# Beyond Augmented-Action Surrogates for Multi-Expert Learning-to-Defer

**Yannis Montreuil** [1] **Axel Carlier** [2] **Lai Xing Ng** [3] **Wei Tsang Ooi** [1]

## Abstract

A learning-to-defer (L2D) system decides, for each input, whether to predict on its own or to hand it to one of several available experts. The very well established recipe trains classifier and router jointly by treating the $K$ classes and $J$ experts as competing actions in one shared $(K+J)$-action geometry. Subsequent work has proposed a series of incremental fixes within this geometry; we show that each still suffers, to varying severity, from an optimization-level pathology (target distortion, gradient amplification, winner-take-all starvation, set-mass collapse, or class–expert coupling) even under statistical consistency. We step outside the augmented-action family entirely and propose a *decoupled surrogate*: a softmax classifier head and an independent sigmoid head per expert, mirroring the two natural objects of the problem. We show that per-sample updates are then coordinatewise and the class–expert Hessian block is identically zero, and prove an excess-risk bound with calibration constant $\max\{2\sqrt{2}, \sqrt{2J/\lambda}\}$—to our knowledge the first multi-expert L2D guarantee whose constant does not grow with the expert pool when the per-expert weight is held fixed. On controlled synthetic studies and on CIFAR-10, CIFAR-10H, and Covertype, it is the only method in our comparison that remains stable as the expert pool grows, preserves rare specialists, and improves over a standalone classifier on every real-data benchmark.

[1]School of Computing, National University of Singapore, Singapore [2]Fédération ENAC ISAE-SUPAERO ONERA, Université de Toulouse, France [3]Institute for Infocomm Research, A*STAR, Singapore. Correspondence to: Yannis Montreuil <yannis.montreuil@u.nus.edu>.

*Accepted to the ICML 2026 Workshop on Statistical Frameworks for Uncertainty in Agentic Systems*, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

## 1. Introduction

Learning-to-Defer (L2D) augments a classifier with the option of handing an input to an expert (Madras et al., 2018; Mozannar & Sontag, 2020). In the multi-expert setting, this means deciding, for each example, whether to predict directly or defer to one among several available experts. The Bayes rule is simple. It compares two conditional objects: the class posterior $\eta_k(x) = \Pr(Y=k \mid X=x)$ and the expert utilities $\alpha_j(x) = \Pr(M_j=Y \mid X=x)$. Predict when the best class posterior exceeds the best expert utility, and defer otherwise. The difficulty lies in designing a surrogate that learns this comparison faithfully from data.

Most existing multi-expert surrogates follow one common route. They introduce an augmented action space containing the $K$ class actions and the $J$ defer-to-expert actions, learn a single score vector $a(x) \in \mathbb{R}^{K+J}$ over that space, and derive both class and expert supervision from those same shared scores (Mozannar & Sontag, 2020; Verma et al., 2022; Cao et al., 2024; Mao et al., 2025d; Liu et al., 2026). This route has delivered statistically consistent surrogates, but it is also a strong modeling choice. The Bayes rule compares a categorical posterior with expert utilities, whereas the augmented-action reduction places classes and experts inside one shared action geometry. Prior work already shows that this geometry can produce unbounded and misaligned defer-probability estimates (Verma et al., 2022; Cao et al., 2024), and that adding experts can cause systematic underfitting as the surrogate reward structure distorts the gradient budget (Liu et al., 2026). Existing fixes restore the correct probability range or cap the reward to one expert, but they address only one symptom at a time without examining the *optimization* path that leads there.

This is the distinction that motivates the paper. A surrogate can be judged along two axes: what it learns at the population level, and how it distributes supervision during training. Consistency tells us about the first axis: the conditional minimizer recovers the Bayes decision. Bounded probability estimation tells us that the learned scores live on the correct range. Neither says anything about how overlap among experts reshapes the gradient budget of a sample, whether correct but non-winning experts are actively suppressed, or whether the class and expert updates remain

entangled. In multi-expert L2D, these local effects are consequences of the surrogate design itself, and they become more pronounced as the number of experts increases.

Our analysis makes this concrete. For additive augmented cross-entropy, the conditional target is distorted by the total expert utility $\sum_j \alpha_j(x)$: the surrogate learns a flattened version of the Bayes quantities rather than the quantities themselves (Proposition 14). At the sample level, the gradient budget is amplified by $1 + |\mathcal{J}(y, m)|$, where $\mathcal{J}(y, m)$ is the realized set of correct experts (Proposition 4). The distortion is therefore statistical and optimization-level at once. The loss targets the wrong object, and it allocates more update mass to regions where experts already agree. PiCCE (Liu et al., 2026) removes the amplification by selecting a single correct expert per sample, but then pushes every other correct expert *down*, creating a starvation effect that can suppress rare specialists (Proposition 5). Mao et al. (2025d) avoid amplification but learn only the total acceptable-set mass, with no ranking inside that set (Proposition 6). A-SM (Cao et al., 2024) corrects the target and produces bounded estimates, but retains a class–expert gradient coupling that can be $O(J)$ (Proposition 7). OvA (Verma et al., 2022) fully decouples expert gradients, but treats the class posterior as $K$ independent binary tasks rather than one categorical distribution (Proposition 19). The issue is therefore not an isolated defect in any single loss: the augmented-action family repeatedly trades one mismatch for another.

We propose a *decoupled surrogate* that departs from the augmented-action family entirely. The class head uses a softmax, producing $p(x) \in \Delta^K$. Each expert head uses an independent sigmoid, producing $u_j(x) \in (0, 1)$. This factorization aligns both axes at once and is, to our knowledge, the first multi-expert L2D surrogate to simultaneously: *(i)* identify $(\eta, \alpha)$ at the population optimum with a categorical class posterior $p \in \Delta^K$; *(ii)* have fully decoupled, uniformly bounded optimization geometry (no amplification, no starvation, no class–expert coupling); *(iii)* admit a consistency bound with calibration constant $\max\{2\sqrt{2}, \sqrt{2J/\lambda}\}$, which is invariant in $J$ whenever the per-expert weight $\lambda/J$ is held fixed (equivalently, $\lambda$ scales linearly with $J$), and equals $2\sqrt{2}$ whenever $\lambda/J \geq 1/4$.

**Contributions.**

- We identify a common failure mode of augmented-action multi-expert surrogates. For five representative losses, we prove exact target and gradient/curvature statements showing how shared action geometry causes target distortion, gradient amplification, winner-take-all starvation, acceptable-set ranking loss, class–expert coupling, or improper class geometry (Section 4).
- We propose a decoupled surrogate that estimates the two Bayes objects on their native scales: a categorical class posterior and independent expert utilities. We prove that its population minimizer is exactly $(\eta, \alpha)$ and that its gradient and Hessian are block-decoupled and uniformly bounded (Section 5).
- We prove a quantitative consistency bound: decoupled surrogate excess risk controls L2D excess risk with calibration constant $\max\{2\sqrt{2}, \sqrt{2J/\lambda}\}$. Hence the constant is independent of $J$ when the per-expert weight $\lambda/J$ is fixed, unlike the available quantitative augmented-action bounds (Theorem 10).
- We validate the decoupled surrogate on synthetic benchmarks targeting each predicted pathology and on three real-data benchmarks, showing that the predicted pathologies appear in practice and that the decoupled surrogate remains stable as the expert pool grows (Section 6).

## 2. Related Work

Learning-to-defer extends selective prediction and classification with a reject option (Chow, 1970; Bartlett & Wegkamp, 2008; Geifman & El-Yaniv, 2017; Cortes et al., 2016; Cao et al., 2022; Cortes et al., 2024b) by allowing a learner to route inputs to external experts rather than abstaining altogether. The modern L2D formulation was introduced for single-expert systems by Madras et al. (2018) and Mozannar & Sontag (2020), and has since grown into a broader literature on surrogate design and statistical guarantees (Mao et al., 2024a; Charusaie et al., 2022; Mozannar et al., 2023; Mao et al., 2024e; 2025d; Montreuil et al., 2025b; 2026a). Appendix A gives an extended related-work discussion.

**Surrogate design and underfitting.** Early predictor-rejector frameworks separate the classification and rejection functions, but Ni et al. (2019) show that such surrogates are inconsistent in the multiclass case. Mozannar & Sontag (2020) introduce the first consistent multiclass surrogate via an augmented action space that jointly scores the $K$ class actions and $J$ defer-to-expert actions. Subsequent work refines this shared-vector design. Verma et al. (2022) compare coupled softmax and one-vs-all (OvA) parameterizations and show that the softmax defer-probability estimate is unbounded, whereas OvA remains bounded. Cao et al. (2024) show that symmetric reductions can produce invalid defer probabilities and propose an asymmetric softmax parameterization with a multi-expert extension. Mao et al. (2025d) design a single-stage surrogate family with formal consistency guarantees for multi-action deferral. A complementary line studies optimization pathologies: Narasimhan et al. (2022) first identify classifier underfitting in the single-expert setting and propose a post-hoc correction. Still in the single-expert setting, Liu et al. (2024) propose a training-time mitigation. More recently, Liu et al. (2026) show that in the multi-expert setting, even redundant experts can in-

duce underfitting through the additive expert reward, and propose a winner-take-all correction (PiCCE).

**Positioning.** Our work differs from the above in scope: rather than proposing an incremental fix within the augmented-action family, we show that the family as a whole exhibits a pattern in which each surrogate trades a fix on one axis (statistical target or optimization geometry) for a failure on the other. The decoupled surrogate departs from the shared-action-space design entirely and avoids this trade-off by construction.

## 3. Preliminaries

**Multi-expert learning to defer.** Following the modern learning-to-defer literature (Madras et al., 2018; Mozannar & Sontag, 2020; Mao et al., 2024e; Montreuil et al., 2026a), let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space supporting a triple $(X, Y, \mathbf{M})$ with distribution $\mathcal{D}$ on $\mathcal{X} \times [K] \times [K]^J$, where $\mathcal{X} \subseteq \mathbb{R}^d$ is the input space, $[K] \coloneqq \{1, \dots, K\}$ is the label space, and $\mathbf{M} = (M_1, \dots, M_J)$ records the predictions of $J$ experts. We write $(x, y, m)$ for a realization of $(X, Y, \mathbf{M})$.

The learner consists of a classifier $h : \mathcal{X} \to [K]$ and a router $r : \mathcal{X} \to \{0\} \cup [J]$. The decision $r(x) = 0$ means that the learner predicts with $h(x)$, whereas $r(x) = j$ means that the instance is routed to expert $j$. In the classification-specialized setting considered throughout the paper (Mozannar & Sontag, 2020; Verma et al., 2022; Cao et al., 2024; Mao et al., 2024e; Liu et al., 2026), the only question is therefore which of the $K$ class predictions and the $J$ expert actions yields the smallest conditional error.

**Definition 1** (Multi-Expert Defer Loss)**.** The multi-expert defer loss incurred on a realization $(x, y, m)$ is

$$\ell_\perp(h, r;\, x, y, m) = \mathbf{1}\{h(x) \neq y\}\mathbf{1}\{r(x) = 0\} + \sum_{j \in [J]} \mathbf{1}\{m_j \neq y\}\mathbf{1}\{r(x) = j\}. \quad (1)$$

Its population risk is $\mathcal{E}_\perp(h, r) \coloneqq \mathbb{E}[\ell_\perp(h, r;\, X, Y, \mathbf{M})]$.

At each input, the learner either predicts on its own and incurs the usual misclassification loss, or it transfers control to one of the experts and pays that expert's error on the same example. The population objective is to choose the better option pointwise in expectation.

**Bayes quantities.** The Bayes rule is governed by two families of conditional probabilities. The first is the class posterior $\eta_k(x) \coloneqq \Pr(Y = k \mid X = x)$ for $k \in [K]$. The second is the vector of expert utilities $\alpha_j(x) \coloneqq \Pr(M_j = Y \mid X = x)$ for $j \in [J]$. These quantities are the natural objects of the defer problem. Predicting class $k$ incurs conditional risk $1 - \eta_k(x)$, while deferring to expert $j$ incurs conditional risk $1 - \alpha_j(x)$. Thus the optimal action depends only on the largest class posterior and the largest expert utility, not on an augmented distribution over $K + J$ actions.

**Lemma 2** (Bayes Rule for Multi-Expert L2D)**.** *For every $x \in \mathcal{X}$, a Bayes-optimal decision is obtained by*

$$\begin{aligned} r^\star(x) &= 0 \quad \textit{if } \max_k \eta_k(x) \geq \max_j \alpha_j(x), \\ r^\star(x) &\in \arg\max_{j \in [J]} \alpha_j(x) \quad \textit{otherwise.} \end{aligned} \quad (2)$$

*with $h^\star(x) \in \arg\max_{k \in [K]} \eta_k(x)$.*

Lemma 2 reveals the basic statistical structure of the problem. The class posterior $\eta(x)$ is a categorical distribution and therefore lies in the simplex $\Delta^K$. By contrast, the expert utilities $\alpha_1(x), \dots, \alpha_J(x)$ are $J$ conditional probabilities in $[0, 1]$ with no analogous normalization constraint. The Bayes-sufficient object is therefore the pair $(\eta(x), \alpha(x)) \in \Delta^K \times [0, 1]^J$.

**Augmented-action viewpoint.** Because the loss in Definition 1 is discontinuous, gradient-based learning must proceed through a surrogate (Zhang, 2002; Bartlett et al., 2006; Steinwart, 2007). The dominant one-stage reduction, introduced by Mozannar & Sontag (2020), casts the problem as prediction over one enlarged action space $\mathcal{A} \coloneqq [K] \cup \{\perp_1, \dots, \perp_J\}$ and learns one score for each action.

**Definition 3** (Augmented-Action Surrogate Family)**.** An *augmented-action surrogate* is specified by a score map $a : \mathcal{X} \to \mathbb{R}^{K+J}$ together with a samplewise loss of the form

$$\Phi^{\text{aug}}(a;\, x, y, m) \;=\; \Phi^{\text{aug}}_{\text{cls}}(a;\, x, y) \;+\; \Phi^{\text{aug}}_{\text{exp}}(a;\, x, y, m),$$

and a prediction rule that selects one augmented action from the same shared score vector $a(x)$, either by $\arg\max_i a_i(x)$ or by the argmax of a maxima-preserving transform of $a(x)$.

This definition covers the baselines studied in the paper: additive cross-entropy (Mozannar & Sontag, 2020), Verma's OvA surrogate (Verma et al., 2022), A-SM (Cao et al., 2024), PiCCE (Liu et al., 2026), and Mao25 (Mao et al., 2025d). They differ in how they build the class and expert terms, but they all learn defer decisions through one shared action geometry over classes and experts.

## 4. The Augmented-Action Mismatch

We now analyze the augmented-action surrogates of Definition 3 along two axes: *(i)* the *conditional target* at the population optimum, and *(ii)* the *local geometry* — how the surrogate distributes gradient mass during training. Consistency and bounded estimation address axis (i); we

show that axis (ii) matters equally, and that the augmented-action family repeatedly trades a failure on one axis for a failure on the other. Throughout, $\mathcal{J}(y,m) := \{j \in [J] : m_j = y\}$ denotes the set of correct experts and $q_i(x) := \exp(a_i)/\sum_{r=1}^{K+J}\exp(a_r)$. Detailed proofs, curvature calculations, and worked examples for the baseline analyses are collected in Appendix D–I.

### 4.1. Additive Cross-Entropy: Target Distortion and Gradient Amplification

We begin with the simplest augmented-action surrogate, additive cross-entropy (Mozannar & Sontag, 2020), which rewards the correct class and every correct expert through one shared softmax:

$$\Phi^{\mathrm{CE}}(a;\, x, y, m) = -\log q_y(x) - \sum_{j\in\mathcal{J}(y,m)} \log q_{K+j}(x). \tag{3}$$

Prior work has already shown that additive CE can produce unbounded defer-probability estimates (Verma et al., 2022) and can underfit when experts are redundant (Liu et al., 2026). This makes it a useful lens for the two mismatches studied in this paper: on axis (i), it distorts the *target*; on axis (ii), it induces a multiplicity pathology that worsens as the number of experts grows. The resulting failure is therefore not merely one of probability validity. Equally importantly, the shared augmented geometry alters the per-sample gradients and curvature in a manner that systematically favors regions of high expert overlap.

