = \gamma^{(\mathrm{e})} > 0$.

G.1 Assumptions and Definitions

To prove consistency results, we utilize a standard surrogate-regret analysis utilized in Agarwal [2014]. To do so, we consider the class of *strongly proper* loss functions which will allow us to transfer the L2D regret (derived from its risk defined in Eq. (1)) to the regret of DR CPE losses.

Definition G.1 (Agarwal [2014, Definition 7]). *A (proper) composite loss function* (λ, Φ) *is* β*-strongly proper composite if*

$$L(\eta, v; \lambda) - L(\eta, \Phi(\eta); \lambda) \geq \frac{\beta}{2} \cdot \left(\Phi^{-1}(v) - \eta\right)^2 \tag{44}$$

where $L(\eta, v; \lambda)$ *corresponds to the pointwise loss of* λ*, given by*

$$L(\eta, v; \lambda) \stackrel{\text{def}}{=} \eta \lambda_{+1}(v) + (1 - \eta)\lambda_{-1}(v). \tag{45}$$

The second term on the LHS of Eq. (44) for a proper composite loss corresponds to the pointwise Bayes risk of the loss function, *i.e.*, it is the minimum obtainable pointwise loss possible. The definition essentially ensure strong concavity of the Bayes pointwise risk [Agarwal, 2014, Theorem 10].

When we examine the pointwise minimizer of the marginal ideal objective $\mathcal{L}_{\mathrm{m}}(s; \gamma, \gamma)$ with shared temperature $\gamma = \gamma^{(\mathrm{e})} > 0$. In this shared temperature case, we can explicitly state the optimal scorer in terms of the un-normalized marginal reweightings.

Lemma G.2. *Let* (λ, Φ) *be a proper composite loss function. Then, the following is a minimizer for* $s^\star_{\mathrm{m}} \in \inf_s \mathcal{L}_{\mathrm{m}}(s; \gamma, \gamma)$*:*

$$s^\star_{\mathrm{m}}(x; \gamma, \gamma) = \Phi(p_{\mathrm{m}}(x; \gamma)), \tag{46}$$

where

$$p_{\mathrm{m}}(x; \gamma) \stackrel{\text{def}}{=} \frac{\tilde{w}_{\mathrm{m}}(x; \gamma)}{\tilde{w}_{\mathrm{m}}(x; \gamma) + \tilde{w}^{(\mathrm{e})}_{\mathrm{m}}(x; \gamma)}. \tag{47}$$

Proof. We write the marginal DR loss objective as

$$\begin{aligned}
&\mathcal{L}_{\mathrm{m}}(s; \gamma, \gamma) \\
&= \mathbb{E}_{\mathrm{P}_{\mathrm{x}}}\left[\tilde{w}_{\mathrm{m}}(\mathsf{X}; \gamma) \cdot (\lambda_{+1} \circ s)(\mathsf{X}) + \tilde{w}^{(\mathrm{e})}_{\mathrm{m}}(\mathsf{X}; \gamma) \cdot (\lambda_{-1} \circ s)(\mathsf{X})\right] \\
&= \mathbb{E}_{\mathrm{P}_{\mathrm{x}}}\left[\left(\tilde{w}_{\mathrm{m}}(\mathsf{X}; \gamma) + \tilde{w}^{(\mathrm{e})}_{\mathrm{m}}(\mathsf{X}; \gamma)\right) \cdot \left(p_{\mathrm{m}}(\mathsf{X}; \gamma) \cdot (\lambda_{+1} \circ s)(\mathsf{X}) + (1 - p_{\mathrm{m}}(\mathsf{X}; \gamma)) \cdot (\lambda_{-1} \circ s)(\mathsf{X})\right)\right] \\
&= \mathbb{E}_{\mathrm{P}_{\mathrm{x}}}\left[\left(\tilde{w}_{\mathrm{m}}(\mathsf{X}; \gamma) + \tilde{w}^{(\mathrm{e})}_{\mathrm{m}}(\mathsf{X}; \gamma)\right) \cdot L(p_{\mathrm{m}}(\mathsf{X}; \gamma), s(\mathsf{X}); \lambda)\right].
\end{aligned} \tag{48}$$

However, as (λ, Φ) is a proper composite loss function, then the pointwise optimal scorer must be given by $s^\star_{\mathrm{m}} = \Phi \circ p_{\mathrm{m}}(\cdot; \gamma)$, as required. □

A further helpful lemma that we will use is to relate the Chow's optimal deferral mechanism with the induced marginal ideal distribution probability p_{m} in Eq. (47).

Lemma G.3. *The optimal rejector given by Lemma 2.1* $r^\star(x; c, \mathrm{P})$ *can be stated as*

$$r^\star(x; c, \mathrm{P}) = \left[\!\!\left[p_{\mathrm{m}}(x; \gamma) \leq \sigma\left(-\frac{c}{\gamma}\right)\right]\!\!\right], \tag{49}$$

where $\sigma(t) \stackrel{\text{def}}{=} (1 + \exp(-t))^{-1}$ *is the sigmoid function.*

Proof. Notice that we directly have

$$\begin{aligned}
p_{\mathrm{m}}(x; \gamma) &= \frac{\tilde{w}_{\mathrm{m}}(x; \gamma)}{\tilde{w}_{\mathrm{m}}(x; \gamma) + \tilde{w}^{(\mathrm{e})}_{\mathrm{m}}(x; \gamma)} \\
&= \frac{\exp\left(-\frac{\mathbb{E}_{\mathsf{Y} \sim \eta^\star(x)}[\ell(x, \mathsf{Y}, h(x))]}{\gamma}\right)}{\exp\left(-\frac{\mathbb{E}_{\mathsf{Y} \sim \eta^\star(x)}[\ell(x, \mathsf{Y}, h(x))]}{\gamma}\right) + \exp\left(-\frac{\mathbb{E}_{\mathsf{Y} \sim \eta^\star(x)}\left[\ell(x, \mathsf{Y}, h^{(\mathrm{e})}(x))\right]}{\gamma}\right)} \\
&= \left(1 + \exp\left(\frac{\mathbb{E}_{\mathsf{Y} \sim \eta^\star(x)}\left[\ell(x, \mathsf{Y}, h(x)) - \ell(x, \mathsf{Y}, h^{(\mathrm{e})}(x))\right]}{\gamma}\right)\right)^{-1} \\
&= \sigma\left(-\frac{\mathbb{E}_{\mathsf{Y} \sim \eta^\star(x)}\left[\ell(x, \mathsf{Y}, h(x)) - \ell(x, \mathsf{Y}, h^{(\mathrm{e})}(x))\right]}{\gamma}\right).
\end{aligned} \tag{50}$$

Finally, the invertibility of σ gives the desired result. □

For convenience, we also define the follow risk related quantities for the DR CPE loss functions (for $\mathrm{i} \in \{\mathrm{j}, \mathrm{m}\}$):

$$\mathcal{L}_{\mathrm{i}}^{\star}(\gamma, \gamma^{(\mathrm{e})}) \stackrel{\text{def}}{=} \inf_{s} \mathcal{L}_{\mathrm{i}}(s; \gamma, \gamma^{(\mathrm{e})}); \tag{51}$$

$$\mathcal{E}_{\mathrm{i}}(s; \gamma, \gamma^{(\mathrm{e})}) \stackrel{\text{def}}{=} \mathcal{L}_{\mathrm{i}}(s; \gamma, \gamma^{(\mathrm{e})}) - \mathcal{L}_{\mathrm{i}}^{\star}(\gamma, \gamma^{(\mathrm{e})}); \tag{52}$$

$$\Delta_{\mathrm{mj}}(\gamma, \gamma^{(\mathrm{e})}) \stackrel{\text{def}}{=} \mathcal{L}_{\mathrm{j}}^{\star}(\gamma, \gamma^{(\mathrm{e})}) - \mathcal{L}_{\mathrm{m}}^{\star}(\gamma, \gamma^{(\mathrm{e})}), \tag{53}$$

which correspond to the Bayes-optimal risk, excess risk, and joint-marginal optimality gap of the DR CPE losses, respectively. The infimum is over all measurable scorers. As per the notation of the main text, we make the dependence of the composite loss (λ, Φ) implicit.

G.2 Helper Lemma

To prove the general consistency result, we first prove a general lemma which connects the deferral regret with the explicit disagreements of deferral decisions. This corresponds to a generalization of Jitkrittum et al. [2023, Corollary 3.2].

Lemma G.4. *For a deferral cost $c \geq 0$ and any rejector $r \colon \mathcal{X} \to \{0, 1\}$, the regret is given by*

$$\mathcal{L}(h, r; c) - \mathcal{L}(h, r^{\star}; c) = \mathbb{E}_{\mathsf{X} \sim \mathrm{P}_{\mathrm{x}}} \left[|\alpha(\mathsf{X})| \cdot [\![r(\mathsf{X}) \neq r^{\star}(\mathsf{X}; c, \mathrm{P})]\!] \right] \tag{54}$$

where $\alpha(x) = \mathbb{E}_{\mathsf{Y} \sim \eta^{\star}(x)} \left[\ell(x, \mathsf{Y}, h(x)) - \ell(x, \mathsf{Y}, h^{(\mathrm{e})}(x)) \right] - c$.

Proof. For a fixed classification model $h \colon \mathcal{X} \to \mathbb{R}^L$ and expert model $h^{(\mathrm{e})} \colon \mathcal{X} \to \mathbb{R}^L$, one can write the L2D risk (Eq. (1)) as

$$\begin{aligned}
\mathcal{L}(h, r; c) &= \mathbb{E}_{\mathsf{X}, \mathsf{Y} \sim \mathrm{P}} \left[(1 - r(\mathsf{X})) \cdot \ell(\mathsf{X}, \mathsf{Y}, h(\mathsf{X})) + r(\mathsf{X}) \cdot \ell(\mathsf{X}, \mathsf{Y}, h^{(\mathrm{e})}(\mathsf{X})) \right] + c \cdot \mathbb{P}_{\mathsf{X} \sim \mathrm{P}_{\mathrm{x}}}[r(\mathsf{X}) = 1] \\
&= \mathbb{E}_{\mathsf{X}, \mathsf{Y} \sim \mathrm{P}} \left[\ell(\mathsf{X}, \mathsf{Y}, h(\mathsf{X})) \right] + \mathbb{E}_{\mathsf{X}, \mathsf{Y} \sim \mathrm{P}} \left[r(\mathsf{X}) \left(\ell(\mathsf{X}, \mathsf{Y}, h^{(\mathrm{e})}(\mathsf{X})) - \ell(\mathsf{X}, \mathsf{Y}, h(\mathsf{X})) + c \right) \right] \\
&= \mathbb{E}_{\mathsf{X}, \mathsf{Y} \sim \mathrm{P}} \left[\ell(\mathsf{X}, \mathsf{Y}, h(\mathsf{X})) \right] - \mathbb{E}_{\mathsf{X} \sim \mathrm{P}_{\mathrm{x}}} \left[r(\mathsf{X}) \alpha(\mathsf{X}) \right].
\end{aligned}$$

In the second equality, we are using the fact that r returns binary outputs.

Thus, the regret is given by

$$\mathcal{L}(h, r; c) - \mathcal{L}(h, r^{\star}; c) = \mathbb{E}_{\mathsf{X} \sim \mathrm{P}_{\mathrm{x}}} \left[\alpha(\mathsf{X}) \cdot (r^{\star}(\mathsf{X}; c, \mathrm{P}) - r(\mathsf{X})) \right].$$

Finally, from Lemma 2.1, notice that $\alpha(x) \geq 0$ iff $r^{\star}(x; c, \mathrm{P}) = 1$. Otherwise, $r^{\star}(x; c, \mathrm{P}) = 0$. Thus we have,

$$\begin{aligned}
\alpha(x) \cdot (r^{\star}(x; c, \mathrm{P}) - r(x)) &= \begin{cases} |\alpha(x)| \cdot (r^{\star}(x; c, \mathrm{P}) - r(x)) & \alpha(x) \geq 0, r^{\star}(x; c, \mathrm{P}) = 1 \\ |\alpha(x)| \cdot (r(x) - r^{\star}(x; c, \mathrm{P})) & \alpha(x) \leq 0, r^{\star}(x; c, \mathrm{P}) = 0 \end{cases} \\
&= \begin{cases} |\alpha(x)| \cdot (1 - r(x)) & \alpha(x) \geq 0, r^{\star}(x; c, \mathrm{P}) = 1 \\ |\alpha(x)| \cdot (r(x) - 0) & \alpha(x) \leq 0, r^{\star}(x; c, \mathrm{P}) = 0 \end{cases} \\
&= \begin{cases} |\alpha(x)| \cdot [\![r(x) = 0]\!] & \alpha(x) \geq 0, r^{\star}(x; c, \mathrm{P}) = 1 \\ |\alpha(x)| \cdot [\![r(x) = 1]\!] & \alpha(x) \leq 0, r^{\star}(x; c, \mathrm{P}) = 0 \end{cases} \\
&= \begin{cases} |\alpha(x)| \cdot [\![r(x) \neq r^{\star}(x)]\!] & \alpha(x) \geq 0, r^{\star}(x; c, \mathrm{P}) = 1 \\ |\alpha(x)| \cdot [\![r(x) \neq r^{\star}(x)]\!] & \alpha(x) \leq 0, r^{\star}(x; c, \mathrm{P}) = 0 \end{cases} \\
&= |\alpha(x)| \cdot [\![r(x) \neq r^{\star}(x; c, \mathrm{P})]\!].
\end{aligned}$$

Taking an expectation over x yields the desired result. □

G.3 Proof of Theorem 4.4 (Approximate Consistency Bound)

Theorem G.5. *Suppose that the loss ℓ is bounded in $[0, B]$, the composite loss (λ, Φ) is β-strongly proper with non-negative partial losses $\lambda_{+1}, \lambda_{-1} \geq 0$, and we consider a shared temperature $\gamma = \gamma^{(\mathrm{e})} > 0$. Define the rejector*

$$r_{c,\gamma}(x) \stackrel{\text{def}}{=} [\![(\Phi^{-1} \circ s)(x) \leq \tau_{c,\gamma}]\!], \qquad \tau_{c,\gamma} \stackrel{\text{def}}{=} \sigma\left(-\frac{c}{\gamma} \right). \tag{55}$$

Then

$$\mathcal{L}(h, r_{c,\gamma}; c) - \mathcal{L}(h, r^\star; c) \leq K_{\gamma,B,c} \sqrt{\frac{\exp(B/\gamma)}{\beta} \left[\mathcal{E}_{\mathrm{j}}(s;\gamma,\gamma) + \Delta_{\mathrm{mj}}(\gamma,\gamma)\right]}, \tag{56}$$

where

$$K_{\gamma,B,c} \stackrel{\text{def}}{=} \frac{\gamma}{\sigma\left(-\frac{\max\{B,c\}}{\gamma}\right)\left(1 - \sigma\left(-\frac{\max\{B,c\}}{\gamma}\right)\right)}. \tag{57}$$

Proof. The proof follows similarly to Agarwal [2014, Theorem 13].