At the population level, additive CE does not target $(\eta, \alpha)$ directly. Its conditional optimum normalizes both the class and expert terms by the same factor $1 + U(x)$, where $U(x) = \sum_j \alpha_j(x)$, so the target itself changes with the total expert overlap. If experts $2, \ldots, J$ are redundant copies of expert 1, the Bayes decision is unchanged, yet the normalization grows linearly in $J$ and pushes every softmax probability toward zero. This same normalization underlies the unbounded reconstruction issue identified by Cao et al. (2024). We defer the exact conditional minimizer statement to Appendix D.2.

**Gradient and curvature amplification (axis (ii)).** The target distortion is only part of the issue. The samplewise geometry is at least as important, because it determines how the surrogate allocates update mass during training.

**Proposition 4** (Samplewise amplification in additive CE)**.** *For one sample $(x, y, m)$ with $\mathcal{J} = \mathcal{J}(y, m)$, the gradient and Hessian of $\Phi^{\mathrm{CE}}$ satisfy*

$$\frac{\partial \Phi^{\mathrm{CE}}}{\partial a_i} = \big(1 + |\mathcal{J}|\big)\, q_i(x) - \mathbf{1}\{i = y\} - \mathbf{1}\{i = K{+}j,\ j \in \mathcal{J}\}. \tag{4}$$

$$\nabla_a^2 \Phi^{\mathrm{CE}} \;=\; (1 + |\mathcal{J}|)\big(\mathrm{Diag}(q) - qq^\top\big), \tag{5}$$

*with largest eigenvalue $\lambda_{\max}(\nabla_a^2 \Phi^{\mathrm{CE}}) \le (1 + |\mathcal{J}|)/2$.*

Both parts carry the same multiplicity factor $1 + |\mathcal{J}|$. If six of eight experts are correct on a sample, additive CE treats it as $7\times$ more important than standard CE would, and the curvature jumps from $1/2$ to $7/2$. Since samples with many correct experts tend to be easy, high-agreement regions, the optimizer is both biased *toward* those regions and poorly conditioned there — pulled away from the decision boundaries where the classify-versus-defer choice is most consequential.

### 4.2. Why Winner-Take-All Fails

The amplification of §4.1 arises because the additive reward fires once for the true class *and* once for every correct expert. A natural repair is to reward exactly one correct expert per sample. This is the strategy of PiCCE (Liu et al., 2026), instantiated with cross-entropy as the base multiclass loss. Let $j^\star$ be the deterministically tie-broken maximizer of $a_{K+j}$ over $j \in \mathcal{J}(y, m)$ when this set is nonempty. Then

$$\Phi^{\mathrm{PiCCE}}(a;\, x, y, m) = -\log q_y(x) - \mathbf{1}\{\mathcal{J} \neq \varnothing\} \log q_{K+j^\star}(x). \tag{6}$$

With at most two log-loss terms active, the total positive gradient mass is capped at 2 regardless of how many experts are correct, and the Hessian is $\nabla_a^2 \Phi^{\mathrm{PiCCE}} = 2(\mathrm{Diag}(q) - qq^\top)$, so $\lambda_{\max} \le 1$ uniformly (Appendix E.4). On axis (ii), the multiplicity explosion of (5) is gone.

**Starvation of non-winning experts (axis (ii)).** The $\arg\max$ selection solves the amplification problem but creates a new gradient pathology.

**Proposition 5** (Winner-take-all starvation in PiCCE)**.** *Under PiCCE with $\mathcal{J}(y, m) \neq \varnothing$, the gradient on any logit is $\partial\Phi^{\mathrm{PiCCE}}/\partial a_i = 2q_i - \mathbf{1}\{i = y\} - \mathbf{1}\{i = K{+}j^\star\}$. For any correct expert $k \in \mathcal{J}$ with $k \neq j^\star$,*

$$\frac{\partial \Phi^{\mathrm{PiCCE}}}{\partial a_{K+k}} \;=\; 2q_{K+k}(x) \;>\; 0. \tag{7}$$

*Gradient descent pushes expert $k$'s logit* downward*, even though it is correct.*

The mechanism is simple: in the shared softmax, the two log-loss terms reward only the true class $y$ and the winning expert $j^\star$. Every other action receives only the repulsive softmax redistribution $2q_i$, which pushes its logit down. For classes $k \neq y$ this repulsion is desirable. For correct experts $k \in \mathcal{J} \setminus \{j^\star\}$ it is not: these experts are being punished despite being right, purely because a rival expert had a higher logit.

**Self-reinforcing lock-in and a moving target.** Starvation compounds across iterations. If $j_1$ starts with a marginally higher logit than $j_2$ on a shared region, $j_1$ wins the $\arg\max$ and receives gradient $2q_{K+j_1} - 1$, while $j_2$ is pushed down ($2q_{K+j_2} > 0$). In the non-saturated regime $q_{K+j_1} < 1/2$, the winner is pulled up and the gap widens; even after saturation, the non-winning correct expert continues to receive the repulsive positive gradient. This positive feedback converts initialization noise into deterministic suppression. This also corrupts the statistical target on axis (i). PiCCE does not estimate $(\eta, \alpha)$ but $(\eta, \gamma(a, \cdot))$, where $\gamma_j(a, x) = \Pr(j = j^\star \text{ and } j \in \mathcal{J} \mid X = x)$ depends on the current scores through the winner map (Appendix E.3). As starvation suppresses experts, $\gamma_j$ shifts mass toward surviving winners, further entrenching their dominance. The target is therefore *moving* rather than fixed by the data distribution.

### 4.3. Three Refined Surrogates, Three Residual Gaps

Three further baselines target different defects. Each closes one gap but opens another.

**Mao25: correct set, no ranking.** The surrogate of Mao et al. (2025d) penalizes the total mass assigned outside the acceptable set: $\Phi^{\text{Mao}} = 1 - \big(q_y + \sum_{j \in \mathcal{J}} q_{K+j}\big)$. It comes with formal consistency guarantees (Mao et al., 2025d). Beyond those guarantees, we observe that the loss is *linear* in the total acceptable mass $S_{\mathcal{S}}$, making it a set-valued analogue of multiclass MAE (Ghosh et al., 2017).

**Proposition 6** (Acceptable-set mass in Mao25)**.** *The gradient is $\partial\Phi^{\text{Mao}}/\partial a_i = q_i\big(S_{\mathcal{S}} - \mathbf{1}\{i \in \mathcal{S}(y,m)\}\big)$, where $\mathcal{S}(y,m) = \{y\} \cup \{K+j : j \in \mathcal{J}\}$ and $S_{\mathcal{S}} = \sum_{c \in \mathcal{S}} q_c$. The loss depends on the acceptable set only through $S_{\mathcal{S}}$, and for acceptable coordinates the gradient has the common factor $S_{\mathcal{S}} - 1$. Thus the samplewise update distinguishes acceptable from unacceptable actions, but it contains no term that prefers one acceptable expert over another on the basis of expert quality.*

Because the samplewise supervision cannot express preferences inside the acceptable set, Mao25 does not recover expert utilities directly. The surrogate can identify a Bayes-optimal action, but it does not estimate the full vector $(\eta, \alpha)$ needed for utility-level comparison among experts — a limitation on axis (i). The same MAE-like structure also weakens optimization: the total attractive gradient mass on acceptable actions is $\sum_{i \in \mathcal{S}} -\partial\Phi^{\text{Mao}}/\partial a_i = S_{\mathcal{S}}(1 - S_{\mathcal{S}})$, which is at most $1/4$ and vanishes as $S_{\mathcal{S}} \to 0$. So when the model currently assigns little mass to the acceptable set, the corrective signal is weak, consistent with the optimization behavior commonly associated with MAE-like losses (Zhang & Sabuncu, 2018).

**A-SM: correct target, coupled geometry.** The asymmetric softmax of Cao et al. (2024) defines class probabilities $\xi_k = \exp(a_k)/\sum_{k'} \exp(a_{k'})$ over classes only and expert estimates via a sigmoid $\psi_j = \sigma\big(a_{K+j} - \log \sum_{k' \neq k^\star} \exp(a_{k'})\big)$ with $k^\star \in \arg\max_k a_k$ (Definition 17). At the population optimum $\xi_k^\star = \eta_k$ and $\psi_j^\star = \alpha_j$: axis (i) is fully corrected. But $\psi_j$ still depends on the class logits, coupling the two estimation problems on axis (ii).

**Proposition 7** (Class–expert gradient coupling in A-SM)**.** *Fix $k^\star \in \arg\max_{k \in [K]} a_k$. For any $r \neq k^\star$, let $\pi_r = \exp(a_r)/\sum_{k \neq k^\star} \exp(a_k)$. Then*

$$\frac{\partial \Phi^{\text{ASM}}}{\partial a_r} = \underbrace{(\xi_r - \mathbf{1}\{r = y\})}_{\textit{class term}} - \pi_r \underbrace{\sum_{j=1}^{J} \big(\psi_j - \mathbf{1}\{m_j = y\}\big)}_{\textit{expert leakage}} . \tag{8}$$

The leakage term vanishes in expectation at the optimum but can be $O(J)$ on individual samples. The mixed Hessian block has operator norm $\leq \sqrt{J}/4$ (Appendix G.4), so the coupling grows with the expert pool. A-SM is the strongest augmented baseline on axis (i); its residual weakness is entirely on axis (ii).

**OvA: decoupled gradients, improper class posterior.** OvA (Verma et al., 2022) abandons the shared simplex entirely, replacing it by $K + J$ independent binary logistic losses (Definition 18). Because every coordinate is optimized independently, the amplification, starvation, and coupling pathologies are all absent: axis (ii) is fully resolved. Its remaining gap is on axis (i): the class head consists of $K$ independent sigmoids, so the estimates need not lie in $\Delta^K$ away from the optimum (Cao et al., 2024).

**Synthesis.** Table 1 summarizes the picture. Every baseline repairs some properties but introduces or retains others; the shared $(K+J)$-action geometry is the common thread. Two design principles emerge, one for each axis: *separate* the class and expert estimation problems (no shared normalization, no gradient coupling), addressing axis (ii); and *match* each head to the statistical type of its target (softmax for $\eta \in \Delta^K$, independent sigmoids for $\alpha_j \in [0, 1]$), addressing axis (i). The next section shows that the decoupled surrogate satisfies both.

## 5. The Decoupled Surrogate

The analysis of Section 4 traces every pathology to one architectural choice: entangling class and expert estimates inside a shared normalization. Two design principles emerged: *separate* the two estimation problems so that gradients do

*Table 1.* Structural properties of multi-expert L2D surrogates. ✓ = satisfied; ✗ = violated. Columns (I)/(II)/(TH): statistical target / local geometry / theoretical guarantee. Each entry cites the proposition, definition, equation, or appendix subsection that establishes it.

| | | Add. CE | PiCCE | Mao25 | A-SM | OvA | **Decoupled** |
|---|---|---|---|---|---|---|---|
| Conditional minimizer recovers the Bayes pair $(\eta, \alpha)$ | (I) | ✗ (14) | ✗ (E.3) | ✗ (6) | ✓ (17) | ✓ (19) | ✓ (10) |
| Per-sample gradient and Hessian free of $O(\lvert\mathcal{J}\rvert)$ prefactor | (II) | ✗ (4) | ✓ (E.4) | ✓ (6) | ✓ (7) | ✓ (H.3) | ✓ (9) |
| No winner-take-all suppression of correct experts | (II) | ✓ (4) | ✗ (5) | ✓ (6) | ✓ (7) | ✓ (19) | ✓ (9) |
| No class–expert coupling in the per-sample gradient | (II) | ✗ (4) | ✗ (5) | ✗ (6) | ✗ (7) | ✓ (H.3) | ✓ (12) |
| Class head is a categorical posterior throughout training | (I) | ✗ (3) | ✗ (3) | ✗ (3) | ✓ (17) | ✗ (19) | ✓ (8) |
| Quantitative surrogate-to-target excess-risk bound | (TH) | ✓† | ✗ | ✓‡ | ✗ | ✗ | ✓ (10) |

†Comp-sum reduction (Mao et al., 2024e), constant $\sqrt{2(J+1)}$. ‡Single-stage family (Mao et al., 2025d), constant $K+J$.

not interfere, and *match* each head to the statistical type of its target so that the conditional minimizer recovers $(\eta, \alpha)$ directly. The decoupled surrogate implements both.

### 5.1. Formulation and Prediction

Let $w : \mathcal{X} \to \mathbb{R}^K$ and $s : \mathcal{X} \to \mathbb{R}^J$ denote the classifier and expert score maps. The decoupled surrogate estimates the class posterior and the expert utilities on their native probability scales: $p_k(x) = \exp(w_k(x))/\sum_{\ell=1}^{K}\exp(w_\ell(x))$ and $u_j(x) = \sigma(s_j(x))$, where $\sigma$ is the logistic sigmoid. The class estimate $p(x) \in \Delta^K$ is a proper categorical distribution; each expert estimate $u_j(x) \in (0,1)$ is an independent probability. These types mirror the Bayes-sufficient pair $(\eta, \alpha) \in \Delta^K \times [0,1]^J$ exactly — no shared simplex forces classes and experts to compete for probability mass. We record the corresponding bounded cost-sensitive formulation in Appendix C; throughout the main text we restrict attention to the classification-specialized zero-one defer case.

**Definition 8** (Decoupled surrogate)**.** For $\lambda > 0$, the decoupled surrogate is

$$\Phi_\lambda^{\mathrm{dec}}(w, s;\, x, y, m) \coloneqq -\log p_y(x) - \frac{\lambda}{J}\sum_{j=1}^{J}\Big[t_j \log u_j(x) + (1-t_j)\log\big(1-u_j(x)\big)\Big], \tag{9}$$

where $t_j \coloneqq \mathbf{1}\{m_j = y\}$.

The first term is the standard multiclass log-loss for the classifier. The second is an average of $J$ independent Bernoulli log-losses, one per expert, with binary targets $t_j$. The $\lambda/J$ weight has a single interpretation: $\lambda$ is the total expert-side budget, split evenly across the $J$ experts so that the per-expert weight $\lambda/J$ is the quantity that actually controls gradients, curvature, and the surrogate-regret transfer constant (Section 5.3). Any $\lambda > 0$ yields the same conditional minimizer, so $\lambda$ governs only the relative learning speed between the two heads, not the target being estimated. Appendix C.4 discusses this choice in detail.

At test time, the decoupled surrogate applies the plug-in Bayes comparison directly in probability space:

$$\begin{aligned} r_{\mathrm{dec}}(x) &= 0 \quad \text{if } \max_k p_k(x) \geq \max_j u_j(x), \\ r_{\mathrm{dec}}(x) &\in \arg\max_{j\in[J]} u_j(x) \quad \text{otherwise.} \end{aligned} \tag{10}$$

### 5.2. Gradient Structure

The per-sample gradient of the decoupled surrogate resolves every pathology identified in Section 4. We prove every result in Appendix C.2.

**Proposition 9** (Per-sample gradient of the decoupled surrogate)**.** *For one sample* $(x, y, m)$,

$$\begin{aligned} \frac{\partial \Phi_\lambda^{\mathrm{dec}}}{\partial w_r} &= p_r(x) - \mathbf{1}\{r = y\}, \\ \frac{\partial \Phi_\lambda^{\mathrm{dec}}}{\partial s_j} &= \frac{\lambda}{J}\big(u_j(x) - t_j\big). \end{aligned} \tag{11}$$

Three properties follow immediately, each directly addressing a failure mode of the augmented-action family:

- *No amplification.* The gradient for expert $j$ depends only on $j$'s own prediction $u_j$ and target $t_j$. Adding more correct experts changes nothing about any individual gradient. The $O(\lvert\mathcal{J}\rvert)$ factor of Proposition 4 is absent.
- *No starvation.* If expert $j$ is correct, then $\partial\Phi/\partial s_j = (\lambda/J)(u_j - 1) < 0$ whenever $u_j < 1$: expert $j$'s logit is pushed *up*, regardless of what other experts predict or whether a rival has a higher score. Every correct expert receives positive reinforcement. The winner-take-all suppression of Proposition 5 is absent.
- *No coupling.* The classifier gradient involves only the class probabilities $p(x)$ and the label $y$; it is completely independent of the expert predictions and targets. Conversely, each expert gradient depends only on its own head. The class–expert leakage of Proposition 7 is absent.

The second-order geometry confirms this decoupling (Appendix C.3). Because the class and expert terms depend on disjoint parameters, the Hessian is block-diagonal with a

*zero* mixed block:

$$\nabla^2_{(w,s)}\Phi^{\mathrm{dec}}_\lambda = \begin{pmatrix} \mathrm{Diag}(p) - pp^\top & 0 \\ 0 & \frac{\lambda}{J} D_u \end{pmatrix}. \tag{12}$$

where $D_u = \mathrm{Diag}\big(u_j(1-u_j)\big)$. The class block is the standard softmax covariance with $\lambda_{\max} \le 1/2$; the expert block is diagonal with entries bounded by $\lambda/(4J)$. Contrast this with additive CE, whose curvature scales as $(1+|\mathcal{J}|)/2$ (Proposition 4), and with A-SM, whose mixed block grows as $\sqrt{J}$ (Proposition 7). OvA shares the decoupled geometry, but its class head consists of $K$ independent sigmoids whose outputs need not lie in $\Delta^K$: away from the population optimum, several class estimates can simultaneously exceed $0.5$, so $\max_k \hat{\eta}_k$ is uncalibrated against $\max_j u_j$ in the classify-versus-defer comparison. The decoupled surrogate is the only one in our comparison that combines full decoupling with a categorical class posterior $p \in \Delta^K$.

### 5.3. Consistency and Surrogate-Regret Transfer

The structural separation translates into a clean statistical guarantee. Because the conditional surrogate risk decomposes into a multiclass cross-entropy for $p$ plus $J$ independent Bernoulli cross-entropies for $(u_1, \ldots, u_J)$, each subproblem has a unique minimizer: $p^\star(x) = \eta(x)$ and $u_j^\star(x) = \alpha_j(x)$ for all $j \in [J]$ (Appendix C.5).

The following theorem makes this guarantee quantitative: any reduction in the decoupled surrogate excess risk transfers, at a controlled rate, to the true L2D excess risk.

**Theorem 10** (Consistency bound for the decoupled surrogate)**.** *Let $p\colon \mathcal{X} \to \Delta^K$ and $u\colon \mathcal{X} \to [0,1]^J$ be measurable, and let $f_{p,u} = (h_{\mathrm{dec}}, r_{\mathrm{dec}})$ be the plug-in policy of (10). For any distribution $\mathcal{D}$,*

$$\mathcal{E}_\perp(f_{p,u}) - \mathcal{E}_\perp^\star \le \max\left\{2\sqrt{2}, \sqrt{\frac{2J}{\lambda}}\right\} \cdot \sqrt{\mathcal{E}_{\mathrm{dec}}(p,u) - \mathcal{E}^\star_{\mathrm{dec}}}\,. \tag{13}$$

In particular, the theorem implies Bayes consistency: along any sequence of measurable estimates $(p_n, u_n)$ whose decoupled surrogate excess risk $\mathcal{E}_{\mathrm{dec}}(p_n,u_n) - \mathcal{E}^\star_{\mathrm{dec}}$ tends to zero, the right-hand side of (13) tends to zero, and therefore the L2D excess risk $\mathcal{E}_\perp(f_{p_n,u_n}) - \mathcal{E}_\perp^\star$ also tends to zero.

The bound has the $\sqrt{\cdot}$ calibration form (Mao et al., 2023b). The constant $\max\{2\sqrt{2}, \sqrt{2J/\lambda}\}$ equals $2\sqrt{2}$ whenever the per-expert weight satisfies $\lambda/J \ge 1/4$, and grows as $\sqrt{2J/\lambda}$ otherwise. The $1/J$ normalization in the surrogate (9) is precisely what makes this constant $J$-independent: scaling $\lambda$ with $J$ (equivalently, holding the per-expert weight $\lambda/J$ fixed) leaves the transfer guarantee unchanged as the expert pool grows, mirroring the gradient stability of Proposition 9.

**Comparison with existing bounds.** Two prior quantitative references are relevant. First, Mao et al. (2024e) analyze the earlier comp-sum reduction that includes the Add. CE surrogate (Mozannar & Sontag, 2020); in our classification-specialized setting, that yields the transfer constant $\sqrt{2(J+1)}$. Second, the Mao25 (Mao et al., 2025d) single-stage surrogate has an excess-risk constant that scales with the augmented action dimension $K+J$. The decoupled surrogate is the only multi-expert surrogate in our comparison whose transfer constant is $O(1)$ *in* $J$ (for fixed per-expert weight $\lambda/J$). For the remaining baselines (PiCCE, A-SM, OvA), the cited analyses establish Bayes consistency but do not give an explicit transfer constant (Appendix I).

## 6. Experiments

We evaluate the decoupled surrogate on four synthetic suites (each targeting one proposition from Section 4) and three real-data benchmarks spanning vision, human annotators, and tabular model routing. All methods share the same backbone and training budget within each benchmark; full details are in the appendix.

**Synthetic suites (Appendix K).** Each suite isolates one surrogate pathology by varying only the structural quantity relevant to the corresponding proposition while keeping the Bayes problem analytically known. *(i)* Nested redundant experts, targeting Add. CE amplification (Prop. 4): at $J$=24, defer regret $0.0002$ vs. $0.348$ for Add. CE (best baseline: PiCCE at $0.238$). *(ii)* Rare specialist, targeting PiCCE starvation (Prop. 5): on shared-correct samples, the decoupled surrogate selects the specialist with probability $0.994$ vs. $0$ for PiCCE. *(iii)* Shared acceptability, targeting Mao25's set-mass collapse (Prop. 6): the decoupled surrogate reaches defer regret $0.0008$ vs. $0.154$ for Mao25. *(iv)* Class–expert coupling, targeting A-SM (Prop. 7): A-SM's mixed class–expert Hessian block grows from $0.12$ at $J$=1 to $0.48$ at $J$=17 ($\approx 0.12\sqrt{J}$), while the decoupled surrogate's is identically zero by construction (Appendix K.4).

**CIFAR-10 with synthetic experts (Appendix L).** We train a ResNet-12 end-to-end on CIFAR-10 with $J$ nested redundant experts whose correctness is drawn from a shared latent (Section K.1). Table 2 reports system and classifier accuracy. The decoupled surrogate is the only method in our comparison that *improves* over the standalone classifier at every $J$; Add. CE, A-SM, and OvA degrade as the redundant pool grows. A-SM's classifier accuracy collapses from .831 ($J$=8) to .680 ($J$=32), consistent with the softmax-coupling pathology of Proposition 7.

**CIFAR-10H with human annotators (Appendix M).** We freeze a pretrained ResNet-12 and train only the L2D head on CIFAR-10H (Peterson et al., 2019) with $J$ real an-

*Table 2.* CIFAR-10 with redundant synthetic experts (mean±std, 4 seeds). Best in **bold**; second-best underlined.

| | Sys. Acc. ↑ | | | Cls. Acc. ↑ | | |
|---|---|---|---|---|---|---|
| Method | $J$=8 | $J$=16 | $J$=32 | $J$=8 | $J$=16 | $J$=32 |
| **Decoupled** | $\mathbf{.918}_{\pm.001}$ | $\mathbf{.919}_{\pm.003}$ | $\mathbf{.919}_{\pm.002}$ | $\underline{.909}_{\pm.001}$ | $\underline{.909}_{\pm.002}$ | $\mathbf{.908}_{\pm.003}$ |
| Classif. only | $\underline{.910}_{\pm.002}$ | $\underline{.911}_{\pm.001}$ | $\underline{.908}_{\pm.002}$ | $\mathbf{.910}_{\pm.002}$ | $\mathbf{.911}_{\pm.001}$ | $\mathbf{.908}_{\pm.002}$ |
| OvA | $.901_{\pm.001}$ | $.890_{\pm.002}$ | $.880_{\pm.003}$ | $.875_{\pm.001}$ | $.859_{\pm.004}$ | $.844_{\pm.003}$ |
| PiCCE | $.858_{\pm.007}$ | $.860_{\pm.011}$ | $.859_{\pm.008}$ | $.888_{\pm.008}$ | $.895_{\pm.006}$ | $.880_{\pm.014}$ |
| A-SM | $.860_{\pm.002}$ | $.827_{\pm.003}$ | $.783_{\pm.006}$ | $.831_{\pm.002}$ | $.772_{\pm.003}$ | $.680_{\pm.009}$ |
| Add. CE | $.815_{\pm.002}$ | $.790_{\pm.013}$ | $.712_{\pm.015}$ | $.902_{\pm.004}$ | $.895_{\pm.003}$ | $.872_{\pm.009}$ |
| Mao25 | $.703_{\pm.000}$ | $.703_{\pm.000}$ | $.703_{\pm.000}$ | $.105_{\pm.024}$ | $.111_{\pm.010}$ | $.105_{\pm.025}$ |

notators subsampled from the crowd (5 draws, Table 3). The decoupled surrogate leads system accuracy at every $J$ with stable classifier quality ($\geq$.889); A-SM collapses from .890 ($J$=5) to .470 ($J$=20) and OvA becomes unstable (std .203 at $J$=10).

**Covertype with model experts (Appendix N).** On Forest CoverType (Blackard & Dean, 1999) with $J$=5 heterogeneous pretrained model experts (oracle $+9.8$ pp over the best single model), the decoupled surrogate reaches system accuracy $0.934$ and classifier accuracy $0.941$—the only L2D method in our comparison to improve over the standalone classifier ($0.929$); all baselines degrade and Mao25 collapses to $100\%$ deferral.

## 7. Conclusion

Every augmented-action surrogate for multi-expert L2D trades a fix on one axis for a failure on the other. The decoupled surrogate resolves both by estimating $\eta$ with a softmax and each $\alpha_j$ with an independent sigmoid, yielding coordinatewise gradients, a zero class–expert Hessian block, and a consistency bound whose constant is $J$-independent for fixed $\lambda/J$. Across the synthetic suites and real-data benchmarks, it is the only method in our comparison that avoids amplification and preserves rare specialists, while improving over the standalone classifier on the real-data benchmarks.

## References


Awasthi, P., Mao, A., Mohri, M., and Zhong, Y. Multi-class h-consistency bounds. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Bartlett, P., Jordan, M., and McAuliffe, J. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101:138–156, 02 2006. doi: 10.1198/016214505000000907.

Bartlett, P. L. and Wegkamp, M. H. Classification with a reject option using a hinge loss. *The Journal of Machine Learning Research*, 9:1823–1840, June 2008.

Blackard, J. A. and Dean, D. J. Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. *Computers and Electronics in Agriculture*, 24(3):131–151, 1999. doi: 10.1016/S0168-1699(99)00046-0.

Cao, Y., Cai, T., Feng, L., Gu, L., Gu, J., An, B., Niu, G., and Sugiyama, M. Generalizing consistent multi-class classification with rejection to be compatible with arbitrary losses. *Advances in neural information processing systems*, 35:521–534, 2022.

Cao, Y., Mozannar, H., Feng, L., Wei, H., and An, B. In defense of softmax parametrization for calibrated and consistent learning to defer. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2024. Curran Associates Inc.

Charusaie, M.-A., Mozannar, H., Sontag, D., and Samadi, S. Sample efficient learning of predictors that complement humans, 2022.

Chow, C. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, January 1970. doi: 10.1109/TIT.1970.1054406.

Cortes, C., DeSalvo, G., and Mohri, M. Learning with rejection. In Ortner, R., Simon, H. U., and Zilles, S. (eds.), *Algorithmic Learning Theory*, pp. 67–82, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46379-7.

Cortes, C., DeSalvo, G., and Mohri, M. Theory and algorithms for learning with rejection in binary classification. *Annals of Mathematics and Artificial Intelligence*, 92(2): 277–315, 2024a.

Cortes, C., Mao, A., Mohri, C., Mohri, M., and Zhong, Y. Cardinality-aware set prediction and top-$k$ classification. In *The Thirty-eighth Annual Conference on Neural*

*Table 3.* CIFAR-10H with real human annotators (mean±std, 5 draws). Best in **bold**; second-best underlined.

| | Sys. Acc. ↑ | | | Cls. Acc. ↑ | | |
|---|---|---|---|---|---|---|
| Method | $J$=5 | $J$=10 | $J$=20 | $J$=5 | $J$=10 | $J$=20 |
| **Decoupled** | $\mathbf{.960}_{\pm.005}$ | $\mathbf{.958}_{\pm.003}$ | $\mathbf{.961}_{\pm.004}$ | $\mathbf{.892}_{\pm.001}$ | $\mathbf{.892}_{\pm.003}$ | $.889_{\pm.001}$ |
| OvA | $\underline{.959}_{\pm.005}$ | $.954_{\pm.008}$ | $\underline{.956}_{\pm.006}$ | $.889_{\pm.001}$ | $.799_{\pm.203}$ | $\mathbf{.890}_{\pm.002}$ |
| A-SM | $.959_{\pm.004}$ | $.952_{\pm.007}$ | $.949_{\pm.005}$ | $\underline{.890}_{\pm.002}$ | $.747_{\pm.313}$ | $.470_{\pm.435}$ |
| PiCCE | $.954_{\pm.003}$ | $\underline{.957}_{\pm.004}$ | $.953_{\pm.002}$ | $.889_{\pm.001}$ | $\underline{.890}_{\pm.002}$ | $\underline{.890}_{\pm.001}$ |
| Add. CE | $.953_{\pm.004}$ | $.952_{\pm.004}$ | $.951_{\pm.003}$ | $.889_{\pm.002}$ | $.888_{\pm.001}$ | $.742_{\pm.331}$ |
| Mao25 | $.950_{\pm.005}$ | $.948_{\pm.003}$ | $.948_{\pm.004}$ | $.116_{\pm.133}$ | $.077_{\pm.024}$ | $.045_{\pm.014}$ |

*Information Processing Systems*, 2024b. URL https://openreview.net/forum?id=WAT3qu737X.

Cortes, C., Mao, A., Mohri, M., and Zhong, Y. Balancing the scales: A theoretical and algorithmic framework for learning from imbalanced data. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025a.

Cortes, C., Mao, A., Mohri, M., and Zhong, Y. Improved balanced classification with theoretically grounded loss functions. In *Advances in Neural Information Processing Systems 39 (NeurIPS 2025)*, 2025b.

DeSalvo, G., Mohri, C., Mohri, M., and Zhong, Y. Budgeted multiple-expert deferral, 2025. URL https://arxiv.org/abs/2510.26706.

Geifman, Y. and El-Yaniv, R. Selective classification for deep neural networks. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/4a8423d5e91fda00bb7e46540e2b0cf1-Paper.pdf.

Ghosh, A., Kumar, H., and Sastry, P. S. Robust loss functions under label noise for deep neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

Joshi, S., Parbhoo, S., and Doshi-Velez, F. Learning-to-defer for sequential medical decision-making under uncertainty. *arXiv preprint arXiv:2109.06312*, 2021.

Liu, S., Cao, Y., Zhang, Q., Feng, L., and An, B. Mitigating underfitting in learning to defer with consistent losses. In *International Conference on Artificial Intelligence and Statistics*, pp. 4816–4824. PMLR, 2024.