From strong-properness (Definition G.1), Eq. (48), and Lemma G.2, we can lower bound the marginal DR loss excess risk:

$$\begin{aligned}
&\mathcal{E}_{\mathrm{m}}(s;\gamma,\gamma) \\
&= \mathcal{L}_{\mathrm{m}}(s;\gamma,\gamma) - \mathcal{L}^\star_{\mathrm{m}}(\gamma,\gamma) \\
&= \mathcal{L}_{\mathrm{m}}(s;\gamma,\gamma) - \mathcal{L}_{\mathrm{m}}(s^\star_{\mathrm{m}};\gamma,\gamma) \\
&= \mathbb{E}_{\mathrm{P_x}}\left[\left(\tilde{w}_{\mathrm{m}}(\mathsf{X};\gamma) + \tilde{w}^{(\mathrm{e})}_{\mathrm{m}}(\mathsf{X};\gamma)\right) \cdot \left(L(p_{\mathrm{m}}(\mathsf{X};\gamma), s(\mathsf{X});\lambda) - L(p_{\mathrm{m}}(\mathsf{X};\gamma), s^\star_{\mathrm{m}}(\mathsf{X});\lambda)\right)\right] \\
&= \mathbb{E}_{\mathrm{P_x}}\left[\left(\tilde{w}_{\mathrm{m}}(\mathsf{X};\gamma) + \tilde{w}^{(\mathrm{e})}_{\mathrm{m}}(\mathsf{X};\gamma)\right) \cdot \left(L(p_{\mathrm{m}}(\mathsf{X};\gamma), s(\mathsf{X});\lambda) - L(p_{\mathrm{m}}(\mathsf{X};\gamma), \Phi(p_{\mathrm{m}}(\mathsf{X};\gamma));\lambda)\right)\right] \\
&\geq \frac{\beta}{2} \cdot \mathbb{E}_{\mathrm{P_x}}\left[\left(\tilde{w}_{\mathrm{m}}(\mathsf{X};\gamma) + \tilde{w}^{(\mathrm{e})}_{\mathrm{m}}(\mathsf{X};\gamma)\right) \cdot \left(\Phi^{-1}(s(\mathsf{X})) - p_{\mathrm{m}}(x;\gamma)\right)^2\right].
\end{aligned} \tag{58}$$

Furthermore, from the bounded loss $\ell \in [0, B]$ assumption, we have that

$$\tilde{w}_{\mathrm{m}}(\mathsf{X};\gamma) + \tilde{w}^{(\mathrm{e})}_{\mathrm{m}}(\mathsf{X};\gamma) \in \left[2\exp\left(-\frac{B}{\gamma}\right), 2\right]. \tag{59}$$

Thus we have

$$\mathbb{E}_{\mathrm{P_x}}\left[\left(\Phi^{-1}(s(\mathsf{X})) - p_{\mathrm{m}}(x;\gamma)\right)^2\right] \leq \frac{\exp(B/\gamma)}{\beta} \cdot \mathcal{E}_{\mathrm{m}}(s;\gamma,\gamma) \tag{60}$$

On the other hand, we also seek to bound $|\alpha(x)|$. Notice that we can state $|\alpha(x)|$ in terms of the logit function (*i.e.*, the inverse sigmoid σ^{-1}):

$$\begin{aligned}
|\alpha(x)| &= \left|\mathbb{E}_{\mathsf{Y}\sim\eta^\star(x)}\left[\ell(x,\mathsf{Y},h(x)) - \ell(x,\mathsf{Y},h^{(\mathrm{e})}(x))\right] - c\right| \\
&= \gamma \cdot \left|\frac{\mathbb{E}_{\mathsf{Y}\sim\eta^\star(x)}\left[\ell(x,\mathsf{Y},h(x)) - \ell(x,\mathsf{Y},h^{(\mathrm{e})}(x))\right]}{\gamma} - \frac{c}{\gamma}\right| \\
&= \gamma \cdot \left|\sigma^{-1}\left(p_{\mathrm{m}}(x;\gamma)\right) - \sigma^{-1}\left(\sigma\left(-\frac{c}{\gamma}\right)\right)\right|
\end{aligned}$$

where we are using Eq. (50).

Again by the boundedness of ℓ, we have that $p_{\mathrm{m}} \in [\sigma(-B/\gamma), \sigma(B/\gamma)]$. Furthermore, one will note that we have that both

$$p_{\mathrm{m}}, \sigma\left(-\frac{c}{\gamma}\right) \in \left[\sigma\left(-\frac{\max\{B,c\}}{\gamma}\right), 1 - \sigma\left(-\frac{\max\{B,c\}}{\gamma}\right)\right]. \tag{61}$$

Thus, by the differentiability of the logit function σ^{-1} and the mean-value-theorem, we have that

$$\begin{aligned}
|\alpha(x)| &\leq \frac{\gamma}{\sigma\left(-\frac{\max\{B,c\}}{\gamma}\right)\left(1 - \sigma\left(-\frac{\max\{B,c\}}{\gamma}\right)\right)} \cdot \left|p_{\mathrm{m}}(x;\gamma) - \sigma\left(-\frac{c}{\gamma}\right)\right| \\
&= K_{\gamma,B,c} \cdot \left|p_{\mathrm{m}}(x;\gamma) - \tau_{c,\gamma}\right|.
\end{aligned} \tag{62}$$

Combining Lemma G.4 and Eq. (62) yields

$$\begin{aligned}
\mathcal{L}(h, r_{c,\gamma}; c) - \mathcal{L}(h, r^\star; c) &= \mathbb{E}_{\mathsf{X}\sim\mathrm{P_x}}\left[|\alpha(\mathsf{X})| \cdot [\![r_{c,\gamma}(\mathsf{X}) \neq r^\star(\mathsf{X}; c, \mathrm{P})]\!]\right] \\
&\leq K_{\gamma,B,c} \cdot \mathbb{E}_{\mathsf{X}\sim\mathrm{P_x}}\left[|p_{\mathrm{m}}(x;\gamma) - \tau_{c,\gamma}| \cdot [\![r_{c,\gamma}(\mathsf{X}) \neq r^\star(\mathsf{X}; c, \mathrm{P})]\!]\right].
\end{aligned}$$

Finally, notice that as $r_{c,\gamma}(x) = [\![(\Phi^{-1} \circ s)(x) \le \tau_{c,\gamma}]\!]$, whenever $r_{c,\gamma}$ and $r^\star$ disagree, then $p_{\mathrm{m}}(x)$ and $(\Phi^{-1} \circ s)(x)$ must lie on the opposite sides of $\tau_{c,\gamma}$ (see Lemma G.3). Hence for x satisfying $r_{c,\gamma}(x) \neq r^\star(x; c, \mathrm{P})$, we have

$$|p_{\mathrm{m}}(x;\gamma) - \tau_{c,\gamma}| \le \left|p_{\mathrm{m}}(x;\gamma) - (\Phi^{-1} \circ s)(x)\right|.$$

Thus using Eq. (60) and Cauchy–Schwartz, we have

$$\begin{aligned}
&\mathcal{L}(h, r_{c,\gamma}; c) - \mathcal{L}(h, r^\star; c) \\
&\le K_{\gamma,B,c} \cdot \mathbb{E}_{\mathsf{X}\sim\mathrm{P}_{\mathsf{X}}}\left[|p_{\mathrm{m}}(x;\gamma) - \tau_{c,\gamma}| \cdot [\![r_{c,\gamma}(\mathsf{X}) \neq r^\star(\mathsf{X}; c, \mathrm{P})]\!]\right] \\
&\le K_{\gamma,B,c} \cdot \mathbb{E}_{\mathsf{X}\sim\mathrm{P}_{\mathsf{X}}}\left[\left|p_{\mathrm{m}}(x;\gamma) - (\Phi^{-1} \circ s)(x)\right| \cdot [\![r_{c,\gamma}(\mathsf{X}) \neq r^\star(\mathsf{X}; c, \mathrm{P})]\!]\right] \\
&\le K_{\gamma,B,c} \cdot \mathbb{E}_{\mathsf{X}\sim\mathrm{P}_{\mathsf{X}}}\left[\left|p_{\mathrm{m}}(x;\gamma) - (\Phi^{-1} \circ s)(x)\right|\right] \\
&\le K_{\gamma,B,c} \cdot \sqrt{\mathbb{E}_{\mathsf{X}\sim\mathrm{P}_{\mathsf{X}}}\left[\left(p_{\mathrm{m}}(x;\gamma) - (\Phi^{-1} \circ s)(x)\right)^2\right]} \\
&\le K_{\gamma,B,c} \cdot \sqrt{\frac{\exp(B/\gamma)}{\beta} \cdot \mathcal{E}_{\mathrm{m}}(s;\gamma,\gamma)} \\
&= K_{\gamma,B,c} \cdot \sqrt{\frac{\exp(B/\gamma)}{\beta} \cdot [\mathcal{E}_{\mathrm{j}}(s;\gamma,\gamma) + \mathcal{E}_{\mathrm{m}}(s;\gamma,\gamma) - \mathcal{E}_{\mathrm{j}}(s;\gamma,\gamma)]} \\
&= K_{\gamma,B,c} \cdot \sqrt{\frac{\exp(B/\gamma)}{\beta} \cdot \left[\mathcal{E}_{\mathrm{j}}(s;\gamma,\gamma) + \mathcal{L}^\star_{\mathrm{j}}(\gamma,\gamma) - \mathcal{L}^\star_{\mathrm{m}}(\gamma,\gamma) - \mathcal{L}_{\mathrm{j}}(s;\gamma,\gamma) + \mathcal{L}_{\mathrm{m}}(s;\gamma,\gamma)\right]} \\
&\le K_{\gamma,B,c} \cdot \sqrt{\frac{\exp(B/\gamma)}{\beta} \cdot [\mathcal{E}_{\mathrm{j}}(s;\gamma,\gamma) + \Delta_{\mathrm{mj}}(\gamma,\gamma)]}.
\end{aligned}$$

The final inequality follows from Lemma 4.3. □

G.4 Variance and Finite-Sample Bounds

In the following, we consider different methods for controlling the general approximate consistency upper bound presented in Theorem 4.4. Although we do not explicitly state it, the sample complexity discussion directly applies to the results of Theorem 4.5 (and its extended discussion in Appendix H).

Variance control of the approximation term. Under Theorem 4.4, we present a short discussion on how the marginal-joint ideal optimality gap can be upper bounded by a variance term. To do this, we leverage Theorem F.2.

Under the assumptions of Theorem G.5, let $s^\star_{\mathrm{m}} = \Phi \circ p_{\mathrm{m}}(\cdot;\gamma)$ be the marginal optimal scorer. Now short-hand its weighted variance:

$$V^\star_{\mathrm{m}} \stackrel{\mathrm{def}}{=} \mathbb{E}_{\mathsf{X}\sim\mathrm{P}_{\mathsf{X}}}\left[(\lambda_{+1} \circ s^\star_{\mathrm{m}})(\mathsf{X}) \cdot \mathbb{V}_{\eta^\star(\mathsf{X})}[\ell(\mathsf{X},\mathsf{Y},h(\mathsf{X}))] + (\lambda_{-1} \circ s^\star_{\mathrm{m}})(\mathsf{X}) \cdot \mathbb{V}_{\eta^\star(\mathsf{X})}[\ell(\mathsf{X},\mathsf{Y},h^{(\mathrm{e})}(\mathsf{X}))]\right]. \quad (63)$$

Thus, Theorem F.2 gives

$$0 \le \Delta_{\mathrm{mj}}(\gamma,\gamma) \le \mathcal{L}_{\mathrm{j}}(s^\star_{\mathrm{m}};\gamma,\gamma) - \mathcal{L}_{\mathrm{m}}(s^\star_{\mathrm{m}};\gamma,\gamma) \le \frac{V^\star_{\mathrm{m}}}{2\gamma^2}. \quad (64)$$

Note that the factor $1/\gamma^2$ can make the bound loose as $\gamma \to 0$. Thus, the low temperature connection in Proposition 4.6 does not establish near-optimality in that limit.

Finite-sample bounds. Through standard methods (*e.g.*, Rademacher complexity), we can convert the consistency bound in Theorem 4.4 (and Theorem 4.5) into finite sample bounds.

Fix the base model h, the expert $h^{(\mathrm{e})}$, and the temperatures (which may be unequal). Let $\mathsf{Z}_i = (\mathsf{X}_i, \mathsf{Y}_i)$ for $i \in \{1,\dots,n\}$ denote n i.i.d. samples from P. Denote a hypothesis class for the scorers by $\mathcal{S}$. We

define

$$g_s(x,y;\gamma,\gamma^{(\mathrm{e})}) \stackrel{\text{def}}{=} \exp\left(-\frac{\ell(x,y,h(x))}{\gamma}\right)(\lambda_{+1}\circ s)(x) + \exp\left(-\frac{\ell(x,y,h^{(\mathrm{e})}(x))}{\gamma^{(\mathrm{e})}}\right)(\lambda_{-1}\circ s)(x),$$

$$\mathcal{L}_{\mathrm{j}}(s;\gamma,\gamma^{(\mathrm{e})}) = \mathbb{E}_{\mathsf{Z}\sim\mathsf{P}}[g_s(\mathsf{Z};\gamma,\gamma^{(\mathrm{e})})], \qquad \widehat{\mathcal{L}}_{\mathrm{j}}(s;\gamma,\gamma^{(\mathrm{e})}) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} g_s(\mathsf{Z}_i;\gamma,\gamma^{(\mathrm{e})}).$$

Define the approximation error, empirical optimization error, and uniform estimation errors, respectively, by

$$A_{\mathcal{S},\gamma,\gamma^{(\mathrm{e})}} \stackrel{\text{def}}{=} \inf_{s\in\mathcal{S}} \mathcal{L}_{\mathrm{j}}(s;\gamma,\gamma^{(\mathrm{e})}) - \inf_{s} \mathcal{L}_{\mathrm{j}}(s;\gamma,\gamma^{(\mathrm{e})}),$$

$$\epsilon_{\mathrm{opt}}(\hat{s};\gamma,\gamma^{(\mathrm{e})}) \stackrel{\text{def}}{=} \widehat{\mathcal{L}}_{\mathrm{j}}(\widehat{s};\gamma,\gamma^{(\mathrm{e})}) - \inf_{s\in\mathcal{S}} \widehat{\mathcal{L}}_{\mathrm{j}}(s;\gamma,\gamma^{(\mathrm{e})}), \qquad \widehat{s}\in\mathcal{S},$$

$$U_{n,\gamma,\gamma^{(\mathrm{e})}} \stackrel{\text{def}}{=} \sup_{s\in\mathcal{S}} \left|\widehat{\mathcal{L}}_{\mathrm{j}}(s;\gamma,\gamma^{(\mathrm{e})}) - \mathcal{L}_{\mathrm{j}}(s;\gamma,\gamma^{(\mathrm{e})})\right|.$$

Let $\mathcal{G}_{\mathrm{j}} = \{g_s \colon s\in\mathcal{S}\}$. We define the (absolute) Rademacher complexity [Bartlett and Mendelson, 2002, Definition 2]

$$\mathfrak{R}_n(\mathcal{G}_{\mathrm{j}}) \stackrel{\text{def}}{=} \mathbb{E}_{\mathsf{Z}_{1:n},\varepsilon_{1:n}}\left[\sup_{s\in\mathcal{S}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i g_s(\mathsf{Z}_i)\right|\right], \tag{65}$$

where the ε_i are independent uniform signs. Define $\mathfrak{R}_n(\mathcal{S})$ analogously using $s(\mathsf{X}_i)$. All complexity constants below use the absolute Rademacher convention and $1/n$ normalization in Eq. (65).