Liu, S., Cao, Y., Feng, L., An, B., and Ong, L. When more experts hurt: Underfitting in multi-expert learning to defer. *arXiv preprint arXiv:2602.17144*, 2026.

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

Mao, A. *Theory and Algorithms for Learning with Multi-Class Abstention and Multi-Expert Deferral*. PhD thesis, New York University, 2025.

Mao, A., Mohri, C., Mohri, M., and Zhong, Y. Two-stage learning to defer with multiple experts. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a. URL https://openreview.net/forum?id=GIlsH0T4b2.

Mao, A., Mohri, M., and Zhong, Y. Cross-entropy loss functions: Theoretical analysis and applications. In *International conference on Machine learning*, pp. 23803–23828. PMLR, 2023b.

Mao, A., Mohri, M., and Zhong, Y. Predictor-rejector multi-class abstention: Theoretical analysis and algorithms. In *International Conference on Algorithmic Learning Theory*, pp. 822–867. PMLR, 2024a.

Mao, A., Mohri, M., and Zhong, Y. $h$-consistency bounds: Characterization and extensions. *Advances in Neural Information Processing Systems*, 36, 2024b.

Mao, A., Mohri, M., and Zhong, Y. $\mathcal{H}$-consistency guarantees for regression. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 34712–34737, 2024c.

Mao, A., Mohri, M., and Zhong, Y. Multi-label learning with stronger consistency guarantees. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2024)*, 2024d.

Mao, A., Mohri, M., and Zhong, Y. Principled approaches for learning to defer with multiple experts. In *ISAIM*, 2024e.

Mao, A., Mohri, M., and Zhong, Y. Realizable $h$-consistent and bayes-consistent loss functions for learning to defer.

*Advances in neural information processing systems*, 37: 73638–73671, 2024f.

Mao, A., Mohri, M., and Zhong, Y. Regression with multi-expert deferral. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024g.

Mao, A., Mohri, M., and Zhong, Y. A universal growth rate for learning with smooth surrogate losses. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2024)*, 2024h.

Mao, A., Mohri, M., and Zhong, Y. Theoretically grounded loss functions and algorithms for score-based multi-class abstention. In Dasgupta, S., Mandt, S., and Li, Y. (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4753–4761. PMLR, 02–04 May 2024i. URL https://proceedings.mlr.press/v238/mao24a.html.

Mao, A., Mohri, M., and Zhong, Y. Beyond tsybakov: Model margin noise and $\mathcal{H}$-consistency bounds, 2025a.

Mao, A., Mohri, M., and Zhong, Y. Enhanced $\mathcal{H}$-consistency bounds. In *Proceedings of the 36th International Conference on Algorithmic Learning Theory (ALT 2025)*, 2025b.

Mao, A., Mohri, M., and Zhong, Y. Principled algorithms for optimizing generalized metrics in binary classification. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025c.

Mao, A., Mohri, M., and Zhong, Y. Mastering multiple-expert routing: Realizable $h$-consistency and strong guarantees for learning to defer. In *Forty-second International Conference on Machine Learning*, 2025d.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of machine learning*. MIT Press, 2012.

Montreuil, Y., Carlier, A., Ng, L. X., and Ooi, W. T. Adversarial robustness in two-stage learning-to-defer: Algorithms and guarantees. In *Forty-second International Conference on Machine Learning*, 2025a.

Montreuil, Y., Heng, Y. S., Carlier, A., Ng, L. X., and Ooi, W. T. A two-stage learning-to-defer approach for multi-task learning. In *Forty-second International Conference on Machine Learning*, 2025b.

Montreuil, Y., Carlier, A., Ng, L. X., and Ooi, W. T. Why ask one when you can ask $k$? learning-to-defer to the top-$k$ experts. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=mGbEv4kVoG.

Montreuil, Y., Dang, H. D., Meyer, M., Ng, L. X., Carlier, A., and Ooi, W. T. Online learning-to-defer with varying experts. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026b. URL https://openreview.net/forum?id=1lix8ppUJ7.

Montreuil, Y., Heng, Y. S., Ng, L. X., Carlier, A., and Ooi, W. T. Optimal query allocation in extractive QA with LLMs: A learning-to-defer framework with theoretical guarantees. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026c. URL https://openreview.net/forum?id=kEVupwepTq.

Montreuil, Y., Ng, L. X., Carlier, A., Ooi, W. T., and Letian, Y. Adversarial robustness in one-stage learning-to-defer. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026d. URL https://openreview.net/forum?id=gF8tv1SaR2.

Montreuil, Y., Yu, L., Carlier, A., Ng, L. X., and Ooi, W. T. Learning to defer in non-stationary time series via switching state-space models. *arXiv preprint arXiv:2601.22538*, 2026e.

Mozannar, H. and Sontag, D. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Mozannar, H., Lang, H., Wei, D., Sattigeri, P., Das, S., and Sontag, D. A. Who should predict? exact algorithms for learning to defer to humans. In *International Conference on Artificial Intelligence and Statistics*, 2023. URL https://api.semanticscholar.org/CorpusID:255941521.

Narasimhan, H., Jitkrittum, W., Menon, A. K., Rawat, A., and Kumar, S. Post-hoc estimators for learning to defer to an expert. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 29292–29304. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/bc8f76d9caadd48f77025b1c889d2e2d-Paper-Conference.pdf.

Nguyen, C. C., Do, T.-T., and Carneiro, G. Probabilistic learning to defer: Handling missing expert annotations and controlling workload distribution. In *The Thirteenth International Conference on Learning Representations*, 2025.

Ni, C., Charoenphakdee, N., Honda, J., and Sugiyama, M. On the calibration of multiclass classification with rejection. *Advances in neural information processing systems*, 32, 2019.

Peterson, J. C., Battleday, R. M., Griffiths, T. L., and Russakovsky, O. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9617–9626, 2019.

Steinwart, I. How to compare different loss functions and their risks. *Constructive Approximation*, 26:225–287, 2007. URL https://api.semanticscholar.org/CorpusID:16660598.

Strong, J., Men, Q., and Noble, J. A. Trustworthy and practical ai for healthcare: A guided deferral system with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 28413–28421, 2025.

Strong, J., Saha, P., Ibrahim, Y., Ouyang, C., and Noble, A. Identity-free deferral for unseen experts. In *The Fourteenth International Conference on Learning Representations*, 2026.

Tailor, D., Patra, A., Verma, R., Manggala, P., and Nalisnick, E. Learning to defer to a population: A meta-learning approach. In Dasgupta, S., Mandt, S., and Li, Y. (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 3475–3483. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/tailor24a.html.

Tewari, A. and Bartlett, P. L. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8(36):1007–1025, 2007. URL http://jmlr.org/papers/v8/tewari07a.html.

Verma, R., Barrejon, D., and Nalisnick, E. Learning to defer to multiple experts: Consistent surrogate losses, confidence calibration, and conformal ensembles. In *International Conference on Artificial Intelligence and Statistics*, 2022. URL https://api.semanticscholar.org/CorpusID:253237048.

Wei, Z., Cao, Y., and Feng, L. Exploiting human-ai dependence for learning to defer. In *Forty-first International Conference on Machine Learning*, 2024.

Zhang, T. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, 32, 12 2002. doi: 10.1214/aos/1079120130.

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

Zhong, Y. *Fundamental Novel Consistency Theory: $\mathcal{H}$-Consistency Bounds*. PhD thesis, New York University, 2025.

# A. Extended Related Work

Learning-to-defer extends selective prediction and classification with a reject option (Chow, 1970; Bartlett & Wegkamp, 2008; Geifman & El-Yaniv, 2017; Cortes et al., 2016; Cao et al., 2022; Cortes et al., 2024a) and is also closely related to the abstention literature (Mao et al., 2024a;i; Cortes et al., 2024b). The distinguishing feature of L2D is that abstention is replaced by routing to an external expert. The modern L2D literature begins with adaptive deferral in the single-expert setting (Madras et al., 2018; Mozannar & Sontag, 2020).

**One-stage methods.** One-stage methods learn prediction and deferral jointly. Madras et al. (2018) introduced adaptive deferral in the single-expert setting, while Mozannar & Sontag (2020) gave the modern multiclass formulation by casting learning to defer as prediction over an augmented action space. Most subsequent one-stage work refines this augmented-action view rather than departing from it. Verma et al. (2022) compare coupled softmax and one-vs-all surrogates for multiple experts; Cao et al. (2024) analyze the calibration pathology of symmetric softmax reductions and propose an asymmetric softmax alternative; Charusaie et al. (2022) study complementarity and sample efficiency; Mozannar et al. (2023) analyze realizability and exact optimization; Wei et al. (2024) incorporate human–AI dependence into surrogate design; and Liu et al. (2024) show that even statistically consistent single-expert surrogates can underfit when non-zero expert-query costs introduce redundant label smoothing. In the multi-expert setting, Mao et al. (2024e) derive quantitative guarantees for comp-sum-style reductions, Mao et al. (2024f; 2025d) develop consistent loss families for learning to defer, and Liu et al. (2026) identify redundancy-driven underfitting and propose PiCCE as a correction.

More recent extensions that continue to optimize prediction and deferral jointly also fit naturally in the one-stage category. These include top-$k$ deferral (Montreuil et al., 2026a), adversarially robust one-stage L2D (Montreuil et al., 2026d), non-stationary time-series deferral (Montreuil et al., 2026e), missing-annotation variants (Nguyen et al., 2025), deferral to a population or to unseen experts (Tailor et al., 2024; Strong et al., 2026), and application-driven one-stage systems in sequential medicine and healthcare (Joshi et al., 2021; Strong et al., 2025). Overall, the one-stage literature is best viewed as a sequence of refinements and extensions of the augmented-action formulation introduced by Mozannar & Sontag (2020), first in the single-expert $K + 1$ setting and then in the multi-expert $K + J$ setting.

**Two-stage methods.** By contrast, *two-stage* methods first train a predictor and then learn a defer rule or router on top of the frozen predictor. This includes post-hoc estimators for deferring to a single expert (Narasimhan et al., 2022), principled two-stage formulations for multiple experts (Mao et al., 2023a), and extensions to multitask and regression settings (Montreuil et al., 2025b; Mao et al., 2024g). Recent extensions of the two-stage line also include adversarial robustness guarantees (Montreuil et al., 2025a), online learning with varying experts (Montreuil et al., 2026b), budgeted L2D (DeSalvo et al., 2025), and extractive question answering (Montreuil et al., 2026c).

The distinction is important: one-stage methods redesign the joint prediction–deferral surrogate, whereas two-stage methods optimize only the routing rule after the predictor has been fixed. Our decoupled parameterization is introduced here as a one-stage surrogate, but the same separation between class prediction and expert-utility estimation can also be used within a two-stage pipeline.

**Quantitative consistency bounds.** Classical Bayes consistency (Bartlett et al., 2006; Tewari & Bartlett, 2007; Zhang, 2002; Steinwart, 2007) guarantees that the conditional minimizer of the surrogate recovers the Bayes decision. Quantitative versions turn surrogate excess risk into target excess risk through an explicit calibration function. This line includes multiclass and cross-entropy analyses (Awasthi et al., 2022; Mao et al., 2024b; 2023b), smooth surrogates (Mao et al., 2024h), regression (Mao et al., 2024c), multi-label learning (Mao et al., 2024d), abstention (Mao et al., 2024i;a; Cortes et al., 2024b), imbalanced and balanced binary classification (Cortes et al., 2025a;b; Mao et al., 2025c), sharper enhanced bounds (Mao et al., 2025b), and noise-adaptive refinements (Mao et al., 2025a); see Zhong (2025); Mao (2025) for unified treatments. Theorem 10 gives such a $\sqrt{\cdot}$-form bound for the decoupled surrogate, with calibration constant $J$-independent for fixed per-expert weight $\lambda/J$.

**Positioning.** Our work is closest to the one-stage literature, but differs in aim. Rather than proposing another repair within an augmented-action parameterization, we compare representative one-stage surrogates along two axes—population target and local optimization geometry—and then propose a decoupled surrogate outside the shared-action-space design. In that sense, our contribution is both comparative and constructive: we isolate a recurrent tradeoff within existing multi-expert surrogates and show that the decoupled surrogate avoids it while retaining a quantitative consistency guarantee.

## B. Proof of Lemma 2

**Lemma 2** (Bayes Rule for Multi-Expert L2D)**.** *For every $x \in \mathcal{X}$, a Bayes-optimal decision is obtained by*

$$\begin{aligned} r^\star(x) &= 0 \quad \textit{if } \max_k \eta_k(x) \geq \max_j \alpha_j(x), \\ r^\star(x) &\in \arg\max_{j\in[J]} \alpha_j(x) \quad \textit{otherwise.} \end{aligned} \tag{2}$$

*with $h^\star(x) \in \arg\max_{k\in[K]} \eta_k(x)$.*

**Proof** We prove the statement directly from the true defer loss (1). Fix an input $x \in \mathcal{X}$ and condition on the event $X = x$. For a decision pair $(h, r)$, the corresponding conditional true deferral risk is

$$\mathcal{C}_\perp(h, r \mid x) \;:=\; \mathbb{E}[\ell_\perp(h, r;\, X, Y, \mathbf{M}) \mid X = x]\,.$$

Substituting the definition of $\ell_\perp$ gives

$$\begin{aligned} \mathcal{C}_\perp(h, r \mid x) &= \mathbb{E}\left[\mathbf{1}\{h(x) \neq Y\}\mathbf{1}\{r(x) = 0\} + \sum_{j=1}^{J} \mathbf{1}\{M_j \neq Y\}\mathbf{1}\{r(x) = j\} \;\middle|\; X = x\right] \\ &= \mathbf{1}\{r(x) = 0\}\Pr\big(h(x) \neq Y \mid X = x\big) + \sum_{j=1}^{J} \mathbf{1}\{r(x) = j\}\Pr\big(M_j \neq Y \mid X = x\big). \end{aligned} \tag{14}$$

Thus, once $x$ is fixed, the learner is simply choosing one action among the $K$ class actions and the $J$ defer actions, and each action has an explicit conditional error probability.

Take a class label $k \in [K]$ and consider the action "predict class $k$", that is, choose $h(x) = k$ and $r(x) = 0$. By (14), its conditional risk is

$$\mathcal{C}_\perp(k, 0 \mid x) = \Pr(Y \neq k \mid X = x).$$

Since $\eta_k(x) = \Pr(Y = k \mid X = x)$ and the label space is multiclass,

$$\Pr(Y \neq k \mid X = x) = 1 - \Pr(Y = k \mid X = x) = 1 - \eta_k(x).$$

Therefore the best classifier action at $x$ is any

$$h^\star(x) \in \arg\max_{k\in[K]} \eta_k(x),$$

and its conditional risk is

$$\min_{k\in[K]} \mathcal{C}_\perp(k, 0 \mid x) = 1 - \max_{k\in[K]} \eta_k(x).$$

Now take an expert index $j \in [J]$ and consider the action "defer to expert $j$", that is, choose $r(x) = j$. Again by (14),

$$\mathcal{C}_\perp(\perp_j \mid x) = \Pr(M_j \neq Y \mid X = x).$$

By definition of the expert utility $\alpha_j(x) = \Pr(M_j = Y \mid X = x)$, we obtain

$$\Pr(M_j \neq Y \mid X = x) = 1 - \alpha_j(x).$$

Hence the best defer action at $x$ is any

$$j^\star(x) \in \arg\max_{j\in[J]} \alpha_j(x),$$

and its conditional risk is

$$\min_{j\in[J]} \mathcal{C}_\perp(\perp_j \mid x) = 1 - \max_{j\in[J]} \alpha_j(x).$$

**Step 3: compare the best class action and the best defer action.** The Bayes decision at $x$ is the action with the smallest conditional true deferral risk among all $K + J$ available actions. Combining the two displays above,

$$\begin{aligned}\inf_{h,r} \mathcal{C}_{\perp}(h, r \mid x) &= \min\Big\{1 - \max_{k\in[K]} \eta_k(x),\ 1 - \max_{j\in[J]} \alpha_j(x)\Big\} \\ &= 1 - \max\Big\{\max_{k\in[K]} \eta_k(x),\ \max_{j\in[J]} \alpha_j(x)\Big\}.\end{aligned}$$

Therefore:

- if $\max_{k\in[K]} \eta_k(x) \geq \max_{j\in[J]} \alpha_j(x)$, then the best class action is at least as good as every defer action, so one may choose $r^\star(x) = 0$ together with any $h^\star(x) \in \arg\max_{k\in[K]} \eta_k(x)$;
- if $\max_{k\in[K]} \eta_k(x) < \max_{j\in[J]} \alpha_j(x)$, then deferring is strictly better, and one may choose $r^\star(x) \in \arg\max_{j\in[J]} \alpha_j(x)$.

This is exactly the rule stated in (2). ■

## C. The Decoupled Surrogate

We present the decoupled surrogate first because it is the reference point for all of the baseline diagnoses that follow. The later sections can be read as controlled departures from this aligned formulation. The decoupled surrogate abandons the augmented-action reduction and instead learns the class posterior and the expert utilities in their native forms.

### C.1. Formulation

**Cost-sensitive formulation.** Before specializing to the zero-one defer loss, it is useful to record the more general bounded-cost version of the decoupled construction. Let $c_j(x, y) \in [0, 1]$ denote the cost of routing $(x, y)$ to expert $j$, and let

$$\tau_j(x, y) \coloneqq 1 - c_j(x, y) \in [0, 1]$$

denote the corresponding utility target. The cost-sensitive decoupled surrogate is

$$\Phi_\lambda^{\mathrm{dec,cs}}(w, s;\, x, y) \coloneqq -\log p_y(x) - \frac{\lambda}{J} \sum_{j=1}^{J} \Big(\tau_j(x, y) \log u_j(x) \ +\ \big(1 - \tau_j(x, y)\big) \log(1 - u_j(x))\Big).$$

Thus the class side is unchanged, while each expert head is trained with a Bernoulli cross-entropy against a possibly soft utility target $\tau_j(x, y)$.

The decoupled surrogate uses one softmax head for the classifier and one sigmoid head for each expert. The softmax head models the categorical class posterior, while the expert heads model the Bernoulli events "expert $j$ is correct on this input".

**Definition 11** (Decoupled surrogate)**.** Let $w(x) \in \mathbb{R}^K$ denote class logits and $s(x) \in \mathbb{R}^J$ denote expert logits. Define

$$p_k(x) \coloneqq \frac{\exp(w_k(x))}{\sum_{\ell=1}^{K} \exp(w_\ell(x))}, \qquad u_j(x) \coloneqq \sigma(s_j(x)).$$

For $\lambda > 0$, define

$$\Phi_\lambda^{\mathrm{dec}}(w, s;\, x, y, m) \coloneqq -\log p_y(x) - \frac{\lambda}{J} \sum_{j=1}^{J} \Big(t_j \log u_j(x) \ +\ (1 - t_j) \log(1 - u_j(x))\Big),$$

where $t_j \coloneqq \mathbf{1}\{m_j = y\}$.

Prediction compares the classifier confidence $\max_{k\in[K]} p_k(x)$ with the largest estimated expert utility $\max_{j\in[J]} u_j(x)$ and defers whenever the latter is larger. This is the classification-specialized version of the general bounded-cost formulation above. In the present paper,

$$c_j(x, y) = \mathbf{1}\{m_j \neq y\}, \qquad \tau_j(x, y) = \mathbf{1}\{m_j = y\},$$

so the expert-side target reduces exactly to binary correctness.

Two structural features are already visible from the definition. First, the class side remains a categorical distribution throughout training because $p(x) \in \Delta^K$ for every parameter value. Second, the expert side is coordinatewise: each $u_j(x)$ is a bounded Bernoulli probability attached to one expert only. The rest of the section formalizes the consequences of this design.

### C.2. Per-Sample Gradient Structure

**Proposition 9** (Per-sample gradient of the decoupled surrogate)**.** *For one sample* $(x, y, m)$,

$$\begin{aligned} \frac{\partial \Phi_\lambda^{\mathrm{dec}}}{\partial w_r} &= p_r(x) - \mathbf{1}\{r = y\}, \\ \frac{\partial \Phi_\lambda^{\mathrm{dec}}}{\partial s_j} &= \frac{\lambda}{J}\big(u_j(x) - t_j\big). \end{aligned} \tag{11}$$

**Proof** The samplewise decoupled loss is the sum of one multiclass cross-entropy term and $J$ independent Bernoulli cross-entropy terms. We derive the two pieces separately.

**Classifier head.** The classifier contribution is

$$\Phi_{\mathrm{cls}}(w; y) \;=\; -\log p_y \;=\; -w_y + \log \sum_{\ell=1}^{K} e^{w_\ell}.$$

Differentiate with respect to $w_r$:

$$\frac{\partial \Phi_{\mathrm{cls}}}{\partial w_r} \;=\; -\mathbf{1}\{r = y\} \;+\; \frac{e^{w_r}}{\sum_{\ell=1}^{K} e^{w_\ell}} \;=\; p_r - \mathbf{1}\{r = y\}.$$

This is the usual multiclass cross-entropy gradient.

**Expert heads.** For expert $j$, write $t_j = \mathbf{1}\{m_j = y\}$. Its contribution is

$$\Phi_j(s_j; t_j) \;=\; -t_j \log u_j - (1 - t_j)\log(1 - u_j), \qquad u_j = \sigma(s_j).$$

Differentiate with respect to $s_j$. By the chain rule,

$$\frac{\partial \Phi_j}{\partial s_j} \;=\; \left(-\frac{t_j}{u_j} + \frac{1 - t_j}{1 - u_j}\right)\frac{du_j}{ds_j}.$$

Since $\frac{du_j}{ds_j} = u_j(1 - u_j)$, we obtain

$$\frac{\partial \Phi_j}{\partial s_j} \;=\; \left(-\frac{t_j}{u_j} + \frac{1 - t_j}{1 - u_j}\right) u_j(1 - u_j) \;=\; u_j - t_j.$$

Multiplying by the prefactor $\lambda/J$ gives

$$\frac{\partial \Phi_\lambda^{\mathrm{dec}}}{\partial s_j} \;=\; \frac{\lambda}{J}(u_j - t_j).$$

This completes the gradient calculation. The important structural point is that the classifier residuals and the expert residuals live on disjoint coordinates: the class gradient depends only on $(w, y)$, and the $j$-th expert gradient depends only on $(s_j, t_j)$. ■

This proposition isolates the first-order reason the decoupled surrogate behaves differently from the augmented-action surrogates. A sample contributes one classifier residual and $J$ expert residuals, but these residuals do not compete through a shared normalization. Correct experts are reinforced independently of one another, and expert-side errors cannot directly distort the class-side update.

### C.3. Per-Sample Curvature Structure

We now derive the Hessian explicitly. The goal is to show not only that the curvature is bounded, but also that the class and expert directions remain decoupled at second order.

Start with the classifier block. From the softmax Jacobian,

$$\frac{\partial p_r}{\partial w_\ell} \;=\; p_r(\mathbf{1}\{r=\ell\} - p_\ell),$$

so the classifier Hessian is

$$\nabla_w^2(-\log p_y) \;=\; \mathrm{Diag}(p) - pp^\top.$$

This is the usual softmax covariance matrix. It captures competition only among the class coordinates themselves. For the expert side, differentiate the gradient $\frac{\lambda}{J}(u_j - t_j)$ once more:

$$\frac{\partial^2 \Phi_\lambda^{\mathrm{dec}}}{\partial s_j^2} \;=\; \frac{\lambda}{J}\frac{du_j}{ds_j} \;=\; \frac{\lambda}{J}u_j(1-u_j),$$

and for $j \neq \ell$,

$$\frac{\partial^2 \Phi_\lambda^{\mathrm{dec}}}{\partial s_j\,\partial s_\ell} \;=\; 0,$$

because $\Phi_j$ depends only on $s_j$.

Finally, the class term depends only on $w$, and the expert terms depend only on $s$. Therefore every mixed class–expert derivative is zero:

$$\frac{\partial^2 \Phi_\lambda^{\mathrm{dec}}}{\partial w_r\,\partial s_j} \;=\; 0 \qquad \text{for all } r \in [K],\ j \in [J].$$

Hence the full Hessian is block diagonal:

$$\nabla^2_{(w,s)}\Phi_\lambda^{\mathrm{dec}} \;=\; \begin{pmatrix} \mathrm{Diag}(p) - pp^\top & 0 \\ 0 & \frac{\lambda}{J}\,\mathrm{Diag}\big(u_j(1-u_j)\big) \end{pmatrix}.$$

The block-diagonal form makes the spectral analysis immediate. The eigenvalues of a block-diagonal matrix are the union of the eigenvalues of its diagonal blocks, so the largest eigenvalue of the full Hessian is the larger of the largest class-block eigenvalue and the largest expert-block eigenvalue.

For the class block, $\mathrm{Diag}(p) - pp^\top$ is a softmax covariance matrix. Its largest eigenvalue is at most $1/2$ (Mohri et al., 2012). Intuitively, this is the intrinsic curvature scale of ordinary multiclass cross-entropy. The decoupled surrogate does not increase that scale.

For the expert block, the matrix is diagonal, so its eigenvalues are exactly the diagonal entries

$$\frac{\lambda}{J}u_j(1-u_j), \qquad j \in [J].$$

Since $0 \le u_j(1-u_j) \le 1/4$ for every $j$, each expert-side eigenvalue is bounded by $\lambda/(4J)$. Therefore

$$\lambda_{\max}\big(\nabla^2_{(w,s)}\Phi_\lambda^{\mathrm{dec}}\big) \;\le\; \max\left\{\frac{1}{2},\ \frac{\lambda}{4J}\right\}.$$

This inequality is informative in two ways. First, the class-side curvature is uniformly bounded by the same constant as ordinary multiclass cross-entropy. Second, the expert-side curvature becomes smaller as the number of experts grows if the total expert weight $\lambda$ is held fixed. In particular, adding experts does not create the kind of growing stiffness that appears in additive CE, nor the mixed class–expert second-order interaction that appears in A-SM.

This is the local-geometric signature of the decoupled surrogate. There is no multiplicity factor, no winner-take-all routing effect, and no mixed class–expert block. Each expert curvature term is also bounded by $(\lambda/J)/4$, so adding more experts does not by itself create an increasingly stiff optimization landscape. The classifier head and expert heads interact only through the final routing rule, not through the samplewise training geometry.

### C.4. The Role of $\lambda/J$ and How to Choose $\lambda$

The previous two subsections already show that the quantity that matters structurally is not $\lambda$ by itself, but the per-expert weight $\lambda/J$. This quantity controls the scale of every expert-side update. In the gradient formula,

$$\frac{\partial \Phi_\lambda^{\mathrm{dec}}}{\partial s_j} \;=\; \frac{\lambda}{J}\,(u_j - t_j),$$

the magnitude of the $j$-th expert residual is proportional to $\lambda/J$. In the Hessian,

$$\frac{\partial^2 \Phi_\lambda^{\mathrm{dec}}}{\partial s_j^2} \;=\; \frac{\lambda}{J}\,u_j(1-u_j),$$

the expert-side curvature is also proportional to $\lambda/J$. Thus the same quantity governs both first-order and second-order expert-side scale.

This observation is practically important when the number of experts varies. If one were to keep $\lambda$ fixed while increasing $J$, then $\lambda/J$ would decrease. Every expert gradient would shrink, every expert curvature term would shrink, and the optimization would place progressively less emphasis on fitting the expert utilities. In other words, a fixed $\lambda$ implicitly weakens the expert task as more experts are added. That is usually not the intended comparison across different expert pools.

By contrast, keeping the per-expert weight $\lambda/J$ fixed means scaling $\lambda$ with $J$. Under this scaling, the per-expert gradient magnitude remains on the same order, the per-expert curvature bound remains $(\lambda/J)/4$, and the transfer constant in Theorem 10 remains

$$\max\left\{2\sqrt{2}, \sqrt{\frac{2J}{\lambda}}\right\},$$

which is independent of $J$ as long as $\lambda/J$ is held fixed.

**One hyperparameter, two lenses.** It is worth emphasizing that $\lambda$ and $\lambda/J$ are not two independent knobs: they are the same quantity, parametrized differently. We keep both names because they serve different roles:

- $\lambda$ is the user-facing coefficient that appears in the loss definition (9) and is tuned on a validation split, matching the convention of single-expert L2D ($J$=1, $\lambda = \lambda/J$).
- $\lambda/J$ is the analysis-facing per-expert weight that appears in gradients, curvature bounds, and the transfer constant, and is the quantity one should keep fixed when comparing settings with different numbers of experts.

In practice, the user tunes $\lambda$ and reads the theory in terms of $\lambda/J$; the two views coincide whenever $J$ is fixed.

The existing theory does *not* identify a universally optimal choice of $\lambda$. The conditional minimizer is the same for every $\lambda > 0$, so there is no population-level "best" coefficient in the sense of changing the target being estimated. What $\lambda/J$ changes is the relative optimization emphasis between the classifier head and the expert heads in finite samples and under function-class constraints. That tradeoff depends on the optimization method, the capacity of the two heads, the amount of data, and the statistical difficulty of estimating $\eta$ versus the $\alpha_j$'s. Without additional assumptions on those ingredients, one should not expect a theorem to single out one universally optimal value.

The present theory nevertheless gives useful guidance.

First, Theorem 10 shows that the transfer constant equals $2\sqrt{2}$ whenever $\lambda/J \geq 1/4$, and deteriorates only when $\lambda/J < 1/4$. So values below $1/4$ are theoretically less attractive: they downweight the expert task and simultaneously weaken the calibration constant.

Second, the gradient and curvature formulas show what increasing $\lambda/J$ does. Larger $\lambda/J$ strengthens expert-side updates and increases expert-side curvature linearly, but it does *not* introduce class–expert coupling or amplification. So increasing $\lambda/J$ is not dangerous in the way increasing the number of experts is for the augmented-action baselines; it simply gives the expert task more optimization weight.

Third, because the class gradient is unaffected by $\lambda/J$, changing $\lambda/J$ only changes the balance between the classifier loss and the average expert loss. This suggests a simple practical rule:

- tune the per-expert weight $\lambda/J$, not $\lambda$ alone;
- keep $\lambda/J$ fixed when comparing settings with different numbers of experts;
- use $\lambda/J \geq 1/4$ as a natural default range if no stronger prior is available.

Among such values, the final choice should be selected by validation on the target routing objective. If the classifier is already strong and the main challenge is accurate expert selection, a larger $\lambda/J$ is often sensible. If the classifier head is harder to fit or the expert labels are especially noisy, a more moderate $\lambda/J$ may work better. The key point is that these are finite-sample optimization considerations, not changes in the population target.

### C.5. Consistency Bound

We now prove the main statistical guarantee of the decoupled surrogate.

**Theorem 10** (Consistency bound for the decoupled surrogate)**.** *Let $p\colon \mathcal{X} \to \Delta^K$ and $u\colon \mathcal{X} \to [0,1]^J$ be measurable, and let $f_{p,u} = (h_{\mathrm{dec}}, r_{\mathrm{dec}})$ be the plug-in policy of (10). For any distribution $\mathcal{D}$,*

$$\mathcal{E}_{\perp}(f_{p,u}) - \mathcal{E}_{\perp}^{\star} \leq \max\left\{2\sqrt{2}, \sqrt{\frac{2J}{\lambda}}\right\} \cdot \sqrt{\mathcal{E}_{\mathrm{dec}}(p,u) - \mathcal{E}_{\mathrm{dec}}^{\star}}\,. \tag{13}$$

**Proof** Throughout this proof we abbreviate the per-expert weight as $\beta \coloneqq \lambda/J$; this is a local shorthand for the fraction appearing in (9), not a separate hyperparameter. Fix $x$. We define

$$\delta_{\perp}(p,u \mid x) \coloneqq \mathcal{C}_{\perp}(f_{p,u} \mid x) - \mathcal{C}_{\perp}^{*}(x)$$

for the conditional L2D regret and

$$\delta_{\mathrm{dec}}(p,u \mid x) \coloneqq \mathcal{C}_{\mathrm{dec}}(p,u \mid x) - \mathcal{C}_{\mathrm{dec}}^{*}(x)$$

for the conditional surrogate excess.