We can now provide a standard excess risk upper bound consisting of the discussed errors.

Proposition G.6. *Suppose that, for every $s\in\mathcal{S}$, $g_s(\mathsf{Z})$ almost surely lies in a common interval of length $B_{\mathrm{j}} < \infty$. Then, for $\delta\in(0,1)$, with probability at least $1-\delta$,*

$$\mathcal{E}_{\mathrm{j}}(\widehat{s};\gamma,\gamma^{(\mathrm{e})}) \le A_{\mathcal{S},\gamma,\gamma^{(\mathrm{e})}} + \epsilon_{\mathrm{opt}}(\hat{s};\gamma,\gamma^{(\mathrm{e})}) + 4\mathfrak{R}_n(\mathcal{G}_{\mathrm{j}}) + 2B_{\mathrm{j}}\sqrt{\frac{\log(2/\delta)}{2n}}. \tag{66}$$

Proof. The proof is based on the strategy of Bartlett and Mendelson [2002, Theorem 8].

One can compute the standard upper bound on the excess risk as

$$\mathcal{E}_{\mathrm{j}}(\widehat{s};\gamma,\gamma^{(\mathrm{e})}) \le A_{\mathcal{S},\gamma,\gamma^{(\mathrm{e})}} + 2U_{n,\gamma,\gamma^{(\mathrm{e})}} + \epsilon_{\mathrm{opt}}(\hat{s};\gamma,\gamma^{(\mathrm{e})}).$$

The symmetrization and McDiarmid argument Bartlett and Mendelson [2002, Theorem 8], applied in both directions after rescaling the loss range to $[0,1]$ yields

$$U_{n,\gamma,\gamma^{(\mathrm{e})}} \le 2\mathfrak{R}_n(\mathcal{G}_{\mathrm{j}}) + B_{\mathrm{j}}\sqrt{\frac{\log(2/\delta)}{2n}}.$$

Combining the bounds yields the result. □

For the common temperature setting of Theorem 4.4, denote $r_{\widehat{s},c,\gamma}(x) = [\![(\Phi^{-1}\circ\widehat{s})(x) \le \tau_{c,\gamma}]\!]$. Combining Theorem 4.4 and Proposition G.6 yields

$$\mathcal{L}(h, r_{\widehat{s},c,\gamma}; c) - \mathcal{L}(h, r^\star; c) \tag{67}$$

$$\le K_{\gamma,B,c}\sqrt{\frac{\exp(B/\gamma)}{\beta}\left[A_{\mathcal{S},\gamma,\gamma} + \epsilon_{\mathrm{opt}}(\hat{s};\gamma,\gamma) + 4\mathfrak{R}_n(\mathcal{G}_{\mathrm{j}}) + 2B_{\mathrm{j}}\sqrt{\frac{\log(2/\delta)}{2n}} + \Delta_{\mathrm{mj}}(\gamma,\gamma)\right]}.$$

Scorer complexity. One can further simplify the bound in Eq. (67) if additional assumptions are made over the scorer hypothesis class $s\in\mathcal{S}$.

Suppose $|s| \le M$ uniformly over $s\in\mathcal{S}$. If $t\mapsto g_t(z)$ is L_g-Lipschitz on $[-M,M]$ and g_0 is bounded, we can use a contraction argument on $g_s - g_0$ to yield [Bartlett and Mendelson, 2002, Theorem 12(4)]

$$\mathfrak{R}_n(\mathcal{G}_{\mathrm{j}}) \le 2L_g\mathfrak{R}_n(\mathcal{S}) + \sqrt{\frac{\mathbb{E}[g_0(\mathsf{Z};\gamma)^2]}{n}}, \tag{68}$$

where $g_t(z)$ denotes the per-example loss with scalar score t.

It should be noted that $\mathfrak{R}_n(\mathcal{S})$ for standard model classes will shrink as the number of samples increase. For example, under appropriate assumptions, ReLU MLP can be shown to have its Rademacher term shrink at a rate of $\mathcal{O}(n^{-1/2})$ [Golowich et al., 2018, Theorem 1]. Under the square-root transfer of Theorem 4.4 (and Theorem 4.5), this yields an overall rate of $\mathcal{O}(n^{-1/4})$ for the corresponding term—a fast rate would require further assumptions. If $A_{\mathcal{S},\gamma,\gamma} = \epsilon_{\text{opt}}(\hat{s};\gamma,\gamma) = \Delta_{\text{mj}} = 0$, Eq. (67) gives an $n^{-1/4}$ L2D regret rate and the sufficient sample scaling $n = \mathcal{O}(\epsilon^{-4}\log(2/\delta))$ for regret at most ϵ.

H Exact Consistency of DR CPE Losses

In the following, we prove the exact consistency result of Theorem 4.5. To do so, we will construct a mapping between deferral costs c and temperatures $\gamma, \gamma^{(\text{e})}$ which will result in an exact consistency bound. Throughout this section, we consider ℓ to be the zero-one loss function. We first give a bound for strongly proper composite DR losses and then specialize to the LSIF [Kanamori et al., 2009].

H.1 Choice of loss functions

The LSIF loss function is described by the following partials [Kanamori et al., 2009]:

$$\lambda_{+1}(v) = -v; \qquad \text{and} \qquad \lambda_{-1}(v) = \frac{1}{2}v^2. \tag{69}$$

For the zero-one loss function, we define the following quantities:

$$p_h(x) \stackrel{\text{def}}{=} \mathbb{P}_{\text{P}}[f(\mathsf{X}) = \mathsf{Y} \mid \mathsf{X} = x]; \qquad \text{and} \qquad p_e(x) \stackrel{\text{def}}{=} \mathbb{P}_{\text{P}}[f^{(\text{e})}(\mathsf{X}) = \mathsf{Y} \mid \mathsf{X} = x]. \tag{70}$$

Notably, when considering a zero-one loss function for ℓ, the following conditional risks simplify:

$$\mathbb{E}_{\mathsf{Y}\sim\eta^\star(x)}\left[\ell(x,\mathsf{Y},h(x))\right] = 1 - p_h(x), \qquad \mathbb{E}_{\mathsf{Y}\sim\eta^\star(x)}\left[\ell(x,\mathsf{Y},h^{(\text{e})}(x))\right] = 1 - p_e(x).$$

As a result, we can simplify α in Lemma G.4 for the zero-one loss to

$$\alpha(x) = p_e(x) - p_h(x) - c. \tag{71}$$

Furthermore, we can simplify the joint ideal weights when ℓ is the zero-one loss function:

$$\begin{aligned} \tilde{w}_{\text{j}}(x;\gamma) &= \mathbb{E}_{\mathsf{Y}\sim\eta^\star(x)}\left[\exp\left(-\frac{\ell(x,\mathsf{Y},h(x))}{\gamma}\right)\right] \\ &= \kappa + (1-\kappa)\cdot p_h(x) \\ \tilde{w}^{(\text{e})}_{\text{j}}(x;\gamma^{(\text{e})}) &= \mathbb{E}_{\mathsf{Y}\sim\eta^\star(x)}\left[\exp\left(-\frac{\ell(x,\mathsf{Y},h^{(\text{e})}(x))}{\gamma^{(\text{e})}}\right)\right] \\ &= \kappa^{(\text{e})} + (1-\kappa^{(\text{e})})\cdot p_e(x), \end{aligned} \tag{72}$$

where for convenience we shorthand

$$\kappa \stackrel{\text{def}}{=} \exp(-1/\gamma), \qquad \text{and} \qquad \kappa^{(\text{e})} \stackrel{\text{def}}{=} \exp(-1/\gamma^{(\text{e})}). \tag{73}$$

These choices of loss functions match that of Proposition I.3.

H.2 Mapping between temperature and deferral costs

In the following, we provide a mapping between temperatures $\gamma, \gamma^{(\text{e})}$ and deferral costs c.

Lemma H.1. *Let $\gamma^{(\text{e})} > 0$ and $c \geq 0$. There exists a unique $\gamma \geq \gamma^{(\text{e})}$ such that*

$$\frac{\kappa}{1-\kappa} - \frac{\kappa^{(\text{e})}}{1-\kappa^{(\text{e})}} = c. \tag{74}$$

Proof. Letting $\frac{\exp(-1/\gamma^{(\mathrm{e})})}{1-\exp(-1/\gamma^{(\mathrm{e})})} + c$ be constant and solving for γ yields the desired result immediately. □

Lemma H.2. *Suppose that the temperature mapping of Lemma H.1 holds. Let ℓ be the zero-one loss function, (λ, Φ) be a strictly proper composite binary loss, and let $s_{\mathrm{j}}^\star$ minimize $\mathcal{L}_{\mathrm{j}}(s; \gamma, \gamma^{(\mathrm{e})})$. Then*

$$(\Phi^{-1} \circ s_{\mathrm{j}}^\star)(x) = \frac{\tilde{w}_{\mathrm{j}}(x;\gamma)}{\tilde{w}_{\mathrm{j}}(x;\gamma) + \tilde{w}_{\mathrm{j}}^{(\mathrm{e})}(x;\gamma^{(\mathrm{e})})} \tag{75}$$

and the following thresholding rejector is equivalent to Chow's Rule

$$\bar{r}^\star_{c,\gamma,\gamma^{(\mathrm{e})}}(x) \stackrel{\text{def}}{=} \left[\!\!\left[(\Phi^{-1} \circ s_{\mathrm{j}}^\star)(x) \leq \frac{\bar{\tau}_{c,\gamma,\gamma^{(\mathrm{e})}}}{1 + \bar{\tau}_{c,\gamma,\gamma^{(\mathrm{e})}}} \right]\!\!\right] = r^\star(x; c, \mathrm{P}) \tag{76}$$

where

$$\bar{\tau}_{c,\gamma,\gamma^{(\mathrm{e})}} \stackrel{\text{def}}{=} \frac{1-\kappa}{1-\kappa^{(\mathrm{e})}}. \tag{77}$$

Proof. We first prove Eq. (75) before showing the Chow's Rule equivalence. We follow a similar proof to Lemma G.2. Simplifying the risk (using the superscript "zo" to denote the usage of the zero-one task loss $\ell = \ell^{\mathrm{zo}}$), we have

$$\begin{aligned}
\mathcal{L}_{\mathrm{j}}^{\mathrm{zo}}(s; \gamma, \gamma^{(\mathrm{e})}) &= \mathbb{E}_{\mathrm{P}_{\mathsf{x}}}\left[\tilde{w}_{\mathrm{j}}(\mathsf{X};\gamma) \cdot (\lambda_{+1} \circ s)(\mathsf{X}) + \tilde{w}_{\mathrm{j}}^{(\mathrm{e})}(\mathsf{X};\gamma^{(\mathrm{e})}) \cdot (\lambda_{-1} \circ s)(\mathsf{X})\right] \\
&= \mathbb{E}_{\mathrm{P}_{\mathsf{x}}}\left[\left(\tilde{w}_{\mathrm{j}}(\mathsf{X};\gamma) + \tilde{w}_{\mathrm{j}}^{(\mathrm{e})}(\mathsf{X};\gamma^{(\mathrm{e})})\right) \cdot L\left(\frac{\tilde{w}_{\mathrm{j}}(\mathsf{X};\gamma)}{\tilde{w}_{\mathrm{j}}(\mathsf{X};\gamma) + \tilde{w}_{\mathrm{j}}^{(\mathrm{e})}(\mathsf{X};\gamma^{(\mathrm{e})})}, s(\mathsf{X}); \lambda\right)\right].
\end{aligned}$$

By strict properness, Eq. (75) yields the unique pointwise minimizer.

Now we examine the Chow's Rule equivalence. Assuming that the temperature mapping of Lemma H.1 holds, we can rewrite the following relationship between the joint ideal weights and the threshold:

$$\begin{aligned}
&\tilde{w}_{\mathrm{j}}(x;\gamma) - \bar{\tau}_{c,\gamma,\gamma^{(\mathrm{e})}} \cdot \tilde{w}_{\mathrm{j}}^{(\mathrm{e})}(x;\gamma^{(\mathrm{e})}) \\
&= \kappa + (1-\kappa) \cdot p_h(x) - \bar{\tau}_{c,\gamma,\gamma^{(\mathrm{e})}} \cdot \left(\kappa^{(\mathrm{e})} + (1-\kappa^{(\mathrm{e})}) \cdot p_e(x)\right) \\
&= \kappa - (1-\kappa) \cdot \frac{\kappa^{(\mathrm{e})}}{1-\kappa^{(\mathrm{e})}} + (1-\kappa) \cdot (p_h(x) - p_e(x)) \\
&= (1-\kappa)\left(\frac{\kappa}{1-\kappa} - \frac{\kappa^{(\mathrm{e})}}{1-\kappa^{(\mathrm{e})}} + p_h(x) - p_e(x)\right) \\
&= (1-\kappa)\left(c + p_h(x) - p_e(x)\right) \\
&= -(1-\kappa)\alpha(x),
\end{aligned}$$

where we are using Eq. (71) to identify α.

As the expert weights are strictly positive, we have

$$\frac{\tilde{w}_{\mathrm{j}}(x;\gamma)}{\tilde{w}_{\mathrm{j}}^{(\mathrm{e})}(x;\gamma^{(\mathrm{e})})} \leq \bar{\tau}_{c,\gamma,\gamma^{(\mathrm{e})}} \iff -(1-\kappa)\alpha(x) \leq 0 \iff \alpha(x) \geq 0.$$

Finally, we consider

$$\begin{aligned}
&\frac{\tilde{w}_{\mathrm{j}}(x;\gamma)}{\tilde{w}_{\mathrm{j}}(x;\gamma) + \tilde{w}_{\mathrm{j}}^{(\mathrm{e})}(x;\gamma^{(\mathrm{e})})} \leq \frac{\bar{\tau}_{c,\gamma,\gamma^{(\mathrm{e})}}}{1 + \bar{\tau}_{c,\gamma,\gamma^{(\mathrm{e})}}} \\
\iff\quad & \tilde{w}_{\mathrm{j}}(x;\gamma) + \bar{\tau}_{c,\gamma,\gamma^{(\mathrm{e})}} \cdot \tilde{w}_{\mathrm{j}}(x;\gamma) \leq \bar{\tau}_{c,\gamma,\gamma^{(\mathrm{e})}} \cdot \tilde{w}_{\mathrm{j}}(x;\gamma) + \bar{\tau}_{c,\gamma,\gamma^{(\mathrm{e})}} \cdot \tilde{w}_{\mathrm{j}}^{(\mathrm{e})}(x;\gamma^{(\mathrm{e})}) \\
\iff\quad & \tilde{w}_{\mathrm{j}}(x;\gamma) \leq \bar{\tau}_{c,\gamma,\gamma^{(\mathrm{e})}} \cdot \tilde{w}_{\mathrm{j}}^{(\mathrm{e})}(x;\gamma^{(\mathrm{e})}) \\
\iff\quad & \frac{\tilde{w}_{\mathrm{j}}(x;\gamma)}{\tilde{w}_{\mathrm{j}}^{(\mathrm{e})}(x;\gamma^{(\mathrm{e})})} \leq \bar{\tau}_{c,\gamma,\gamma^{(\mathrm{e})}} \iff \alpha(x) \geq 0.
\end{aligned} \tag{78}$$

Combined with Lemma 2.1, we see that thresholding the scorer as per Eq. (76) yields Chow's rule. □

H.3 Proof of Theorem 4.5 (Exact Consistency Bound)

The following provides a proof of Theorem 4.5 while also stating the formal statement. The exact relation between costs, temperatures, and thresholding parameters are explicitly given.

Theorem H.3. *Suppose that the loss ℓ is the zero-one loss function, the composite loss (λ, Φ) is β-strongly proper with $\beta > 0$, and the temperatures $\gamma, \gamma^{(e)}$ and deferral cost $c \geq 0$ satisfy the condition of Lemma H.1. Define the rejector*

$$\bar{r}_{c,\gamma,\gamma^{(e)}}(x) = \left[\!\!\left[(\Phi^{-1} \circ s)(x) \leq \frac{\bar{\tau}_{c,\gamma,\gamma^{(e)}}}{(1 + \bar{\tau}_{c,\gamma,\gamma^{(e)}})} \right]\!\!\right]. \tag{79}$$

Then

$$\mathcal{L}^{\text{zo}}(h, \bar{r}_{c,\gamma,\gamma^{(e)}}; c) - \mathcal{L}^{\text{zo}}(h, r^\star; c) \leq \frac{2}{\sqrt{\beta}} \left(\frac{1}{1-\kappa} + \frac{1}{1-\kappa^{(e)}} \right) \sqrt{\mathcal{E}^{\text{zo}}_{\text{j}}(s; \gamma, \gamma^{(e)})}, \tag{80}$$

where the superscript "zo" is to indicate we are considering the zero-one loss task loss $\ell = \ell^{\text{zo}}$.

Proof. For convenience, let $r = \bar{r}_{c,\gamma,\gamma^{(e)}}$ and denote

$$p(x) \stackrel{\text{def}}{=} \frac{\tilde{w}_{\text{j}}(x;\gamma)}{\tilde{w}_{\text{j}}(x;\gamma) + \tilde{w}^{(e)}_{\text{j}}(x;\gamma^{(e)})}. \tag{81}$$

From the proof of Lemma H.2 (and mirroring Eq. (58)), we have that

$$\begin{aligned}
\mathcal{E}^{\text{zo}}_{\text{j}}(s;\gamma,\gamma^{(e)}) &= \mathbb{E}_{\text{P}_{\text{x}}}\Big[\left(\tilde{w}_{\text{j}}(\mathsf{X};\gamma) + \tilde{w}^{(e)}_{\text{j}}(\mathsf{X};\gamma^{(e)}) \right) \\
&\qquad \cdot \left(L\left(p(\mathsf{X}), s(\mathsf{X}); \lambda\right) - L\left(p(\mathsf{X}), \Phi(p(\mathsf{X})); \lambda\right) \right) \Big] \\
&\geq \frac{\beta}{2} \cdot \mathbb{E}_{\text{P}_{\text{x}}} \left[\left(\tilde{w}_{\text{j}}(\mathsf{X};\gamma) + \tilde{w}^{(e)}_{\text{j}}(\mathsf{X};\gamma^{(e)}) \right) \cdot \left((\Phi^{-1} \circ s)(\mathsf{X}) - p(\mathsf{X}) \right)^2 \right],
\end{aligned}$$

where we use strong properness.

Further notice from a similar computation to Eq. (78), we have that

$$\begin{aligned}
\left(\tilde{w}_{\text{j}}(x;\gamma) + \tilde{w}^{(e)}_{\text{j}}(x;\gamma^{(e)}) \right) \cdot \left(p(x) - \frac{\bar{\tau}_{c,\gamma,\gamma^{(e)}}}{1 + \bar{\tau}_{c,\gamma,\gamma^{(e)}}} \right) &= \frac{\tilde{w}_{\text{j}}(x;\gamma) - \bar{\tau}_{c,\gamma,\gamma^{(e)}} \cdot \tilde{w}^{(e)}_{\text{j}}(x;\gamma^{(e)})}{1 + \bar{\tau}_{c,\gamma,\gamma^{(e)}}} \\
&= -\frac{1-\kappa}{1 + \bar{\tau}_{c,\gamma,\gamma^{(e)}}} \alpha(x).
\end{aligned}$$

Now, similar to the proof of Theorem G.5 we examine the values of $p(x)$ and $(\Phi^{-1} \circ s)(x)$ when $r^\star(x) \neq \bar{r}_{c,\gamma,\gamma^{(e)}}(x)$. Through Lemma H.2, the rejectors only disagree when $p(x)$ and $(\Phi^{-1} \circ s)(x)$ lie on opposite sides of the threshold, hence we have $|p(x) - \bar{\tau}_{c,\gamma,\gamma^{(e)}}/(1 + \bar{\tau}_{c,\gamma,\gamma^{(e)}})| \leq |p(x) - (\Phi^{-1} \circ s)(x)|$.

Together, we have

$$\begin{aligned}
&\mathcal{L}^{\text{zo}}(h, \bar{r}_{c,\gamma,\gamma^{(e)}}; c) - \mathcal{L}^{\text{zo}}(h, r^\star; c) \\
&= \mathbb{E}_{\mathsf{X} \sim \text{P}_{\text{x}}} \left[|\alpha(\mathsf{X})| \cdot [\![r(\mathsf{X}) \neq r^\star(\mathsf{X}; c, \text{P})]\!] \right] \\
&= \frac{1 + \bar{\tau}_{c,\gamma,\gamma^{(e)}}}{1-\kappa} \cdot \mathbb{E}_{\mathsf{X} \sim \text{P}_{\text{x}}} \Big[\left(\tilde{w}_{\text{j}}(\mathsf{X};\gamma) + \tilde{w}^{(e)}_{\text{j}}(\mathsf{X};\gamma^{(e)}) \right) \\
&\qquad \cdot \left| p(\mathsf{X}) - \frac{\bar{\tau}_{c,\gamma,\gamma^{(e)}}}{1 + \bar{\tau}_{c,\gamma,\gamma^{(e)}}} \right| \cdot [\![r(\mathsf{X}) \neq r^\star(\mathsf{X}; c, \text{P})]\!] \Big] \\
&\leq \frac{1 + \bar{\tau}_{c,\gamma,\gamma^{(e)}}}{1-\kappa} \cdot \mathbb{E}_{\mathsf{X} \sim \text{P}_{\text{x}}} \left[\left(\tilde{w}_{\text{j}}(\mathsf{X};\gamma) + \tilde{w}^{(e)}_{\text{j}}(\mathsf{X};\gamma^{(e)}) \right) \cdot \left| p(\mathsf{X}) - (\Phi^{-1} \circ s)(\mathsf{X}) \right| \right] \\
&\leq \frac{1 + \bar{\tau}_{c,\gamma,\gamma^{(e)}}}{1-\kappa} \cdot \sqrt{\frac{2}{\beta} \cdot \mathbb{E}_{\mathsf{X} \sim \text{P}_{\text{x}}} \left[\tilde{w}_{\text{j}}(\mathsf{X};\gamma) + \tilde{w}^{(e)}_{\text{j}}(\mathsf{X};\gamma^{(e)}) \right] \cdot \mathcal{E}^{\text{zo}}_{\text{j}}(s;\gamma,\gamma^{(e)})} \\
&\leq \frac{1 + \bar{\tau}_{c,\gamma,\gamma^{(e)}}}{1-\kappa} \cdot \sqrt{\frac{4}{\beta} \mathcal{E}^{\text{zo}}_{\text{j}}(s;\gamma,\gamma^{(e)})}.
\end{aligned}$$

We use Cauchy–Schwarz and strong-properness. The final statement follows from a simplification of $\frac{1+\bar{\tau}_{c,\gamma,\gamma^{(\mathrm{e})}}}{1-\kappa}$. □

We can obtain a sharper bound without using strong properness by considering the LSIF loss function.

Corollary H.4. *Suppose that the loss ℓ is the zero-one loss function, λ is the LSIF loss in Eq. (69), and the temperatures $\gamma, \gamma^{(\mathrm{e})}$ and deferral cost $c \geq 0$ satisfy the condition of Lemma H.1. Define the rejector*

$$\bar{r}_{c,\gamma,\gamma^{(\mathrm{e})}}(x) \stackrel{\text{def}}{=} [\![s(x) \leq \bar{\tau}_{c,\gamma,\gamma^{(\mathrm{e})}}]\!] . \tag{82}$$

Then,

$$\mathcal{L}^{\mathrm{zo}}(h, \bar{r}_{c,\gamma,\gamma^{(\mathrm{e})}}; c) - \mathcal{L}^{\mathrm{zo}}(h, r^\star; c) \leq \frac{\sqrt{2}}{1-\exp(-1/\gamma)} \sqrt{\mathcal{E}^{\mathrm{LSIF}}_{\mathrm{j}}(s; \gamma, \gamma^{(\mathrm{e})})}, \tag{83}$$

*where the superscript "*LSIF*" is to indicate we are considering the LSIF density-ratio loss.*

Proof. Given the LSIF partial losses in Eq. (69), we have that

$$\mathcal{L}^{\mathrm{LSIF}}_{\mathrm{j}}(s; \gamma, \gamma^{(\mathrm{e})}) = \mathbb{E}_{\mathsf{X} \sim \mathrm{P}_{\mathrm{x}}} \left[-\tilde{w}_{\mathrm{j}}(\mathsf{X}; \gamma) \cdot s(\mathsf{X}) + \frac{1}{2} \tilde{w}^{(\mathrm{e})}_{\mathrm{j}}(\mathsf{X}; \gamma^{(\mathrm{e})}) \cdot s(\mathsf{X})^2 \right].$$

Completing the square yields the unique pointwise optimal scorer:

$$s^\star_{\mathrm{j}}(x) = \frac{\tilde{w}_{\mathrm{j}}(x; \gamma)}{\tilde{w}^{(\mathrm{e})}_{\mathrm{j}}(x; \gamma^{(\mathrm{e})})}.$$

Thus we have the excess risk

$$\begin{aligned} \mathcal{E}^{\mathrm{LSIF}}_{\mathrm{j}}(s; \gamma, \gamma^{(\mathrm{e})}) &= \mathcal{L}^{\mathrm{LSIF}}_{\mathrm{j}}(s; \gamma, \gamma^{(\mathrm{e})}) - \mathcal{L}^{\mathrm{LSIF}}_{\mathrm{j}}(s^\star_{\mathrm{j}}; \gamma, \gamma^{(\mathrm{e})}) \\ &= \frac{1}{2} \cdot \mathbb{E}_{\mathsf{X} \sim \mathrm{P}_{\mathrm{x}}} \left[\tilde{w}^{(\mathrm{e})}_{\mathrm{j}}(\mathsf{X}; \gamma^{(\mathrm{e})}) \cdot \left(s(\mathsf{X}) - s^\star_{\mathrm{j}}(\mathsf{X}) \right)^2 \right]. \end{aligned} \tag{84}$$

From the proof of Lemma H.2, we have

$$\begin{aligned} \tilde{w}^{(\mathrm{e})}_{\mathrm{j}}(x; \gamma^{(\mathrm{e})}) \cdot \left(s^\star_{\mathrm{j}}(x) - \bar{\tau}_{c,\gamma,\gamma^{(\mathrm{e})}} \right) &= \tilde{w}_{\mathrm{j}}(x; \gamma) - \bar{\tau}_{c,\gamma,\gamma^{(\mathrm{e})}} \cdot \tilde{w}^{(\mathrm{e})}_{\mathrm{j}}(x; \gamma^{(\mathrm{e})}) \\ &= -(1-\kappa)\alpha(x). \end{aligned}$$

Thus, thresholding $s^\star_{\mathrm{j}}$ at $\bar{\tau}_{c,\gamma,\gamma^{(\mathrm{e})}}$ yields Chow's rule Lemma 2.1. Whenever $\bar{r}_{c,\gamma,\gamma^{(\mathrm{e})}}(x) \neq r^\star(x; c, \mathrm{P})$, the scores $s(x)$ and $s^\star_{\mathrm{j}}(x)$ must lie on the opposite sides of the threshold. Hence $\left| s^\star_{\mathrm{j}}(x) - \bar{\tau}_{c,\gamma,\gamma^{(\mathrm{e})}} \right| \leq \left| s(x) - s^\star_{\mathrm{j}}(x) \right|$. Applying Lemma G.4, followed by Cauchy–Schwarz, and Eq. (84), we have that

$$\begin{aligned} &\mathcal{L}^{\mathrm{zo}}(h, \bar{r}_{c,\gamma,\gamma^{(\mathrm{e})}}; c) - \mathcal{L}^{\mathrm{zo}}(h, r^\star; c) \\ &\leq \frac{1}{1-\kappa} \cdot \mathbb{E}_{\mathsf{X} \sim \mathrm{P}_{\mathrm{x}}} \left[\tilde{w}^{(\mathrm{e})}_{\mathrm{j}}(\mathsf{X}; \gamma^{(\mathrm{e})}) \cdot \left| s(\mathsf{X}) - s^\star_{\mathrm{j}}(\mathsf{X}) \right| \right] \\ &\leq \frac{1}{1-\kappa} \cdot \sqrt{\mathbb{E}_{\mathsf{X} \sim \mathrm{P}_{\mathrm{x}}} \left[\tilde{w}^{(\mathrm{e})}_{\mathrm{j}}(\mathsf{X}; \gamma^{(\mathrm{e})}) \right]} \cdot \sqrt{\mathbb{E}_{\mathsf{X} \sim \mathrm{P}_{\mathrm{x}}} \left[\tilde{w}^{(\mathrm{e})}_{\mathrm{j}}(\mathsf{X}; \gamma^{(\mathrm{e})}) \cdot \left(s(\mathsf{X}) - s^\star_{\mathrm{j}}(\mathsf{X}) \right)^2 \right]} \\ &\leq \frac{\sqrt{2}}{1-\kappa} \cdot \sqrt{\mathcal{E}^{\mathrm{LSIF}}_{\mathrm{j}}(s; \gamma, \gamma^{(\mathrm{e})})}. \end{aligned}$$

As required. □

I Connections to Expert-Comparison Estimation

The joint ideal DR CPE losses connect to various expert-comparison estimation approaches. In the following, we prove Proposition 4.6 and also consider a few different instantiations of ℓ and λ which yields other interesting expert-comparison estimation approaches.