Conditioning on $X = x$ in the samplewise surrogate gives

$$\begin{aligned}\mathcal{C}_{\mathrm{dec}}(p,u \mid x) = &-\sum_{k=1}^{K} \eta_k(x) \log p_k(x) \\ &- \beta \sum_{j=1}^{J} \Big(\alpha_j(x) \log u_j(x) + (1-\alpha_j(x)) \log(1-u_j(x))\Big),\end{aligned} \tag{15}$$

The class term depends only on $p(x)$, and the $j$-th expert term depends only on $u_j(x)$.

For any $q \in \Delta^K$,

$$-\sum_{k=1}^{K} \eta_k(x) \log q_k = H(\eta(x)) + \mathrm{KL}\big(\eta(x) \,\|\, q\big),$$

where $H(\eta(x)) = -\sum_k \eta_k(x) \log \eta_k(x)$. For any $v \in (0,1)$,

$$-\alpha_j(x) \log v - (1-\alpha_j(x)) \log(1-v) = H_{\mathrm{B}}(\alpha_j(x)) + \mathrm{KL}\Big(\mathrm{Bern}(\alpha_j(x)) \big\| \mathrm{Bern}(v)\Big),$$

where $H_{\mathrm{B}}(t) = -t \log t - (1-t)\log(1-t)$ is the Bernoulli entropy. Substituting these identities into (15) yields

$$\begin{aligned}\mathcal{C}_{\mathrm{dec}}(p,u \mid x) = &\, H(\eta(x)) + \beta \sum_{j=1}^{J} H_{\mathrm{B}}(\alpha_j(x)) \\ &+ \mathrm{KL}\big(\eta(x) \,\|\, p(x)\big) + \beta \sum_{j=1}^{J} \mathrm{KL}\Big(\mathrm{Bern}(\alpha_j(x)) \big\| \mathrm{Bern}(u_j(x))\Big).\end{aligned} \tag{16}$$

The first line is independent of $(p, u)$, so it is exactly the conditional minimum. Therefore

$$\delta_{\mathrm{dec}}(p, u \mid x) \;=\; \mathrm{KL}\big(\eta(x) \,\|\, p(x)\big) \;+\; \beta \sum_{j=1}^{J} \mathrm{KL}\Big(\mathrm{Bern}(\alpha_j(x)) \big\| \mathrm{Bern}(u_j(x))\Big). \tag{17}$$

This identity shows that the conditional surrogate excess is exactly the sum of one class estimation error and $J$ expert estimation errors, each measured in the natural KL divergence of its own subproblem. Since every KL term is nonnegative and vanishes only when its two arguments coincide, the conditional surrogate risk is minimized exactly when $p(x) = \eta(x)$ and $u_j(x) = \alpha_j(x)$ for every $j$.

We now translate this conditional estimation error into conditional L2D regret. Define

$$\varepsilon_{\mathrm{cls}}(x) \coloneqq \|p(x) - \eta(x)\|_\infty, \qquad \varepsilon_{\mathrm{exp}}(x) \coloneqq \|u(x) - \alpha(x)\|_\infty.$$

Also let $y^\star \in \arg\max_y \eta_y(x)$, $\hat{y} \in \arg\max_y p_y(x)$, $j^\star \in \arg\max_j \alpha_j(x)$, $\hat{j} \in \arg\max_j u_j(x)$, $a^\star = \eta_{y^\star}(x)$, and $b^\star = \alpha_{j^\star}(x)$.

If both the plug-in rule and the Bayes rule predict, then

$$\delta_\perp(p, u \mid x) = \eta_{y^\star}(x) - \eta_{\hat{y}}(x).$$

Since $\hat{y}$ maximizes $p$, we have $p_{\hat{y}} \geq p_{y^\star}$, and hence

$$\eta_{y^\star} - \eta_{\hat{y}} = (\eta_{y^\star} - p_{y^\star}) + (p_{y^\star} - p_{\hat{y}}) + (p_{\hat{y}} - \eta_{\hat{y}}) \leq 2\varepsilon_{\mathrm{cls}}(x).$$

If the plug-in rule predicts while the Bayes rule defers, then $b^\star > a^\star$ and $p_{\hat{y}} \geq u_{\hat{j}} \geq u_{j^\star}$. Therefore

$$\delta_\perp(p, u \mid x) = \alpha_{j^\star} - \eta_{\hat{y}} \leq (\alpha_{j^\star} - u_{j^\star}) + (u_{j^\star} - p_{\hat{y}}) + (p_{\hat{y}} - \eta_{\hat{y}}) \leq \varepsilon_{\mathrm{exp}}(x) + \varepsilon_{\mathrm{cls}}(x).$$

If both rules defer, then

$$\delta_\perp(p, u \mid x) = \alpha_{j^\star}(x) - \alpha_{\hat{j}}(x).$$

Since $\hat{j}$ maximizes $u$, the same argument gives

$$\alpha_{j^\star} - \alpha_{\hat{j}} \leq 2\varepsilon_{\mathrm{exp}}(x).$$

If the plug-in rule defers while the Bayes rule predicts, then $a^\star > b^\star$ and $u_{\hat{j}} > p_{\hat{y}} \geq p_{y^\star}$. Hence

$$\delta_\perp(p, u \mid x) = \eta_{y^\star} - \alpha_{\hat{j}} \leq (\eta_{y^\star} - p_{y^\star}) + (p_{y^\star} - u_{\hat{j}}) + (u_{\hat{j}} - \alpha_{\hat{j}}) \leq \varepsilon_{\mathrm{cls}}(x) + \varepsilon_{\mathrm{exp}}(x).$$

All four cases are therefore covered by

$$\delta_\perp(p, u \mid x) \leq 2 \max\{\varepsilon_{\mathrm{cls}}(x), \varepsilon_{\mathrm{exp}}(x)\}. \tag{18}$$

We next bound these two sup-norm errors by the conditional surrogate excess. Since $\|\cdot\|_\infty \leq \|\cdot\|_1$, Pinsker's inequality for categorical distributions gives

$$\varepsilon_{\mathrm{cls}}(x) \leq \|p(x) - \eta(x)\|_1 \leq \sqrt{2\,\mathrm{KL}(\eta(x) \| p(x))}.$$

For each expert coordinate, Pinsker's inequality for Bernoulli distributions gives

$$|u_j(x) - \alpha_j(x)| \leq \sqrt{\frac{1}{2}\,\mathrm{KL}\Big(\mathrm{Bern}(\alpha_j(x)) \big\| \mathrm{Bern}(u_j(x))\Big)}.$$

Taking the maximum over $j$ and upper bounding the maximum by the Euclidean norm yields

$$\varepsilon_{\mathrm{exp}}(x) = \max_j |u_j(x) - \alpha_j(x)| \leq \sqrt{\frac{1}{2} \sum_{j=1}^{J} \mathrm{KL}\Big(\mathrm{Bern}(\alpha_j(x)) \big\| \mathrm{Bern}(u_j(x))\Big)}.$$

Combining these inequalities with (17) gives

$$\varepsilon_{\mathrm{cls}}(x) \le \sqrt{2\,\delta_{\mathrm{dec}}(p,u \mid x)}, \qquad \varepsilon_{\mathrm{exp}}(x) \le \sqrt{\frac{\delta_{\mathrm{dec}}(p,u \mid x)}{2\beta}}.$$

Substituting into (18),

$$\begin{aligned}\delta_{\perp}(p,u \mid x) &\le 2\max\{\varepsilon_{\mathrm{cls}}(x),\, \varepsilon_{\mathrm{exp}}(x)\} = \max\{2\varepsilon_{\mathrm{cls}}(x),\, 2\varepsilon_{\mathrm{exp}}(x)\} \\ &\le \max\Big\{2\sqrt{2\,\delta_{\mathrm{dec}}(p,u \mid x)},\; 2\sqrt{\tfrac{\delta_{\mathrm{dec}}(p,u|x)}{2\beta}}\Big\} = \max\Big\{2\sqrt{2},\; \sqrt{\frac{2}{\beta}}\Big\}\sqrt{\delta_{\mathrm{dec}}(p,u \mid x)}.\end{aligned}$$

Defining $\Gamma(t) \coloneqq \max\{2\sqrt{2}, \sqrt{2/\beta}\}\sqrt{t}$, this is the pointwise calibration inequality

$$\delta_{\perp}(p,u \mid x) \le \Gamma\big(\delta_{\mathrm{dec}}(p,u \mid x)\big). \tag{19}$$

Taking expectations over $X$ and using Jensen's inequality, with the concavity of $\Gamma$, gives

$$\mathcal{E}_{\perp}(f_{p,u}) - \mathcal{E}_{\perp}^{\star} = \mathbb{E}_X[\delta_{\perp}(p,u \mid X)] \le \mathbb{E}_X[\Gamma(\delta_{\mathrm{dec}}(p,u \mid X))] \le \Gamma(\mathbb{E}_X[\delta_{\mathrm{dec}}(p,u \mid X)])\,.$$

Since

$$\mathbb{E}_X[\delta_{\mathrm{dec}}(p,u \mid X)] = \mathcal{E}_{\mathrm{dec}}(p,u) - \mathcal{E}_{\mathrm{dec}}^{\star},$$

the claimed bound follows from the definition of $\Gamma$ and $\beta = \lambda/J$. ■

This theorem contains the whole statistical story of the decoupled surrogate. The pointwise KL decomposition gives exact conditional alignment, the calibration inequality turns this alignment into regret control, and Jensen's inequality lifts the pointwise statement to population excess risk.

### C.6. Bayes-Consistency Corollary

**Corollary 12** (Bayes consistency of the decoupled surrogate)**.** *Every exact population minimizer of the decoupled surrogate satisfies $p = \eta$ and $u_j = \alpha_j$ almost surely for all $j \in [J]$, and the induced plug-in rule is Bayes-optimal. More generally, if $\mathcal{E}_{\mathrm{dec}}(p_n, u_n) \to \mathcal{E}_{\mathrm{dec}}^{\star}$, then $\mathcal{E}_{\perp}(f_{p_n,u_n}) \to \mathcal{E}_{\perp}^{\star}$.*

**Proof** The unrestricted class contains $(\eta, \alpha)$. Any exact minimizer $(p, u)$ achieves the population infimum, hence its conditional excess $\delta_{\mathrm{dec}}(p,u \mid x) = 0$ almost surely. By the KL decomposition (17), this forces $p = \eta$ and $u_j = \alpha_j$ a.s. for all $j$, so the plug-in rule is Bayes-optimal by Lemma 2. The approximate statement follows from Theorem 10. ■

## D. Additive Cross-Entropy (Mozannar & Sontag, 2020)

This section collects everything needed to understand the additive cross-entropy baseline in one place. The key point is that additive CE is misaligned on both axes simultaneously: it changes the population target and it rescales the local geometry by the realized number of correct experts.

The section proceeds in the same order as the main paper's two-axis analysis. We first restate the surrogate and show exactly what it estimates at the conditional optimum. We then derive its samplewise gradient and Hessian and make the multiplicity factor $1 + |\mathcal{J}|$ explicit. The closing comparison with the decoupled surrogate and the worked example then summarize why both effects matter.

### D.1. Formulation

The standard multi-expert cross-entropy baseline treats the $K$ classifier actions and the $J$ defer actions as one augmented action space and rewards the true class together with every correct expert.

**Definition 13** (Additive augmented cross-entropy, Mozannar & Sontag (2020))**.** Let $a(x) \in \mathbb{R}^{K+J}$ be an augmented score vector and define

$$q_i(x) \coloneqq \frac{\exp(a_i(x))}{\sum_{r=1}^{K+J}\exp(a_r(x))}, \qquad i \in [K+J].$$

$$\Phi^{\mathrm{CE}}(a; x, y, m) := -\log q_y(x) - \sum_{j:\, m_j = y} \log q_{K+j}(x).$$

Prediction is made by the augmented argmax of $a(x)$. This is the baseline whose additive reward structure leads to the redundancy-driven underfitting effect highlighted in the main text.

### D.2. Conditional Target Distortion

**Proposition 14** (Conditional minimizer of additive CE)**.** *The conditional risk of* CE *at $x$ is uniquely minimized at*

$$q_y^\star(x) = \frac{\eta_y(x)}{1 + U(x)}, \qquad q_{K+j}^\star(x) = \frac{\alpha_j(x)}{1 + U(x)}, \qquad y \in [K],\ j \in [J]. \tag{20}$$

**Proof** We begin by writing the conditional risk as a cross-entropy against a normalized target. Once this normalization is explicit, the optimizer follows from the standard uniqueness of the cross-entropy minimizer over a simplex.

The conditional surrogate risk at a fixed $x$ is

$$\mathcal{C}_{\mathrm{add}}(q \mid x) = -\sum_{y=1}^{K} \eta_y(x) \log q_y - \sum_{j=1}^{J} \alpha_j(x) \log q_{K+j},$$

subject to $q \in \Delta^{K+J}$. Since $\sum_y \eta_y(x) = 1$, the total target mass is $1 + U(x)$ where $U(x) = \sum_j \alpha_j(x)$. Define the normalized target

$$\tilde{t}_y(x) = \frac{\eta_y(x)}{1 + U(x)}, \qquad \tilde{t}_{K+j}(x) = \frac{\alpha_j(x)}{1 + U(x)}.$$

Then $\mathcal{C}_{\mathrm{add}}(q \mid x) = (1 + U(x)) H(\tilde{t}(x), q)$, where $H$ denotes cross-entropy. Cross-entropy over a simplex is uniquely minimized at $q = \tilde{t}$, which gives the stated formula. ■

### D.3. Gradient Amplification and Curvature

**Proposition 4** (Samplewise amplification in additive CE)**.** *For one sample $(x, y, m)$ with $\mathcal{J} = \mathcal{J}(y, m)$, the gradient and Hessian of $\Phi^{\mathrm{CE}}$ satisfy*

$$\begin{aligned} \frac{\partial \Phi^{\mathrm{CE}}}{\partial a_i} &= \big(1 + |\mathcal{J}|\big)\, q_i(x) - \mathbf{1}\{i = y\} \\ &\quad - \mathbf{1}\{i = K{+}j,\ j \in \mathcal{J}\}. \end{aligned} \tag{4}$$

$$\nabla_a^2 \Phi^{\mathrm{CE}} \;=\; (1 + |\mathcal{J}|)\big(\mathrm{Diag}(q) - qq^\top\big), \tag{5}$$

*with largest eigenvalue $\lambda_{\max}(\nabla_a^2 \Phi^{\mathrm{CE}}) \le (1 + |\mathcal{J}|)/2$.*

**Proof** The loss is a scaled log-sum-exp plus linear terms. We therefore derive the gradient first and then the Hessian, keeping track of where the factor $1 + |\mathcal{J}|$ enters.

Write $\mathcal{J} = \mathcal{J}(y, m)$. The loss can be rewritten as

$$\Phi^{\mathrm{CE}}(a; x, y, m) = (1 + |\mathcal{J}|) \log \sum_{r=1}^{K+J} e^{a_r} - a_y - \sum_{j \in \mathcal{J}} a_{K+j}.$$

**Part (a): gradient.** Since $\partial \log \sum_r e^{a_r} / \partial a_i = q_i$, differentiating gives

$$\frac{\partial \Phi^{\mathrm{CE}}}{\partial a_i} = (1 + |\mathcal{J}|)\, q_i - \mathbf{1}\{i = y\} - \mathbf{1}\{i = K + j,\ j \in \mathcal{J}\}.$$

The total positive mass is the sum of the indicator coefficients: $1 + |\mathcal{J}|$.