Lemma I.1. *Let $c \geq 0$. Then*

$$\lim_{\gamma \to 0^+} \exp\left(-\frac{c}{\gamma}\right) = \begin{cases} 0 & c > 0, \\ 1 & c = 0. \end{cases}$$

Furthermore,

$$\lim_{\gamma \to \infty} \exp\left(-\frac{c}{\gamma}\right) = 1.$$

Proof. If $c > 0$, then $c/\gamma \to \infty$ as $\gamma \to 0^+$, and therefore $-c/\gamma \to -\infty$. Since $\exp(x) \to 0$ as $x \to -\infty$, we obtain

$$\lim_{\gamma \to 0^+} \exp\left(-\frac{c}{\gamma}\right) = 0.$$

If instead $c = 0$, then for every $\gamma > 0$ we have

$$\exp\left(-\frac{c}{\gamma}\right) = \exp(0) = 1,$$

and hence

$$\lim_{\gamma \to 0^+} \exp\left(-\frac{c}{\gamma}\right) = 1.$$

Finally, as $\gamma \to \infty$ we have $c/\gamma \to 0$, so by continuity of the exponential function,

$$\lim_{\gamma \to \infty} \exp\left(-\frac{c}{\gamma}\right) = \exp(0) = 1.$$

□

Corollary I.2. *Suppose that ℓ the zero-one loss function. Then*

$$\lim_{\gamma \to 0^+} \mathcal{L}_{\mathrm{j}}(s; \gamma, \gamma) = \mathbb{E}_{\mathrm{P}}\left[[\![f(\mathsf{X}) = \mathsf{Y}]\!] \cdot (\lambda_{+1} \circ s)(\mathsf{X}) + [\![f^{(\mathrm{e})}(\mathsf{X}) = \mathsf{Y}]\!] \cdot (\lambda_{-1} \circ s)(\mathsf{X})\right],$$

Proof. As ℓ is the zero-one loss, for every (x, y) we have

$$\ell(x, y, h(x)) = 1 - [\![f(x) = y]\!], \qquad \ell(x, y, h^{(\mathrm{e})}(x)) = 1 - [\![f^{(\mathrm{e})}(x) = y]\!].$$

Hence, by the preceding Lemma I.1,

$$\lim_{\gamma \to 0^+} \exp\left(-\frac{\ell(\mathsf{X}, \mathsf{Y}, h(\mathsf{X}))}{\gamma}\right) = [\![f(\mathsf{X}) = \mathsf{Y}]\!],$$
$$\lim_{\gamma \to 0^+} \exp\left(-\frac{\ell(\mathsf{X}, \mathsf{Y}, h^{(\mathrm{e})}(\mathsf{X}))}{\gamma}\right) = [\![f^{(\mathrm{e})}(\mathsf{X}) = \mathsf{Y}]\!].$$

Therefore, from Eq. (14),

$$\begin{aligned} \lim_{\gamma \to 0^+} \mathcal{L}_{\mathrm{j}}(s; \gamma, \gamma) = \lim_{\gamma \to 0^+} \mathbb{E}_{\mathrm{P}} \Bigg[& \exp\left(-\frac{\ell(\mathsf{X}, \mathsf{Y}, h(\mathsf{X}))}{\gamma}\right) (\lambda{+}1 \circ s)(\mathsf{X}) \\ & + \exp\left(-\frac{\ell(\mathsf{X}, \mathsf{Y}, h^{(\mathrm{e})}(\mathsf{X}))}{\gamma}\right) (\lambda{-}1 \circ s)(\mathsf{X}) \Bigg] \\ = \mathbb{E}_{\mathrm{P}} & \left[[\![f(\mathsf{X}) = \mathsf{Y}]\!] \cdot (\lambda_{+1} \circ s)(\mathsf{X}) + [\![f^{(\mathrm{e})}(\mathsf{X}) = \mathsf{Y}]\!] \cdot (\lambda_{-1} \circ s)(\mathsf{X}) \right], \end{aligned}$$

where we used dominated convergence in the last step. □

I.1 Proof of Proposition 4.6

Proof of Proposition 4.6. For the squared loss, the partial losses are [Menon and Ong, 2016]

$$\lambda_{+1}(v) = (v-1)^2, \qquad \lambda_{-1}(v) = (v+1)^2.$$

Therefore, from Corollary I.2

$$\lim_{\gamma \to 0^+} \mathcal{L}_{\mathrm{j}}(s;\gamma,\gamma) = \mathbb{E}_{\mathsf{P}}\left[[\![f(\mathsf{X}) = \mathsf{Y}]\!](s(\mathsf{X})-1)^2 + [\![f^{(\mathrm{e})}(\mathsf{X}) = \mathsf{Y}]\!](s(\mathsf{X})+1)^2\right].$$

To identify the optimal scorer of the limiting objective, we condition on $\mathsf{X} = x$ and consider the pointwise conditional risk

$$\begin{aligned} R_x(v) &= \mathbb{E}\left[[\![f(\mathsf{X}) = \mathsf{Y}]\!](v-1)^2 + [\![f^{(\mathrm{e})}(\mathsf{X}) = \mathsf{Y}]\!](v+1)^2 \mid \mathsf{X} = x\right] \\ &= p_h(x)(v-1)^2 + p_e(x)(v+1)^2, \end{aligned}$$

where

$$p_h(x) = \mathbb{P}_{\mathsf{P}}[f(\mathsf{X}) = \mathsf{Y} \mid \mathsf{X} = x], \qquad p_e(x) = \mathbb{P}_{\mathsf{P}}[f^{(\mathrm{e})}(\mathsf{X}) = \mathsf{Y} \mid \mathsf{X} = x].$$

Equivalently, with

$$\Delta(x) = \mathbb{E}[\mathsf{Y}^{(\mathrm{d})} \mid \mathsf{X} = x] = p_h(x) - p_e(x), \qquad \Sigma(x) = p_h(x) + p_e(x),$$

we can write

$$R_x(v) = \Sigma(x)v^2 - 2\Delta(x)v + \Sigma(x).$$

Therefore, whenever $\Sigma(x) > 0$, the unique minimizer satisfies

$$\frac{\partial R_x(v)}{\partial v} = 2\Sigma(x)v - 2\Delta(x) = 0,$$

and hence

$$s_0^\star(x) = \frac{\Delta(x)}{\Sigma(x)}.$$

If $\Sigma(x) = 0$, then $R_x(v) = 0$ for all v, so any value of v is optimal. Finally, for any $\tau \in \mathbb{R}$ and any x with $\Sigma(x) > 0$, we have

$$s_0^\star(x) \le \tau \iff \Delta(x) \le \tau\,\Sigma(x),$$

as claimed. □

I.2 Ratio Expert Comparison

One interesting variant of Proposition 4.6 is when we change the exact DRE loss λ to other well-known DRE loss functions.

Proposition I.3. *Suppose ℓ is the zero-one loss function and λ is the LSIF [Kanamori et al., 2009] loss function. Then the unique pointwise minimizer $s_0^\star(x)$ of the limiting risk satisfies, for any $\tau \in \mathbb{R}$,*

$$[\![s_0^\star(x) \le \tau]\!] = \left[\!\!\left[\mathbb{P}_{\mathsf{P}}[f(\mathsf{X}) = \mathsf{Y} \mid \mathsf{X} = x] \le \tau \mathbb{P}_{\mathsf{P}}[f^{(\mathrm{e})}(\mathsf{X}) = \mathsf{Y} \mid \mathsf{X} = x]\right]\!\!\right]. \tag{85}$$

Proof. For the LSIF loss function, the partial losses are Menon and Ong [2016],

$$\lambda_{+1}(v) = -v, \qquad \lambda_{-1}(v) = \frac{1}{2}v^2.$$

Mirroring the proof of Proposition 4.6, we have the pointwise condition risk

$$R_x(v) = -p_h(x)v + \frac{1}{2}p_e(x)v^2.$$

Therefore, taking the minimizer, we have that

$$\frac{\partial R_x(v)}{\partial v} = -p_h(x) + p_e(x)v = 0,$$

and hence,

$$s_0^\star(x) = \frac{p_h(x)}{p_e(x)}.$$

The result follows immediately. □

Proposition I.4. *Suppose ℓ is the zero-one loss function and λ is the KLIEP [Sugiyama et al., 2008] loss function. Then the unique pointwise minimizer $s_0^\star(x)$ of the limiting risk satisfies, for any $\tau \in \mathbb{R}$,*

$$[\![s_0^\star(x) \leq \tau]\!] = \left[\!\!\left[\mathbb{P}_{\mathsf{P}}[f(\mathsf{X}) = \mathsf{Y} \mid \mathsf{X} = x] \leq \tau \mathbb{P}_{\mathsf{P}}[f^{(\mathrm{e})}(\mathsf{X}) = \mathsf{Y} \mid \mathsf{X} = x]\right]\!\!\right]. \tag{86}$$

Proof. For the KLIEP loss function, the partial losses are Menon and Ong [2016],

$$\lambda_{+1}(v) = -\log v, \qquad \lambda_{-1}(v) = v$$

Mirroring the proof of Proposition 4.6, we have the pointwise condition risk

$$R_x(v) = -p_h(x) \log v + \frac{1}{2} p_e(x) v.$$

Therefore, taking the minimizer, we have that

$$\frac{\partial R_x(v)}{\partial v} = -\frac{1}{v} p_h(x) + p_e(x) = 0,$$

and hence,

$$s_0^\star(x) = \frac{p_h(x)}{p_e(x)}.$$

The result follows immediately. □

It is particularly interesting that two of (arguably) the most well-known DRE losses [Sugiyama et al., 2008, Kanamori et al., 2009] yield the same optimal deferral mechanism, as per Propositions I.3 and I.4. This is an interesting variant of the usual "difference" which is optimized for via Eq. (19). The optimal scorer is also very intuitive: we want to reject whenever our base model's (conditional) accuracy is worse than what the expert would have (up to discount scaling τ).

The logistic loss also yields a similar optimal deferral mechanism.

Proposition I.5. *Suppose ℓ is the zero-one loss function and λ is the logistic loss function. Then the unique pointwise minimizer $s_0^\star(x)$ of the limiting risk satisfies, for any $\tau \in \mathbb{R}$,*

$$[\![s_0^\star(x) \leq \tau]\!] = \left[\!\!\left[\mathbb{P}_{\mathsf{P}}[f(\mathsf{X}) = \mathsf{Y} \mid \mathsf{X} = x] \leq \exp(\tau) \mathbb{P}_{\mathsf{P}}[f^{(\mathrm{e})}(\mathsf{X}) = \mathsf{Y} \mid \mathsf{X} = x]\right]\!\!\right]. \tag{87}$$

Proof. For the logistic loss function, the partial losses are Menon and Ong [2016],

$$\lambda_{+1}(v) = \log(1 + \exp(-v)), \qquad \lambda_{-1}(v) = \log(1 + \exp(v))$$

Mirroring the proof of Proposition 4.6, we have the pointwise condition risk

$$R_x(v) = p_h(x) \log(1 + \exp(-v)) + p_e(x) \log(1 + \exp(v)).$$

Therefore, taking the minimizer, we have that

$$\begin{aligned}
\frac{\partial R_x(v)}{\partial v} &= -\frac{\exp(-v)}{1 + \exp(-v)} p_h(x) + \frac{\exp(v)}{1 + \exp(v)} p_e(x) \\
&= -\frac{1}{1 + \exp(v)} p_h(x) + \frac{\exp(v)}{1 + \exp(v)} p_e(x) \\
&= 0.
\end{aligned}$$

Thus we have

$$-p_h(x) + \exp(v) p_e(x) = 0$$

and hence,

$$s_0^\star(x) = \log\left(\frac{p_h(x)}{p_e(x)}\right).$$

The result follows immediately. □

The optimal deferral mechanisms are essentially the same ratio thresholds reparameterized when comparing Propositions I.3 and I.4 versus Proposition I.5. Despite this, as in practice we are thresholding the scorers (as per Lemma 4.1), KLIEP and LSIF may be more preferable: a small change in the scorer for the logistic loss (Proposition I.5) can cause a large shift in the decision w.r.t. the accuracy of base versus expert.

J Experiment Settings

The following section details the experimental setup in more details. All implementations were done using `PyTorch` for models and training; and `tensorflow_datasets` (`tfds`) to manage datasets. The CIFAR-100 and MedMNIST experiments were completed via a single NVIDIA Tesla V100. The ImageNet experiments are described in Appendix L. To produce error bars, we repeat deferral model training (keeping the base and expert models fixed) over 11 random seeds. The `Conf` baseline does not have error bars as it is derived directly from the base model's confidence.

J.1 Base Datasets

We consider three datasets:

- CIFAR-100 (no license): The standard vision dataset consisting of 50,000 training images and 10,000 test images with 100 classes.
- PathMNIST (CC BY 4.0): A colon pathology image dataset consisting of 89,996 training images; 10,004 validation images; and 7,180 test images. The dataset has 9 class labels.
- DermaMNIST (CC BY-NC 4.0): A smaller dermatoscopy image dataset consisting of 7,007 training images; 1,003 validation images; and 2,005 test images. The dataset has 7 class labels.
- ImageNet (custom research license): A standard vision dataset consisting of 1,281,167 training images and 50,000 validation images with 1,000 classes.

As CIFAR-100 does not have an explicit validation set, we follow Jitkrittum et al. [2023] and generate a validation set consisting of a subselection of 5,000 examples from the training set. In particular, we subsample 50 examples for each of the 100 classes to yield 45,000 training examples and 5,000 validation examples.

J.2 Corruptions

We consider three types of corruptions:

- Label noise: For $k \in \{10, 25\}$, the first k classes are randomly relabeled.
- Long-tail: For $h \in \{50, 25\}$, we specify the first h classes as head classes, and the remaining classes as tail classes. We then subsample the dataset with 500 examples for the head class and 50 examples for the tail class.
- Specialist: We specify a list of specialists class labels. All examples for the specialist class are taken. For all other classes, we subsample $p \in \{20, 10\}$ percent of their examples.

In the label noise and long-tail settings, subsampling are applied to the training partitions (used by both h and $h^{(e)}$). In the specialist setting, h is trained on the clean training partition, while only $h^{(e)}$ subsamples examples from non-specialist classes. In each setting, the held-out partition used to train the deferral scorer and the test set remain clean.

For the specialist class labels, we use the following for each dataset:

- CIFAR-100: The specialist is on mammals, which consists of 25/100 classes. This can be depicted by the following class label dictionary:

```
MAMMAL_SPECIALIST_FINE_LABELS_BY_COARSE = {
  "aquatic_mammals": ("beaver", "dolphin", "otter", "seal", "whale"),
  "large_carnivores": ("bear", "leopard", "lion", "tiger", "wolf"),
  "large_omnivores_and_herbivores": (
    "camel",
    "cattle",
    "chimpanzee",
    "elephant",
    "kangaroo",
  ),
  "medium_mammals": ("fox", "porcupine", "possum", "raccoon", "skunk"),
  "small_mammals": ("hamster", "mouse", "rabbit", "shrew", "squirrel"),
}
```

- PathMNIST: The specialist is on class labels associated with tumors, which consists of 3/9 classes. Its labels are given by:

```
PATH_SPECIALIST = [
  "lymphocytes",
  "cancer-associated stroma",
  "colorectal adenocarcinoma epithelium"
]
```

- DermaMNIST: The specialist is on class labels associated with melanocytic lesions, which consists of 2/7 classes. Its labels are given by:

```
DERMA_SPECIALIST = [
  "melanoma",
  "melanocytic nevi",
]
```

J.3 Parameterized Deferral Models

We follow Jitkrittum et al. [2023]'s post-hoc deferral-model setup for all learned deferral baselines. In all such methods, the base classifier h and expert $h^{(e)}$ are first trained and then frozen; only a separate deferral scorer is fit on the clean deferral-training split.

The parameterized deferral methods learn a lightweight MLP that takes features derived solely from the base model's predictive probability vector $\eta(x) \in \Delta(\mathcal{Y})$. Following Jitkrittum et al. [2023], we use a feature map $v(\eta(x)) \in \mathbb{R}^{|\mathcal{Y}|+11}$ consisting of:

1. The entropy of $\eta(x)$;
2. The top-10 predicted probabilities; and
3. A one-hot encoding of the predicted class $\arg\max_{y'} \eta_{y'}(x)$.

The learned deferral mechanism uses an architecture of

$$g(x) = (\mathrm{FC}_1 \circ \mathrm{FC}_{2^4,\mathrm{ReLU}} \circ \mathrm{FC}_{2^6,\mathrm{ReLU}} \circ v \circ \eta)(x),$$

where FC_k denotes a fully connected layer with k outputs. A rejector then is derived from a thresholding on g (which depends on the approach).

J.4 Optimizers

We use a combination of SGD and Adam [Kingma and Ba, 2014] optimizers.

- Base models
  - All base classifiers are trained with SGD using momentum 0.9 and weight decay 10^{-4}.
  - CIFAR-100: learning rate 1.0, with linear warmup for 15 epochs, then multi-step decay at epochs $\{96, 192, 224\}$. Models are trained for 256 epochs.
  - PathMNIST and DermaMNIST: learning rate 0.1, with linear warmup for 5 epochs, then multi-step decay at epochs $\{60, 90, 110\}$. Models are trained for 128 epochs.
  - ImageNet: MobileNetV2 [Sandler et al., 2018] uses a learning rate of 0.05, with linear warmup for 5 epochs, then multi-step decay at epochs $\{30, 60, 80\}$; EfficientNet-B0 [Tan and Le, 2019] uses a learning rate of 0.1 with cosine annealing over 128 epochs. Both models are trained for 128 epochs.
- Deferral models
  - All learned deferral models use Adam.
  - The optimizer settings are consistent across methods: learning rate 7×10^{-4} and weight decay 10^{-3}.
  - Deferral models are trained for 32 epochs.

J.5 Evaluation

To produce the curves in Section 5, we sort the decision scores and defer the first k examples, with smaller scores favoring deferral. Ties retain the evaluation order through a stable sort, and tabulated values are interpolated at the specified rates. Each curve keeps the scorer and its training hyperparameters fixed. The `Random` baseline is the expected accuracy under uniform random routing, *i.e.*, the linear interpolation between the base and expert accuracies.

J.6 Baselines

We go into more detail regarding the baselines we consider.

Confidence-Based Deferral Perhaps the simplest approach to deferral is to utilize the base model h's perceived confidence of its prediction. At inference time, one can simply threshold this confidence (`Conf`):

$$\text{Conf}(x) = \max_y h_y(x), \tag{88}$$

that is, $[\![\text{Conf}(x) \leq \tau]\!]$ for some threshold $\tau \in [0, 1]$.

Expert-Comparison Estimation These approaches are based on simple prediction targets (typically trained via regression, *i.e.*, squared loss functions) [Kag et al., 2022, Jitkrittum et al., 2023]. As a result, we end up with a model g that approximates a constructed random variable. Deferral can then be established by thresholding g, *i.e.*, $r(x) = [\![d(g(x)) \leq \tau]\!]$ for some τ which we can tune and d is a decision function.

- **Difference in zero-one accuracy** (`Estimate-Diff01`): This baseline considers the difference in zero-one accuracy. This is exactly the target discussed in Eq. (19). For completeness, we repeat the target random variable that is defined:

  $$\mathsf{Y}^{(\mathrm{d})} \stackrel{\text{def}}{=} [\![f(\mathsf{X}) = \mathsf{Y}]\!] - [\![f^{(\mathrm{e})}(\mathsf{X}) = \mathsf{Y}]\!].$$

  Thus, negative values correspond to examples on which the expert is correct and the base model is not. The decision function in this case is just the identity $d(z) = z$.
- **Expert maximum confidence** (`Estimate-MaxProb`) This baseline regresses the expert's maximum predicted class probability:

  $$\mathsf{P}^{(\mathrm{m})} \stackrel{\text{def}}{=} \max_{y \in \mathcal{Y}} \eta_y^{(\mathrm{e})}(\mathsf{X}). \tag{89}$$

  At test time, the learned score is compared against the base model's maximum confidence,

  $$d(g(x)) = \max_{y \in \mathcal{Y}} \eta_y(x) - g(x), \tag{90}$$

  where we remind that η is the base model and $g(x)$ approximates $\max_{y \in \mathcal{Y}} \eta_y^{(\mathrm{e})}(\mathsf{X})$ via Eq. (89). Thus, deferral is favored when the expert appears more confident than the base model.

Two-Stage Surrogate Losses We adapt Mao et al. [2023] for single expert deferral. In particular, we utilize their modified exponential loss function, which is given by (`TwoStage-Exp`):

$$\begin{aligned} \mathcal{L}_{\texttt{two-stage}}(s; c) = \mathbb{E}_{\mathsf{P}} \Bigg[& [\![f(\mathsf{X}) = \mathsf{Y}]\!] \exp\left(s(\mathsf{X}) - \max_y h_y(\mathsf{X}) \right) \\ & + ([\![f^{(\mathrm{e})}(\mathsf{X}) = \mathsf{Y}]\!] - c) \exp\left(\max_y h_y(\mathsf{X}) - s(\mathsf{X}) \right) \Bigg]. \end{aligned} \tag{91}$$

Typically, one defers from thresholding s to the base model's confidence, *i.e.*, $[\![s(x) > \max_y h_y(x)]\!]$. We consider a modification by thresholding over additional values to create a deferral curve $[\![s(x) > \max_y h_y(x) + \tau]\!]$ for $\tau \in \mathbb{R}$. For the reported comparisons, we use $c = 0.05$ and vary τ with the scorer fixed.

K Extended Experiments

In the following, we present an extension of the experiments presented in Section 5. For corrupted datasets, in this appendix section, we signify the strength of the corruption via parenthesis, *e.g.*, "PathMNIST Specialist(20)" retains 20% of the non-specialist training examples for the expert, together with all specialist training examples.

Table 1 presents a tabulated version of Fig. 2. We highlight approaches which are closer to random deferral (the linear interpolation between 0% and 100% deferral rates). As one can see, although DR CPE approaches may not be the best approach for all dataset and deferral combination, it is the most robust to failure (least red) for even its Prob01 approach.

We test over alternate ℓ's used to define the weights in Eq. (14): we consider probabilistic zero-one loss; Generalized Cross Entropy (GCE) [Zhang and Sabuncu, 2018]; and a top-k ($k = 5$) zero-one loss. These results are summarized in Fig. 3. From these plots, we see that for some loss functions, like the probabilistic zero-one loss, there are setting where a higher $\gamma = 1$ temperature can cause a reduction in performance, see PathMNIST Specialist(20). GCE has better performance in the $\gamma = 1$ setting, but it suffers from a significant drop in performance compared to the $\gamma = 0.5$ counterpart and also has a high variance (similar to Prob01 and top-k variants for PathMNIST Specialist(20) with $\gamma = 0.5$).

We also consider changing the target evaluation criterion from accuracy to top-k ($k = 5$). These results are summarized in Fig. 4—we focus on CIFAR-100 for this scenario due to its larger number of ($|\mathcal{Y}| = 100$) class labels. An interesting finding of these results is that performance in accuracy does not necessarily result in the corresponding top-k performance, nor vice-versa. Some of the most dramatic changes here are the performance of the `TwoStage-Exp` approach, which does much better in top-k. This is even when compared against DR CPE with objective specified to top-k accuracy. Despite this, the top-k variant of DR CPE performs well against its GCE variant for specialist CIFAR-100 settings.

Finally, we consider a variation of Fig. 2, where we vary the density-ratio loss λ utilized. From Propositions I.3 to I.5, the Bayes-optimal scorer (and thus deferral curves) should be the same. However, in practice we find that the logistic loss performs consistently worse when compared to other loss functions and suffers from an extremely high variance. The squared loss seems to be the best performing with LSIF and KLIEP performing similarly.

Table 1: Tabulated deferral accuracy results of Fig. 2. Bolded entries are the best performing approaches if they lie outside of the 1 standard deviation ranges of all other approaches (of different types). Red entries are approaches which are closer to random than the best performing approach.