**Part (b): curvature.** We now differentiate once more. Recall that

$$q_i = \frac{e^{a_i}}{\sum_{r=1}^{K+J} e^{a_r}}.$$

Differentiating $q_i$ with respect to $a_r$ gives the standard softmax Jacobian

$$\frac{\partial q_i}{\partial a_r} = q_i(\mathbf{1}\{i = r\} - q_r).$$

Equivalently,

$$\frac{\partial q_i}{\partial a_i} = q_i(1 - q_i), \qquad \frac{\partial q_i}{\partial a_r} = -q_i q_r \quad (r \neq i).$$

Therefore the Hessian of the log-sum-exp term $\log \sum_r e^{a_r}$ is the matrix whose $(i, r)$ entry is $q_i(\mathbf{1}\{i = r\} - q_r)$, namely

$$\nabla_a^2 \log \sum_{r=1}^{K+J} e^{a_r} = \mathrm{Diag}(q) - qq^\top.$$

This matrix is the covariance matrix of a categorical random variable with probabilities $q$, which is why it is positive semidefinite. Since the terms $-a_y$ and $-\sum_{j\in\mathcal{J}} a_{K+j}$ are linear in $a$, they disappear under second differentiation. Hence

$$\nabla_a^2 \mathrm{CE} = (1 + |\mathcal{J}|)(\mathrm{Diag}(q) - qq^\top).$$

It remains to bound the largest eigenvalue. Let $M \coloneqq \mathrm{Diag}(q) - qq^\top$. We apply Gershgorin's circle theorem row by row.

For row $i$, the diagonal entry is

$$M_{ii} = q_i - q_i^2 = q_i(1 - q_i),$$

and every off-diagonal entry is

$$M_{ij} = -q_i q_j, \qquad j \neq i.$$

The Gershgorin radius of row $i$ is the sum of the absolute values of the off-diagonal entries:

$$R_i = \sum_{j\neq i} |M_{ij}| = \sum_{j\neq i} q_i q_j = q_i \sum_{j\neq i} q_j = q_i(1 - q_i),$$

because $\sum_j q_j = 1$. Gershgorin's theorem therefore says that every eigenvalue $\lambda$ of $M$ lies in at least one interval

$$[M_{ii} - R_i,\ M_{ii} + R_i] = [q_i(1 - q_i) - q_i(1 - q_i),\ q_i(1 - q_i) + q_i(1 - q_i)] = [0,\ 2q_i(1 - q_i)].$$

Since the scalar function $u \mapsto u(1 - u)$ is maximized at $u = 1/2$, we have $q_i(1 - q_i) \leq 1/4$ for every $i$. Hence every eigenvalue of $M$ lies in $[0, 1/2]$, and in particular

$$\lambda_{\max}(M) \leq \frac{1}{2}.$$

Finally, multiplying by the prefactor $1 + |\mathcal{J}|$ yields

$$\lambda_{\max}(\nabla_a^2 \mathrm{CE}) \leq \frac{1 + |\mathcal{J}|}{2}.$$

■

### D.4. Head-to-Head with the Decoupled Surrogate

Additive CE and the decoupled surrogate differ on both analytical axes, and the contrast is useful because the two losses are trying to solve the same decision problem in very different ways. Additive CE folds classes and experts into one augmented softmax. The decoupled surrogate instead keeps the class posterior and the expert utilities as separate objects throughout training and prediction. The next two paragraphs make precise why this distinction matters.

**Axis (i): statistical target.** Additive CE does not recover $(\eta, \alpha)$ directly. By Proposition 14, it learns the normalized object

$$\left(\frac{\eta_1}{1+U}, \ldots, \frac{\eta_K}{1+U}, \frac{\alpha_1}{1+U}, \ldots, \frac{\alpha_J}{1+U}\right),$$

so the target itself moves with the total expert overlap $U(x)$. This is the first misalignment. If one duplicates an already-correct expert several times, the underlying Bayes decision problem does not change, but the additive-CE target does: every class and expert coordinate is shrunk by the larger denominator $1 + U(x)$. In other words, additive CE is not estimating the native decision quantities on their original scale.

The decoupled surrogate has no such dependence. Its conditional minimizer is exactly $p^\star(x) = \eta(x)$ and $u_j^\star(x) = \alpha_j(x)$ for every expert (Appendix C.5). Duplicating one expert leaves the target for every other coordinate unchanged. This is the sense in which the decoupled surrogate is target-aligned: the statistical object being estimated is the Bayes object itself, not a renormalized proxy.

**Axis (ii): local geometry.** Additive CE allocates one positive unit of gradient mass to the true class and one additional unit to every correct expert, which yields the multiplicative factor $1 + |\mathcal{J}|$ in both the gradient and the Hessian. The practical meaning is simple: samples on which many experts happen to be correct are up-weighted in both first-order and second-order optimization, regardless of whether those samples are actually more informative.

The decoupled surrogate eliminates this samplewise multiplicity entirely. Its class gradient is the standard multiclass cross-entropy gradient, and each expert receives an independent Bernoulli gradient of size $(\lambda/J)(u_j - t_j)$, unaffected by how many other experts are correct. Thus the decoupled surrogate keeps the optimization problem on the same scale across low-overlap and high-overlap regions of the input space.

Additive CE is problematic for two separate reasons: it aims at the wrong population target, and it also distorts the optimization landscape around each sample. The decoupled surrogate fixes both issues at once by preserving the native target $(\eta, \alpha)$ and by using coordinatewise gradients with no overlap-dependent amplification.

### D.5. Worked Example: Redundant Experts

Consider $K = 2$ classes and $J = 4$ redundant experts. Suppose the Bayes-optimal expert behavior is represented by expert 1 and experts 2,3,4 are exact copies of it. On a sample with true label $y = 1$, assume all four experts are correct, so $|\mathcal{J}| = 4$.

This is the cleanest possible redundancy scenario. From the perspective of the decision problem, there is really only one useful expert behavior here, copied four times. A reasonable surrogate should therefore treat this sample almost the same way as it would treat a sample with only one correct expert. The calculation below shows that additive CE does not.

It is useful to make the axis-(i) issue explicit first. Suppose the class posterior is

$$\eta(x) = (0.6, 0.4),$$

and the underlying useful expert behavior has success probability $0.9$. If there is only one such expert, additive CE targets

$$\left(\frac{0.6}{1+0.9}, \frac{0.4}{1+0.9}, \frac{0.9}{1+0.9}\right) = (0.316,\ 0.211,\ 0.474).$$

If we now replace that one expert by four exact copies, the Bayes decision does not change: the best class utility is still $0.6$ and the best expert utility is still $0.9$. However, additive CE now targets

$$\left(\frac{0.6}{1+4\cdot 0.9}, \frac{0.4}{1+4\cdot 0.9}, \frac{0.9}{1+4\cdot 0.9}, \frac{0.9}{1+4\cdot 0.9}, \frac{0.9}{1+4\cdot 0.9}, \frac{0.9}{1+4\cdot 0.9}\right)$$
$$= (0.130,\ 0.087,\ 0.196,\ 0.196,\ 0.196,\ 0.196).$$

So even before discussing gradients, the surrogate target has already moved. Nothing about the underlying decision problem changed; only the number of redundant copies changed.

Take an uninformative additive-CE initialization with equal logits on the $K + J = 6$ actions, hence $q_i = 1/6$ for all $i$. Proposition 4 gives

$$\frac{\partial \Phi^{\mathrm{CE}}}{\partial a_1} = 5 \cdot \frac{1}{6} - 1 = -\frac{1}{6}, \qquad \frac{\partial \Phi^{\mathrm{CE}}}{\partial a_2} = 5 \cdot \frac{1}{6} = \frac{5}{6},$$

and for every correct expert $j \in \{1, 2, 3, 4\}$,

$$\frac{\partial \Phi^{\mathrm{CE}}}{\partial a_{K+j}} = 5 \cdot \frac{1}{6} - 1 = -\frac{1}{6}.$$

Thus the sample contributes one rewarded class term and four rewarded expert terms, for a total positive mass of $5$. By the Hessian formula proved above, the same sample also carries curvature prefactor $5$. Both quantities are inflated by the redundancy count.

To see why this is a pathology rather than a harmless rescaling, compare it with the single-copy version of the same problem. If only expert $1$ were present, then the same derivation would give total positive mass $2$ rather than $5$, because the sample would reward one class and one expert. So merely replacing one expert by four exact copies multiplies the local scale of the update by a factor of $5/2$, even though no new information has been added.

Now compare the decoupled surrogate on the same sample. Let the class head start at $p = (1/2, 1/2)$ and each expert head at $u_j = 1/2$. Then

$$\frac{\partial \Phi^{\mathrm{dec}}}{\partial w_1} = -\frac{1}{2}, \qquad \frac{\partial \Phi^{\mathrm{dec}}}{\partial w_2} = \frac{1}{2}, \qquad \frac{\partial \Phi^{\mathrm{dec}}}{\partial s_j} = \frac{\lambda}{4}\left(\frac{1}{2} - 1\right) = -\frac{\lambda}{8},$$

for all $j = 1, \ldots, 4$. If the same sample had only one correct expert instead of four, the decoupled surrogate class gradient would be unchanged and the nonzero expert gradient would still be of the same per-coordinate form. In other words, the decoupled surrogate updates each relevant coordinate because it is correct, not because many other coordinates happen to agree with it.

Under additive CE, duplicating one correct expert into four copies changes both what the model is trying to estimate and how strongly this sample pulls on the optimizer. Under the decoupled surrogate, duplication does neither: the class task remains the same, and each expert is trained on its own correctness event with the same local scale. That is why the decoupled surrogate remains stable in redundancy-heavy regimes while additive CE can underfit or become dominated by high-overlap regions.

The problem is not merely that its probability estimates can require post-processing. The deeper issue is structural. Redundant experts change the surrogate target and simultaneously change the optimization geometry, even when the Bayes decision problem itself is unchanged. The decoupled surrogate is designed precisely to remove both dependencies.

## E. PiCCE (Liu et al., 2026)

PiCCE (Liu et al., 2026) was designed to remove the additive amplification of CE (Mozannar & Sontag, 2020) by rewarding only one correct expert per sample. This section shows the resulting tradeoff. The samplewise scale becomes controlled, but the winner-take-all mechanism creates a different pathology: correct experts that fail to win the internal $\arg\max$ receive negative updates.

The logic of the section is parallel to that of additive CE. We first define the surrogate precisely, then isolate the starvation effect at the sample level, then show how the same winner-take-all mechanism induces a moving conditional target, and finally compare the resulting behavior with the decoupled surrogate. Two facts should be kept distinct throughout. First, PiCCE genuinely improves on additive CE in one respect: it removes the overlap-dependent amplification factor. Second, that repair is achieved by forcing correct experts to compete with one another. The point of the section is to show that this second effect is not a minor side effect but the central limitation of the surrogate.

### E.1. Formulation

PiCCE modifies the additive surrogate by rewarding only one correct expert rather than all of them.

**Definition 15** (PiCCE, Liu et al. (2026))**.** Let $a(x) \in \mathbb{R}^{K+J}$ be an augmented score vector and define

$$q_i(x) := \frac{\exp(a_i(x))}{\sum_{r=1}^{K+J} \exp(a_r(x))}, \qquad i \in [K+J].$$

Let

$$j^\star(a; x, y, m) = \arg\max_{j:\, m_j = y} a_{K+j}(x)$$

with deterministic tie-breaking, whenever at least one expert is correct. Then

$$\Phi^{\mathrm{PiCCE}}(a; x, y, m) := -\log q_y(x) - \mathbf{1}\{\exists j \in [J] : m_j = y\} \log q_{K+j^\star(a;x,y,m)}(x).$$

Prediction again uses the augmented argmax over $a(x)$. In our analysis, this winner-take-all correction removes additive amplification but can starve useful experts that are correct yet repeatedly unselected.

### E.2. Winner-Take-All Starvation

**Proposition 5** (Winner-take-all starvation in PiCCE)**.** *Under PiCCE with $\mathcal{J}(y, m) \neq \varnothing$, the gradient on any logit is $\partial\Phi^{\mathrm{PiCCE}}/\partial a_i = 2q_i - \mathbf{1}\{i = y\} - \mathbf{1}\{i = K+j^\star\}$. For any correct expert $k \in \mathcal{J}$ with $k \neq j^\star$,*

$$\frac{\partial \Phi^{\mathrm{PiCCE}}}{\partial a_{K+k}} \;=\; 2q_{K+k}(x) \;>\; 0. \tag{7}$$

*Gradient descent pushes expert $k$'s logit* downward*, even though it is correct.*

**Proof** Once the winning expert is fixed, PiCCE becomes a softmax loss with exactly two rewarded coordinates: the true class and the selected expert. The sign pattern for every other correct expert can therefore be read off directly from the gradient.

Assume $\mathcal{J}(y, m) \neq \varnothing$ and let $j^\star$ be the PiCCE winner. Then

$$\Phi^{\mathrm{PiCCE}}(a; x, y, m) = 2 \log \sum_{r=1}^{K+J} e^{a_r} - a_y - a_{K+j^\star}.$$

Differentiating with respect to $a_i$ gives $\partial\Phi^{\mathrm{PiCCE}}/\partial a_i = 2q_i - \mathbf{1}\{i = y\} - \mathbf{1}\{i = K + j^\star\}$. This formula already shows the winner-take-all structure: only two coordinates receive a subtractive reward term, namely the true class and the selected expert. Every other coordinate sees only the positive softmax term $2q_i$.

For a correct expert $k \neq j^\star$, set $i = K + k$. Since $k \neq j^\star$, neither indicator is active, so

$$\frac{\partial \Phi^{\mathrm{PiCCE}}}{\partial a_{K+k}} = 2q_{K+k}(x) > 0,$$

where the strict positivity uses $q_{K+k} > 0$ (softmax outputs are strictly positive). Hence gradient descent pushes $a_{K+k}$ downward.

The expert $k$ is correct on the current sample, but because it failed to win the internal competition, the update acts as if it were an ordinary non-target coordinate and suppresses its logit. ■

### E.3. Score-Dependent Conditional Target

For completeness we derive the conditional risk formula used in §4.2. Condition on $X = x$ and take expectations over $(Y, M)$. The PiCCE loss rewards expert $j^\star = \arg\max_{j \in \mathcal{J}(Y,M)} a_{K+j}$ whenever $\mathcal{J} \neq \varnothing$. Define $\gamma_j(a, x) = \Pr(j \in \mathcal{J}(Y, M)$ and $j = j^\star \mid X = x)$. The class term is standard:

$$\mathbb{E}[\log q_Y \mid X = x] = \sum_{k=1}^{K} \eta_k(x) \log q_k.$$

For the expert term, partition the event $\{\mathcal{J} \neq \varnothing\}$ according to which expert wins the internal competition. This yields

$$\mathbb{E}[\mathbf{1}\{\mathcal{J} \neq \varnothing\} \log q_{K+j^\star} \mid X = x] = \sum_{j=1}^{J} \gamma_j(a, x) \log q_{K+j}.$$

Substituting both identities gives

$$\mathcal{C}_{\mathrm{PiCCE}}(q, a \mid x) = -\sum_{k=1}^{K} \eta_k(x) \log q_k - \sum_{j=1}^{J} \gamma_j(a, x) \log q_{K+j}.$$

Thus, for a fixed score vector $a$, PiCCE behaves like an augmented cross-entropy whose class-side coefficients are $\eta_k(x)$ and whose expert-side coefficients are $\gamma_j(a, x)$, not $\alpha_j(x)$. Equivalently, viewed as a weighted multiclass log-loss over the augmented simplex, its instantaneous target is proportional to

$$\big(\eta_1(x), \ldots, \eta_K(x), \gamma_1(a, x), \ldots, \gamma_J(a, x)\big).$$

The coefficients $\gamma_j(a, x)$ depend on the score vector $a$ through the winner map $j^\star$, so the expert-side target moves as $a$ is updated. This is the precise sense in which PiCCE chases a moving target. The class-side weights are fixed by the data distribution, but the expert-side weights are redistributed by the current score ordering. This is in contrast to additive CE, where the coefficients are fixed by the distribution alone.