Dataset	Method	5%	10%	15%	20%	25%	50%	75%
CIFAR-100 Clean [14-56]	Conf	66.14 ± 0.00	66.67 ± 0.00	67.58 ± 0.00	$\mathbf{68.21 \pm 0.00}$	68.47 ± 0.00	70.16 ± 0.00	69.64 ± 0.00
	TwoStage-Exp	65.71 ± 0.02	66.45 ± 0.02	66.93 ± 0.02	67.14 ± 0.02	67.39 ± 0.03	68.41 ± 0.03	69.18 ± 0.01
	Estimate-Diff01	66.12 ± 0.07	66.79 ± 0.07	67.38 ± 0.10	67.98 ± 0.10	68.50 ± 0.09	69.93 ± 0.08	69.72 ± 0.07
	Estimate-MaxProb	$\mathbf{66.22 \pm 0.04}$	$\mathbf{66.98 \pm 0.05}$	67.56 ± 0.04	68.01 ± 0.06	68.53 ± 0.07	70.13 ± 0.06	69.79 ± 0.09
	DRCPE (GCE)	66.01 ± 0.03	66.74 ± 0.07	67.51 ± 0.06	68.03 ± 0.06	68.54 ± 0.03	70.12 ± 0.05	69.84 ± 0.05
	DRCPE (Prob01)	66.14 ± 0.07	66.84 ± 0.05	67.58 ± 0.08	68.07 ± 0.07	68.54 ± 0.06	70.07 ± 0.05	69.73 ± 0.06
	Random	65.40 ± 0.00	65.61 ± 0.00	65.82 ± 0.00	66.03 ± 0.00	66.24 ± 0.00	67.29 ± 0.00	68.34 ± 0.00
CIFAR-100 Clean [8-32]	Conf	61.54 ± 0.00	62.46 ± 0.00	63.37 ± 0.00	64.18 ± 0.00	65.11 ± 0.00	68.17 ± 0.00	68.18 ± 0.00
	TwoStage-Exp	61.41 ± 0.03	62.08 ± 0.04	62.98 ± 0.04	63.55 ± 0.07	64.36 ± 0.06	66.74 ± 0.08	67.58 ± 0.04
	Estimate-Diff01	61.76 ± 0.07	62.82 ± 0.09	63.77 ± 0.10	64.56 ± 0.12	65.31 ± 0.08	67.91 ± 0.12	68.09 ± 0.07
	Estimate-MaxProb	61.57 ± 0.05	62.62 ± 0.08	63.63 ± 0.05	64.30 ± 0.08	65.15 ± 0.05	68.24 ± 0.06	68.16 ± 0.04
	DRCPE (GCE)	61.50 ± 0.06	62.39 ± 0.07	63.33 ± 0.09	64.31 ± 0.06	65.23 ± 0.06	68.18 ± 0.06	68.16 ± 0.04
	DRCPE (Prob01)	61.70 ± 0.08	62.71 ± 0.12	63.63 ± 0.06	64.54 ± 0.09	65.43 ± 0.07	68.19 ± 0.05	68.10 ± 0.04
	Random	60.99 ± 0.00	61.36 ± 0.00	61.72 ± 0.00	62.09 ± 0.00	62.45 ± 0.00	64.27 ± 0.00	66.09 ± 0.00
PathMNIST Clean [8-14]	Conf	86.02 ± 0.00	86.46 ± 0.00	86.73 ± 0.00	86.91 ± 0.00	87.03 ± 0.00	86.66 ± 0.00	86.36 ± 0.00
	TwoStage-Exp	85.29 ± 0.08	85.92 ± 0.08	86.64 ± 0.08	$\mathbf{87.10 \pm 0.05}$	$\mathbf{87.31 \pm 0.03}$	$\mathbf{87.05 \pm 0.03}$	86.59 ± 0.03
	Estimate-Diff01	85.98 ± 0.06	86.59 ± 0.06	86.74 ± 0.03	86.87 ± 0.05	87.06 ± 0.04	86.75 ± 0.08	86.57 ± 0.13
	Estimate-MaxProb	$\mathbf{86.05 \pm 0.02}$	86.47 ± 0.01	86.72 ± 0.01	86.94 ± 0.01	87.04 ± 0.01	86.84 ± 0.05	86.65 ± 0.03
	DRCPE (GCE)	86.02 ± 0.04	86.60 ± 0.05	86.75 ± 0.03	86.93 ± 0.03	87.08 ± 0.04	86.85 ± 0.08	86.71 ± 0.09
	DRCPE (Prob01)	86.00 ± 0.05	86.58 ± 0.03	86.75 ± 0.03	86.89 ± 0.04	87.06 ± 0.04	86.87 ± 0.10	86.72 ± 0.07
	Random	85.05 ± 0.00	85.12 ± 0.00	85.18 ± 0.00	85.25 ± 0.00	85.32 ± 0.00	85.65 ± 0.00	85.98 ± 0.00
CIFAR-100 Label-Noise [14-56]	Conf	57.58 ± 0.00	58.08 ± 0.00	58.74 ± 0.00	59.13 ± 0.00	59.76 ± 0.00	61.32 ± 0.00	60.60 ± 0.00
	TwoStage-Exp	57.36 ± 0.02	57.84 ± 0.03	58.25 ± 0.04	58.74 ± 0.03	58.98 ± 0.06	60.05 ± 0.05	60.14 ± 0.02
	Estimate-Diff01	57.93 ± 0.07	58.69 ± 0.10	59.28 ± 0.07	59.80 ± 0.06	60.24 ± 0.11	61.23 ± 0.06	60.78 ± 0.09
	Estimate-MaxProb	57.85 ± 0.04	58.41 ± 0.05	58.88 ± 0.05	59.36 ± 0.05	59.95 ± 0.07	$\mathbf{61.41 \pm 0.06}$	60.85 ± 0.07
	DRCPE (GCE)	57.85 ± 0.09	58.38 ± 0.07	58.92 ± 0.07	59.44 ± 0.12	59.98 ± 0.09	61.26 ± 0.06	60.82 ± 0.06
	DRCPE (Prob01)	57.96 ± 0.07	58.64 ± 0.09	59.45 ± 0.11	$\mathbf{60.01 \pm 0.06}$	$\mathbf{60.47 \pm 0.11}$	61.31 ± 0.06	60.84 ± 0.10
	Random	57.35 ± 0.00	57.48 ± 0.00	57.61 ± 0.00	57.73 ± 0.00	57.86 ± 0.00	58.51 ± 0.00	59.15 ± 0.00
CIFAR-100 Long-Tail [8-32]	Conf	40.32 ± 0.00	40.70 ± 0.00	41.15 ± 0.00	41.43 ± 0.00	41.84 ± 0.00	43.91 ± 0.00	44.31 ± 0.00
	TwoStage-Exp	40.21 ± 0.03	40.70 ± 0.01	40.93 ± 0.02	41.24 ± 0.05	41.54 ± 0.04	42.87 ± 0.05	43.48 ± 0.03
	Estimate-Diff01	40.83 ± 0.06	41.80 ± 0.05	42.60 ± 0.08	$\mathbf{43.28 \pm 0.07}$	$\mathbf{43.91 \pm 0.05}$	$\mathbf{45.63 \pm 0.12}$	$\mathbf{45.82 \pm 0.07}$
	Estimate-MaxProb	40.32 ± 0.04	40.80 ± 0.05	41.21 ± 0.05	41.60 ± 0.05	42.10 ± 0.04	43.88 ± 0.06	44.38 ± 0.11
	DRCPE (GCE)	40.44 ± 0.05	41.18 ± 0.07	41.99 ± 0.14	42.57 ± 0.12	43.00 ± 0.10	44.86 ± 0.06	45.18 ± 0.04
	DRCPE (Prob01)	40.77 ± 0.04	41.76 ± 0.06	42.50 ± 0.07	43.09 ± 0.07	43.51 ± 0.08	45.08 ± 0.07	45.64 ± 0.07
	Random	39.93 ± 0.00	40.15 ± 0.00	40.37 ± 0.00	40.60 ± 0.00	40.82 ± 0.00	41.92 ± 0.00	43.03 ± 0.00
PathMNIST Specialist [8-14]	Conf	$\mathbf{85.96 \pm 0.00}$	86.49 ± 0.00	86.91 ± 0.00	87.10 ± 0.00	87.08 ± 0.00	85.95 ± 0.00	84.83 ± 0.00
	TwoStage-Exp	85.57 ± 0.06	86.56 ± 0.06	$\mathbf{87.30 \pm 0.07}$	$\mathbf{87.74 \pm 0.08}$	$\mathbf{87.64 \pm 0.07}$	86.02 ± 0.12	84.32 ± 0.06
	Estimate-Diff01	84.99 ± 0.00	84.98 ± 0.01	84.98 ± 0.01	84.97 ± 0.02	84.94 ± 0.09	84.58 ± 0.10	83.83 ± 0.25
	Estimate-MaxProb	85.86 ± 0.03	86.53 ± 0.03	86.86 ± 0.02	87.08 ± 0.05	87.09 ± 0.10	87.05 ± 0.13	86.57 ± 0.08
	DRCPE (GCE)	85.68 ± 0.10	86.38 ± 0.05	87.01 ± 0.05	87.05 ± 0.06	87.05 ± 0.10	86.96 ± 0.14	86.57 ± 0.07
	DRCPE (Prob01)	85.14 ± 0.24	85.77 ± 0.42	86.22 ± 0.65	86.23 ± 0.64	86.23 ± 0.63	86.33 ± 0.55	86.40 ± 0.15
	Random	84.90 ± 0.00	84.81 ± 0.00	84.72 ± 0.00	84.64 ± 0.00	84.55 ± 0.00	84.12 ± 0.00	83.68 ± 0.00
DermaMNIST Specialist [8-8]	Conf	76.01 ± 0.00	75.71 ± 0.00	76.11 ± 0.00	75.76 ± 0.00	75.66 ± 0.00	73.32 ± 0.00	72.52 ± 0.00
	TwoStage-Exp	75.97 ± 0.06	75.66 ± 0.05	74.94 ± 0.22	74.38 ± 0.15	74.14 ± 0.17	73.50 ± 0.11	72.62 ± 0.06
	Estimate-Diff01	76.02 ± 0.15	76.32 ± 0.21	76.40 ± 0.27	76.50 ± 0.27	76.43 ± 0.38	76.40 ± 0.36	76.15 ± 0.21
	Estimate-MaxProb	75.81 ± 0.06	76.00 ± 0.12	76.17 ± 0.14	76.37 ± 0.09	76.41 ± 0.06	75.54 ± 0.07	75.49 ± 0.09
	DRCPE (GCE)	75.72 ± 0.07	75.94 ± 0.12	76.34 ± 0.09	76.52 ± 0.10	76.46 ± 0.10	76.10 ± 0.08	76.13 ± 0.09
	DRCPE (Prob01)	75.77 ± 0.11	76.03 ± 0.17	76.47 ± 0.07	76.59 ± 0.09	76.45 ± 0.08	76.14 ± 0.14	76.12 ± 0.14
	Random	75.78 ± 0.00	75.60 ± 0.00	75.42 ± 0.00	75.24 ± 0.00	75.06 ± 0.00	74.16 ± 0.00	73.27 ± 0.00

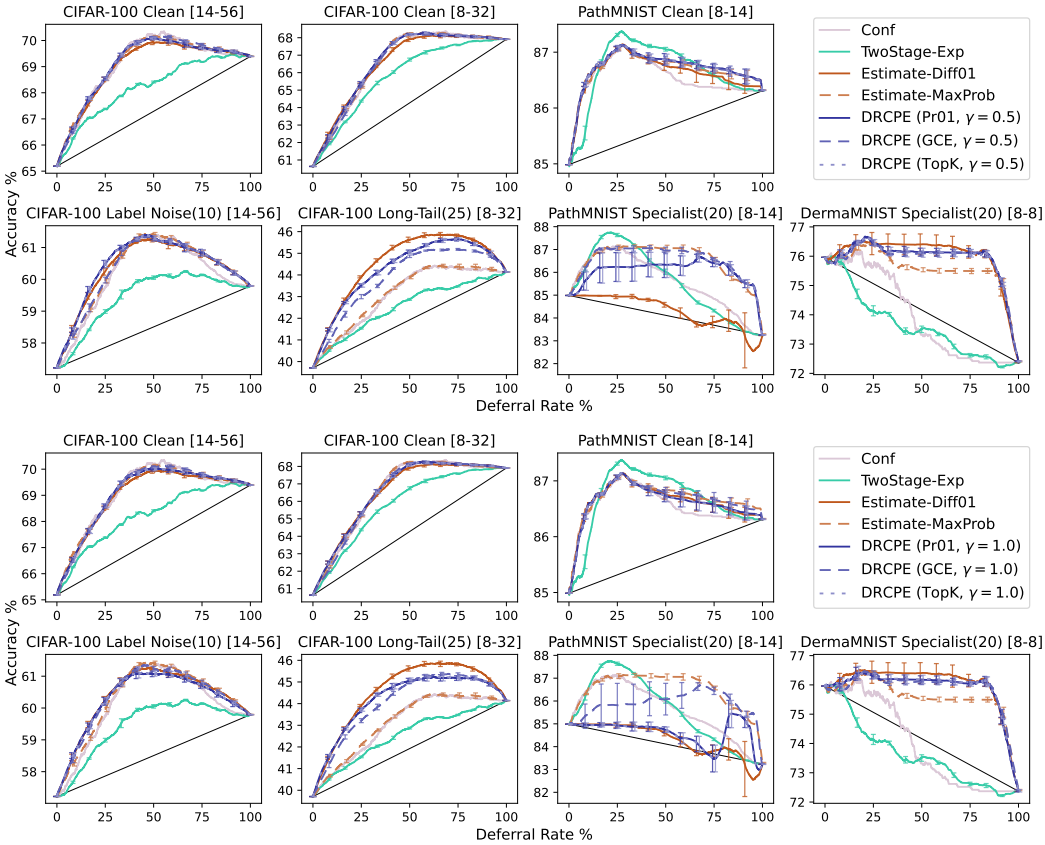

Figure 3: Accuracy-deferral trade-off plots over different ℓs and $\gamma \in \{0.5, 1.0\}$.

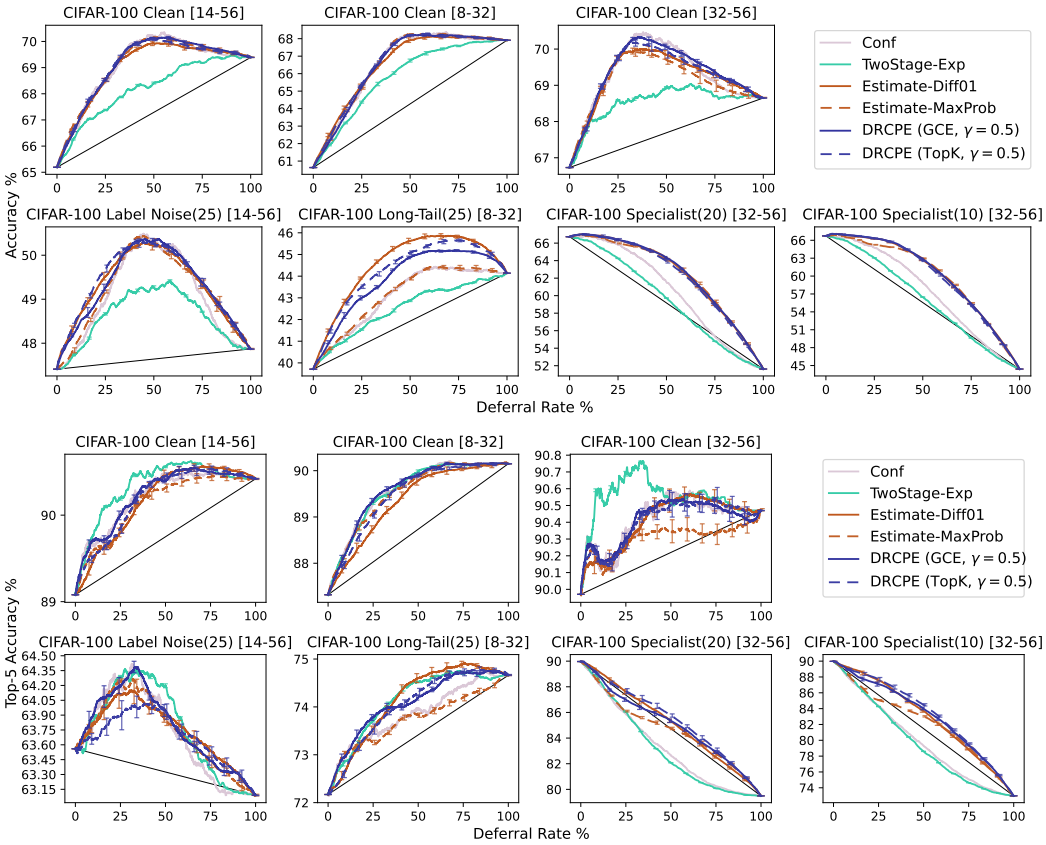

Figure 4: Accuracy-deferral trade-off plots over both accuracy and top-k ($k = 5$) performance.

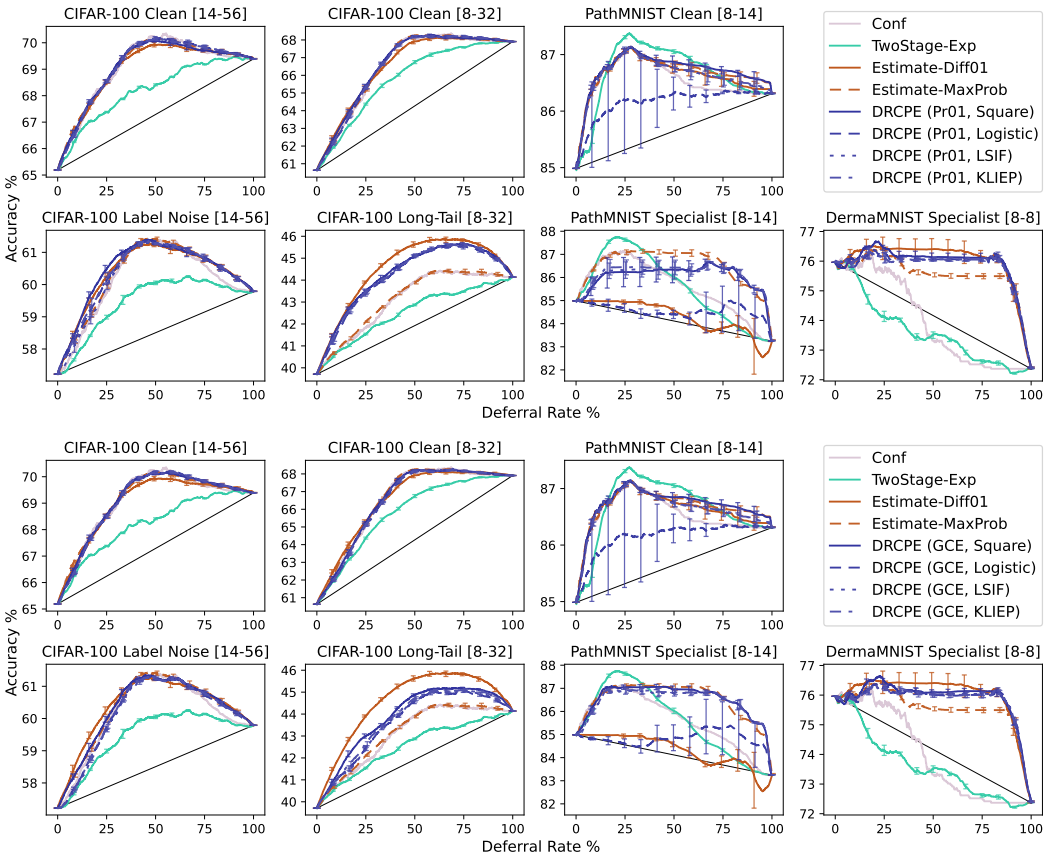

Figure 5: Accuracy-deferral trade-off plots over different DR partial losses (with fixed $\gamma = 0.5$).

L ImageNet Experiments

In the following section, we consider the dog-specialist ImageNet setting examined in Jitkrittum et al. [2023]. The base model h is a MobileNetV2 [Sandler et al., 2018] and the expert $h^{(e)}$ is an EfficientNet-B0 [Kag et al., 2022]. The base model is trained on the clean training split; while the expert trains on all dog examples and a percentage of $p \in \{100, 4, 2\}\%$ non-dog training examples. The deferral mechanism and test set are constructed using the validation set, with a 20:80 split.

L.1 Unequal Temperatures

We consider an experimental setting where we consider unequal temperatures. In particular, we use a specific setting which theoretically match the exact consistency setting described in Theorem 4.5. More specifically, we consider the LSIF [Kanamori et al., 2009] DR loss λ, which yields a nice consistency result as per Corollary H.4. To adhere to this setting, we pick pairs of temperatures $\gamma, \gamma^{(e)}$ (and costs) which corresponds to different deferral costs c (as per the mapping lemma in Lemma H.1): we fix $\gamma = 0.5$ and vary $\gamma^{(e)} \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$.

We consider some "task losses" ℓ which are related to the consistency setting of Corollary H.4. We use the zero-one loss denoted as `DRCPE (Hard01)` (this correspond exactly to theory). We also use a probabilistic variant denoted by `DRCPE (Prob01)`—which uses the L_2 distance between one-hot labels and predicted probabilities. As also consider `DRCPE (Squared)` which is the main experiment's configuration (Section 5) with equal temperatures $\gamma = \gamma^{(e)} = 0.5$ with squared DR loss λ and `Prob01` task loss ℓ.

Tables 2 to 4 summarize these unequal temperature experiments over all tested specialist percentages. For the 100% (non-specialist) regime, `Estimate-MaxProb` and `Estimate-Diff01` perform best or similarly to DR CPE. For the other percentage dog-specialist percentages, `DRCPE (Squared)` (the main experiment's settings) consistently perform best; with the exception of the 4% at 25% deferral rate, where LSIF with `Prob01` and $(0.5, 0.4)$ performs best. In general, temperatures close to the equal temperature regime performs better and `Prob01` typically performs better than their `Hard01` counterparts for the non-100% setting (using the LSIF loss).

In general, these results show that despite `DRCPE Hard01` being the setting which has exact consistency bounds (Appendix H), in practice, for finite sample settings, deviation from theory can yield improvements.

Table 2: ImageNet dog-specialist setting with $p = 100\%$. Base/expert accuracies: $67.57\%/70.79\%$. Entries are accuracy (%) means $\pm$ on standard deviation over 11 random seeds. Bold are highest means.

		Deferral rate (%)			
Approach	$(\gamma, \gamma^{(e)})$	25	50	75	90
Random	–	68.38 ± 0.00	69.18 ± 0.00	69.99 ± 0.00	70.47 ± 0.00
Conf	–	70.01 ± 0.00	70.92 ± 0.00	70.84 ± 0.00	70.80 ± 0.00
TwoStage-Exp	–	69.37 ± 0.01	70.33 ± 0.02	70.65 ± 0.01	70.76 ± 0.00
Estimate-Diff01	–	70.17 ± 0.11	$\mathbf{71.01 \pm 0.05}$	70.88 ± 0.04	$\mathbf{70.82 \pm 0.02}$
Estimate-MaxProb	–	$\mathbf{70.36 \pm 0.03}$	$\mathbf{71.01 \pm 0.02}$	$\mathbf{70.89 \pm 0.02}$	$\mathbf{70.82 \pm 0.01}$
DRCPE (Hard01)	$(0.5, 0.1)$	68.22 ± 0.26	68.49 ± 0.23	69.29 ± 0.13	70.11 ± 0.05
DRCPE (Hard01)	$(0.5, 0.2)$	68.38 ± 0.27	68.64 ± 0.28	69.36 ± 0.14	70.13 ± 0.05
DRCPE (Hard01)	$(0.5, 0.3)$	69.44 ± 0.21	69.74 ± 0.27	69.95 ± 0.20	70.34 ± 0.09
DRCPE (Hard01)	$(0.5, 0.4)$	70.25 ± 0.04	71.00 ± 0.02	70.87 ± 0.02	$\mathbf{70.82 \pm 0.01}$
DRCPE (Hard01)	$(0.5, 0.5)$	70.05 ± 0.03	70.92 ± 0.02	70.85 ± 0.01	70.80 ± 0.01
DRCPE (Prob01)	$(0.5, 0.1)$	67.53 ± 0.01	67.65 ± 0.02	68.69 ± 0.03	69.96 ± 0.01
DRCPE (Prob01)	$(0.5, 0.2)$	67.54 ± 0.01	67.75 ± 0.03	68.86 ± 0.04	70.00 ± 0.02
DRCPE (Prob01)	$(0.5, 0.3)$	68.10 ± 0.23	68.32 ± 0.18	69.19 ± 0.10	70.07 ± 0.03
DRCPE (Prob01)	$(0.5, 0.4)$	70.21 ± 0.10	70.97 ± 0.03	70.85 ± 0.01	70.80 ± 0.00
DRCPE (Prob01)	$(0.5, 0.5)$	69.97 ± 0.07	70.90 ± 0.01	70.84 ± 0.00	70.80 ± 0.00
DRCPE (Squared)	$(0.5, 0.5)$	70.17 ± 0.15	70.96 ± 0.05	70.86 ± 0.02	$\mathbf{70.82 \pm 0.01}$

Table 3: ImageNet dog-specialist setting with $p = 4\%$. Base/expert accuracies: $67.57\%/43.07\%$. Entries are accuracy (%) means $\pm$ on standard deviation over 11 random seeds. Bold are highest means.

		Deferral rate (%)			
Approach	$(\gamma, \gamma^{(e)})$	25	50	75	90
Random	–	61.45 ± 0.00	55.32 ± 0.00	49.20 ± 0.00	45.52 ± 0.00
Conf	–	63.69 ± 0.00	56.87 ± 0.00	48.28 ± 0.00	44.32 ± 0.00
TwoStage-Exp	–	62.08 ± 0.02	54.44 ± 0.04	47.42 ± 0.01	44.46 ± 0.01
Estimate-Diff01	–	66.12 ± 0.46	61.23 ± 0.46	53.86 ± 0.39	48.01 ± 0.18
Estimate-MaxProb	–	65.25 ± 0.06	59.49 ± 0.36	52.32 ± 0.52	47.36 ± 0.33
DRCPE (Prob01)	$(0.5, 0.1)$	64.75 ± 0.32	57.07 ± 0.49	48.84 ± 0.17	44.74 ± 0.04
DRCPE (Prob01)	$(0.5, 0.2)$	65.22 ± 0.36	57.79 ± 0.56	49.13 ± 0.25	44.81 ± 0.06
DRCPE (Prob01)	$(0.5, 0.3)$	65.91 ± 0.15	59.03 ± 0.36	49.81 ± 0.33	44.99 ± 0.11
DRCPE (Prob01)	$(0.5, 0.4)$	$\mathbf{66.51 \pm 0.11}$	60.53 ± 0.17	52.47 ± 0.38	46.87 ± 0.43
DRCPE (Prob01)	$(0.5, 0.5)$	65.61 ± 0.27	60.87 ± 0.41	53.69 ± 0.39	47.98 ± 0.23
DRCPE (Hard01)	$(0.5, 0.1)$	64.67 ± 0.38	57.26 ± 0.50	48.96 ± 0.22	44.78 ± 0.05
DRCPE (Hard01)	$(0.5, 0.2)$	64.71 ± 0.38	57.31 ± 0.52	48.99 ± 0.23	44.78 ± 0.05
DRCPE (Hard01)	$(0.5, 0.3)$	64.79 ± 0.41	57.45 ± 0.56	49.10 ± 0.26	44.81 ± 0.06
DRCPE (Hard01)	$(0.5, 0.4)$	64.93 ± 0.41	57.91 ± 0.54	49.54 ± 0.31	44.96 ± 0.13
DRCPE (Hard01)	$(0.5, 0.5)$	65.75 ± 0.33	59.30 ± 0.36	50.93 ± 0.41	45.90 ± 0.39
DRCPE (Squared)	$(0.5, 0.5)$	66.46 ± 0.25	$\mathbf{61.61 \pm 0.27}$	$\mathbf{54.26 \pm 0.21}$	$\mathbf{48.23 \pm 0.08}$

Table 4: ImageNet dog-specialist setting with $p = 2\%$. Base/expert accuracies: 67.57%/34.10%. Entries are accuracy (%) means $\pm$ on standard deviation over 11 random seeds. Bold are highest means.

		Deferral rate (%)			
Approach	$(\gamma, \gamma^{(e)})$	25	50	75	90
Random	–	59.20 ± 0.00	50.84 ± 0.00	42.47 ± 0.00	37.45 ± 0.00
Conf	–	62.94 ± 0.00	54.44 ± 0.00	43.10 ± 0.00	36.91 ± 0.00
TwoStage-Exp	–	61.18 ± 0.04	51.56 ± 0.04	41.83 ± 0.02	36.76 ± 0.01
Estimate-Diff01	–	65.42 ± 0.51	58.57 ± 0.57	48.15 ± 0.42	40.28 ± 0.25
Estimate-MaxProb	–	64.79 ± 0.14	57.48 ± 0.39	47.00 ± 0.42	39.74 ± 0.21
DRCPE (Hard01)	$(0.5, 0.1)$	63.29 ± 0.51	52.26 ± 0.72	41.23 ± 0.31	36.13 ± 0.07
DRCPE (Hard01)	$(0.5, 0.2)$	63.39 ± 0.52	52.40 ± 0.74	41.29 ± 0.32	36.14 ± 0.08
DRCPE (Hard01)	$(0.5, 0.3)$	63.94 ± 0.42	53.23 ± 0.63	41.68 ± 0.41	36.24 ± 0.11
DRCPE (Hard01)	$(0.5, 0.4)$	64.83 ± 0.22	54.69 ± 0.41	42.97 ± 0.59	36.79 ± 0.32
DRCPE (Hard01)	$(0.5, 0.5)$	65.55 ± 0.11	56.97 ± 0.39	46.15 ± 0.43	39.03 ± 0.34
DRCPE (Prob01)	$(0.5, 0.1)$	63.31 ± 0.48	52.01 ± 0.63	41.03 ± 0.18	36.06 ± 0.05
DRCPE (Prob01)	$(0.5, 0.2)$	64.12 ± 0.32	53.25 ± 0.62	41.47 ± 0.29	36.17 ± 0.07
DRCPE (Prob01)	$(0.5, 0.3)$	65.19 ± 0.05	55.45 ± 0.18	43.68 ± 0.50	36.94 ± 0.36
DRCPE (Prob01)	$(0.5, 0.4)$	65.62 ± 0.06	57.93 ± 0.14	47.68 ± 0.16	40.19 ± 0.07
DRCPE (Prob01)	$(0.5, 0.5)$	65.18 ± 0.20	58.40 ± 0.25	48.08 ± 0.27	40.33 ± 0.13
DRCPE (Squared)	$(0.5, 0.5)$	$\mathbf{65.93 \pm 0.13}$	$\mathbf{59.18 \pm 0.27}$	$\mathbf{48.73 \pm 0.17}$	$\mathbf{40.62 \pm 0.07}$

L.2 Non-Dog Subgroup Evaluation

We further evaluate the 4% specialist setting on only non-dog examples. Here the base model h is substantially more accurate than the expert $h^{(e)}$. Table 5 summarizes these results. Here, `DRCPE (Squared)` performs best with similar insights to Section L.1.

Table 5: ImageNet dog-specialist setting with $p = 4\%$ evaluated on only non-dog examples. Base/expert non-dog accuracies: 67.14%/38.27%. Entries are accuracy (%) means $\pm$ on standard deviation over 11 random seeds. Bold are highest means.

		Deferral rate (%)			
Approach	$(\gamma, \gamma^{(e)})$	25	50	75	90
Random	–	59.92 ± 0.00	52.71 ± 0.00	45.49 ± 0.00	41.16 ± 0.00
Conf	–	62.36 ± 0.00	53.87 ± 0.00	43.77 ± 0.00	39.51 ± 0.00
TwoStage-Exp	–	61.07 ± 0.04	51.76 ± 0.03	43.18 ± 0.01	39.74 ± 0.01
Estimate-Diff01	–	63.08 ± 0.33	57.24 ± 0.51	49.33 ± 0.41	43.27 ± 0.18
Estimate-MaxProb	–	62.80 ± 0.03	55.10 ± 0.35	47.72 ± 0.58	42.61 ± 0.34
DRCPE (Hard01)	$(0.5, 0.1)$	62.42 ± 0.08	53.41 ± 0.27	44.26 ± 0.12	40.00 ± 0.04
DRCPE (Hard01)	$(0.5, 0.2)$	62.41 ± 0.08	53.43 ± 0.28	44.28 ± 0.13	40.00 ± 0.04
DRCPE (Hard01)	$(0.5, 0.3)$	62.40 ± 0.05	53.49 ± 0.28	44.36 ± 0.16	40.02 ± 0.05
DRCPE (Hard01)	$(0.5, 0.4)$	62.43 ± 0.04	53.70 ± 0.29	44.73 ± 0.25	40.15 ± 0.11
DRCPE (Hard01)	$(0.5, 0.5)$	62.63 ± 0.06	54.81 ± 0.29	46.11 ± 0.46	41.10 ± 0.41
DRCPE (Prob01)	$(0.5, 0.1)$	62.45 ± 0.05	53.17 ± 0.22	44.10 ± 0.13	39.97 ± 0.04
DRCPE (Prob01)	$(0.5, 0.2)$	62.54 ± 0.06	53.59 ± 0.37	44.34 ± 0.19	40.02 ± 0.05
DRCPE (Prob01)	$(0.5, 0.3)$	62.64 ± 0.05	54.48 ± 0.30	44.87 ± 0.28	40.18 ± 0.10
DRCPE (Prob01)	$(0.5, 0.4)$	62.93 ± 0.13	56.21 ± 0.27	47.78 ± 0.47	42.08 ± 0.46
DRCPE (Prob01)	$(0.5, 0.5)$	62.66 ± 0.16	56.81 ± 0.47	49.16 ± 0.42	43.26 ± 0.24
DRCPE (Squared)	$(0.5, 0.5)$	$\mathbf{63.32 \pm 0.33}$	$\mathbf{57.68 \pm 0.32}$	$\mathbf{49.77 \pm 0.24}$	$\mathbf{43.52 \pm 0.07}$

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Claims in the abstract and introduction summarize the results of the paper. Introduction gives direct references to theoretical results.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assuamptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: All theoretical results come with a proof in the supplemental material.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Details of baselines and dataset settings are detailed in the supplementary. We detail the exact optimizers and architectures requires to reproduce the results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: GitHub link to code is given.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The training and test details are detailed in the paper and its supplementary material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Error bars are presented in plots, which are derived from 11 experimental runs (all detailed in supplemental material).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Compute resources used are stated.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Code of ethics are followed.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See Appendix B.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Only standard datasets are used and there is no potential harmful model to be released.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Assets are appropriately credited and the license for the data utilized are explicitly stated in the supplemental material.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No]

    Justification: No new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: Core method development does not involve LLMs.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.