Fix a point $x$ and suppose there is positive conditional probability that experts 1 and 2 are both correct. If the current scores satisfy $a_{K+1} > a_{K+2}$ on those outcomes, then that shared probability mass contributes to $\gamma_1(a, x)$. If the score ordering is reversed, the same data mass contributes to $\gamma_2(a, x)$ instead. The data distribution has not changed; only the current score ordering has. That is why the PiCCE expert target is score-dependent rather than intrinsic to the underlying decision problem.

### E.4. Curvature

The winner-take-all repair eliminates the multiplicity factor of additive CE. Whenever $\mathcal{J}(y, m) \neq \varnothing$,

$$\Phi^{\mathrm{PiCCE}}(a; x, y, m) = 2 \log \sum_{r=1}^{K+J} e^{a_r} - a_y - a_{K+j^\star},$$

where $j^\star$ is treated as fixed on the current sample. We now differentiate this expression exactly as in the additive-CE case, but with prefactor 2 instead of $1 + |\mathcal{J}|$.

The only nonlinear term is the log-sum-exp. Its Hessian is

$$\nabla_a^2 \log \sum_{r=1}^{K+J} e^{a_r} = \mathrm{Diag}(q) - qq^\top,$$

while the linear terms $-a_y$ and $-a_{K+j^\star}$ vanish under second differentiation. Therefore

$$\nabla_a^2 \Phi^{\mathrm{PiCCE}} = 2(\mathrm{Diag}(q) - qq^\top),$$

The matrix $\mathrm{Diag}(q) - qq^\top$ is again the softmax covariance matrix. From the additive-CE derivation above, every eigenvalue of this matrix lies in $[0, 1/2]$. Multiplying by the prefactor 2 gives

$$\lambda_{\max}\big(\nabla_a^2 \Phi^{\mathrm{PiCCE}}\big) \leq 1.$$

So PiCCE is indeed well behaved in the second-order sense that additive CE is not. This point is worth stating clearly because it isolates the real failure mode. The problem with PiCCE is not exploding curvature or overlap-dependent scaling. The problem is the direction of the updates: the surrogate suppresses correct experts that fail to win the internal comparison, and this same winner-take-all mechanism makes the expert-side target score-dependent.

### E.5. Head-to-Head with the Decoupled Surrogate

**Axis (i): statistical target.** PiCCE preserves the correct class coefficients $\eta_k(x)$, but on the expert side it replaces $\alpha_j(x)$ by the score-dependent quantity $\gamma_j(a, x)$. The target therefore depends on the optimization trajectory through the current winner map. This is the first key contrast with the decoupled surrogate. Under the decoupled surrogate, the expert target for coordinate $j$ is always the same object, $\alpha_j(x)$, regardless of which expert currently has the highest score. The data determine the target. Under PiCCE, by contrast, shared correct mass is assigned to whichever expert is currently winning the internal competition.

The practical implication is that PiCCE does not merely estimate the wrong expert quantity at the optimum; it can also reshape the quantity it is trying to estimate while training proceeds. The decoupled surrogate does not have this moving-target effect: the class and expert subproblems are proper losses with fixed conditional targets $(\eta, \alpha)$.

**Axis (ii): local geometry.** PiCCE solves the multiplicity problem by capping the number of rewarded terms at two, but it does so by selecting a unique correct expert and repelling every other correct expert. This is not just a bookkeeping choice. It means that on a sample where several experts are simultaneously correct, PiCCE turns agreement into competition.

The decoupled surrogate makes the opposite design choice. It keeps the per-coordinate updates bounded while reinforcing every correct expert independently. Agreement among correct experts therefore does not create either additive amplification or winner-take-all suppression. PiCCE and the decoupled surrogate agree on the need to avoid additive amplification, but they differ fundamentally on whether multiple correct experts should compete with one another or all receive supportive updates.

PiCCE improves on additive CE along one dimension: it removes the $O(|\mathcal{J}|)$ scaling. But it does so by replacing an overlap-dependent magnitude problem with a winner-dependent direction problem. The decoupled surrogate avoids both failures simultaneously: it keeps the local scale bounded and it preserves supportive updates for every correct expert.

### E.6. Worked Example: Generalist and Specialist Both Correct

Consider one generalist expert and one specialist expert. On a sample from the specialist region, suppose both experts are correct, so $\mathcal{J} = \{1, 2\}$, but the current logits satisfy $a_{K+1} > a_{K+2}$, meaning the generalist wins the internal PiCCE competition. This is precisely the situation in which one would like the training signal to preserve both experts: the generalist is correct, but the specialist is also correct on the very region for which it was meant to be useful. PiCCE does not preserve both. It forces them to compete.

To make the signs concrete, suppose the current softmax assigns

$$q_{K+1} = 0.30, \qquad q_{K+2} = 0.20.$$

Then $j^\star = 1$ and

$$\frac{\partial \Phi^{\mathrm{PiCCE}}}{\partial a_{K+1}} = 2q_{K+1} - 1, \qquad \frac{\partial \Phi^{\mathrm{PiCCE}}}{\partial a_{K+2}} = 2q_{K+2}.$$

Substituting the numbers gives

$$\frac{\partial \Phi^{\mathrm{PiCCE}}}{\partial a_{K+1}} = 2(0.30) - 1 = -0.40, \qquad \frac{\partial \Phi^{\mathrm{PiCCE}}}{\partial a_{K+2}} = 2(0.20) = 0.40.$$

Under gradient descent, the generalist logit is pushed upward because its gradient is negative, while the specialist logit is pushed downward because its gradient is positive. So the same sample provides evidence for both experts, yet PiCCE uses that evidence to strengthen the winner and weaken the loser.

This same example also reflects the axis-(i) issue. On any population of such samples, the shared "both correct" mass is assigned to whichever expert currently wins the internal comparison. If the generalist is ahead, that mass contributes to $\gamma_1(a, x)$; if the specialist overtakes it, the same mass contributes to $\gamma_2(a, x)$ instead. The target therefore co-moves with the current score ordering.

For the decoupled surrogate, let the same sample have expert probabilities $u_1 = 0.80$ and $u_2 = 0.75$, with targets $t_1 = t_2 = 1$. Then

$$\frac{\partial \Phi^{\mathrm{dec}}}{\partial s_1} = \frac{\lambda}{2}(0.80 - 1) = -0.10\lambda, \qquad \frac{\partial \Phi^{\mathrm{dec}}}{\partial s_2} = \frac{\lambda}{2}(0.75 - 1) = -0.125\lambda.$$

Both experts are reinforced. The specialist may still trail the generalist at the current iterate, but it is not punished for being correct. The sample acts as supportive evidence for both experts rather than as an internal tournament.

This example is therefore the PiCCE analogue of the redundant-expert example for additive CE. The key point is not merely that PiCCE uses a different formula. The deeper point is structural. Whenever several experts are correct on the same sample, PiCCE converts agreement into competition. The decoupled surrogate preserves the bounded geometry without introducing that competition, which is exactly why it is more robust in the rare-specialist regime studied in the main text.

## F. Mao25 (Mao et al., 2025d)

Mao25 moves away from additive rewards and winner-take-all selection by optimizing only the total mass assigned to the acceptable set. This resolves the two most immediate geometric issues, but it also removes the signal needed to rank experts within that set.

The central question in this section is what object Mao25 is actually trying to fit. The answer is not the full vector $(\eta, \alpha)$, but only the total probability assigned to the acceptable set. That choice makes the local geometry comparatively benign, but it also explains why the surrogate cannot express preferences among multiple acceptable experts.

The positive theory in Mao et al. (2025d) is important to keep in view. In the single-stage multiple-expert setting, their family $L_\Psi$ comes with formal consistency guarantees under mild endpoint conditions on $\Psi$. They further single out the member $\Psi(u) = 1 - u$, denoted $L_{\mathrm{mae}}$ in their paper, because it also admits quantitative excess-risk control. Our criticism is not about those consistency guarantees. It is about the practical optimization behavior of that specific $\Psi(u) = 1 - u$ instantiation: its samplewise loss is linear in acceptable-set mass and therefore has an MAE-like optimization profile.

### F.1. Formulation

The single-stage surrogate of Mao et al. (2025d) defines a family indexed by a decreasing function $\Psi$. In their multiple-expert experiments, they instantiate this family with $\Psi(u) = 1 - u$, which is the version we include in our empirical comparison.

**Definition 16** (Mao25 single-stage surrogate with $\Psi(u) = 1 - u$, Mao et al. (2025d))**.** Let $a(x) \in \mathbb{R}^{K+J}$ be an augmented score vector and define

$$q_i(x) := \frac{\exp(a_i(x))}{\sum_{r=1}^{K+J} \exp(a_r(x))}, \qquad i \in [K+J].$$

Then

$$\Phi^{\mathrm{Mao}}(a; x, y, m) := 1 - \Big( q_y(x) + \sum_{j:\, m_j = y} q_{K+j}(x) \Big).$$

Prediction uses the augmented argmax over $a(x)$. This surrogate depends only on the total mass assigned to the acceptable set consisting of the true class and all correct experts.

This has a simple consequence at the conditional level. Fix an input $x$ and take expectation over the random pair $(Y, M)$.

The class part contributes

$$\mathbb{E}[q_Y \mid X = x] = \sum_{k=1}^{K} \eta_k(x)\, q_k,$$

because $q_k$ is fixed once $x$ is fixed, and only the random label $Y$ remains inside the expectation. Likewise, for the expert part,

$$\mathbb{E}\left[ \sum_{j:\, M_j = Y} q_{K+j} \,\middle|\, X = x \right] = \sum_{j=1}^{J} \alpha_j(x)\, q_{K+j},$$

because expert $j$ contributes exactly on the event $\{M_j = Y\}$, whose conditional probability is $\alpha_j(x)$ by definition.

Substituting these two identities into the loss gives

$$\mathcal{C}_{\mathrm{Mao}}(q \mid x) = 1 - \sum_{k=1}^{K} \eta_k(x) q_k - \sum_{j=1}^{J} \alpha_j(x) q_{K+j}.$$

Now the key point is visible. The conditional risk is a linear function of the augmented probabilities $q$. Minimizing it is therefore the same as maximizing

$$\sum_{k=1}^{K} \eta_k(x) q_k + \sum_{j=1}^{J} \alpha_j(x) q_{K+j}$$

over the simplex $\Delta^{K+J}$. A linear objective over a simplex is minimized at an extreme point, so there always exists a conditional minimizer that puts all mass on one action whose coefficient is maximal among

$$\big(\eta_1(x), \ldots, \eta_K(x), \alpha_1(x), \ldots, \alpha_J(x)\big).$$

If the maximizer is unique, that minimizer is unique and is exactly the Bayes-optimal action. If several actions tie for the maximum, any convex combination over those maximizers is also optimal.

So Mao25 can recover a Bayes-optimal action, but it does not recover the full vector $(\eta, \alpha)$. It identifies which action can be optimal, not the underlying utilities themselves.

### F.2. Acceptable-Set Mass Without Ranking

**Proposition 6** (Acceptable-set mass in Mao25)**.** *The gradient is $\partial\Phi^{\mathrm{Mao}}/\partial a_i = q_i\big(S_{\mathcal{S}} - \mathbf{1}\{i \in \mathcal{S}(y,m)\}\big)$, where $\mathcal{S}(y,m) = \{y\} \cup \{K+j : j \in \mathcal{J}\}$ and $S_{\mathcal{S}} = \sum_{c\in\mathcal{S}} q_c$. The loss depends on the acceptable set only through $S_{\mathcal{S}}$, and for acceptable coordinates the gradient has the common factor $S_{\mathcal{S}} - 1$. Thus the samplewise update distinguishes acceptable from unacceptable actions, but it contains no term that prefers one acceptable expert over another on the basis of expert quality.*

**Proof** The argument has two steps. First we show that the loss depends on the score vector only through the acceptable-set mass. Then we differentiate that mass to obtain the gradient formula.

Let

$$\mathcal{S}(y,m) \coloneqq \{y\} \cup \{K+j : j \in \mathcal{J}(y,m)\}$$

denote the acceptable set. By definition,

$$\Phi^{\mathrm{Mao}}(a; x, y, m) = 1 - \sum_{c\in\mathcal{S}(y,m)} q_c(x) = 1 - S_{\mathcal{S}}(x).$$

Thus the loss depends on $a$ only through the total acceptable mass $S_{\mathcal{S}}(x)$.

To differentiate, use the softmax Jacobian $\partial q_c/\partial a_i = q_c(\mathbf{1}\{c=i\} - q_i)$. Then

$$\begin{aligned}\frac{\partial S_{\mathcal{S}}}{\partial a_i} &= \sum_{c\in\mathcal{S}(y,m)} q_c(\mathbf{1}\{c=i\} - q_i)\\ &= \mathbf{1}\{i\in\mathcal{S}(y,m)\}q_i - q_i \sum_{c\in\mathcal{S}(y,m)} q_c\\ &= q_i\big(\mathbf{1}\{i\in\mathcal{S}(y,m)\} - S_{\mathcal{S}}\big).\end{aligned}$$

and therefore

$$\frac{\partial\Phi^{\mathrm{Mao}}}{\partial a_i} = -\frac{\partial S_{\mathcal{S}}}{\partial a_i} = q_i\big(S_{\mathcal{S}} - \mathbf{1}\{i\in\mathcal{S}(y,m)\}\big).$$

If $i \in \mathcal{S}(y,m)$, then $S_{\mathcal{S}} \le 1$, so the gradient is non-positive. If $i \notin \mathcal{S}(y,m)$, then the gradient is $q_i S_{\mathcal{S}} \ge 0$. Finally, since $\Phi^{\mathrm{Mao}} = 1 - S_{\mathcal{S}}$, any two score vectors with the same $S_{\mathcal{S}}$ incur the same loss regardless of their internal allocation inside $\mathcal{S}(y,m)$.

This formula makes the absence of direct within-set ranking explicit. If $i, r \in \mathcal{S}(y,m)$ are both acceptable actions on the current sample, then

$$\frac{\partial\Phi^{\mathrm{Mao}}}{\partial a_i} = q_i(S_{\mathcal{S}} - 1), \qquad \frac{\partial\Phi^{\mathrm{Mao}}}{\partial a_r} = q_r(S_{\mathcal{S}} - 1).$$

Both coordinates share the same factor $S_{\mathcal{S}} - 1$. Their update magnitudes differ only through the current probabilities $q_i$ and $q_r$, not through any term encoding which acceptable expert is statistically better.

The same formula also reveals why optimization can be weak. Summing the attractive gradient mass over the acceptable set gives

$$\sum_{i\in\mathcal{S}} -\frac{\partial\Phi^{\mathrm{Mao}}}{\partial a_i} = \sum_{i\in\mathcal{S}} q_i(1 - S_{\mathcal{S}}) = S_{\mathcal{S}}(1 - S_{\mathcal{S}}).$$

This quantity is at most $1/4$ and tends to zero as $S_{\mathcal{S}} \to 0$. Thus, when the model currently places very little mass on the acceptable set, the corrective signal toward that set is weak. In the single-label special case $|\mathcal{S}| = 1$, the loss reduces to $1 - q_y$, which is exactly the multiclass MAE form. The same weak-gradient phenomenon is well known for MAE-like losses in noisy-label learning (Zhang & Sabuncu, 2018). ■

### F.3. Curvature

Although Mao25 does not suffer from additive amplification, its second-order structure still depends only on acceptable-set membership and the total acceptable mass. We differentiate the gradient formula step by step.

Write

$$g_i(a) \coloneqq \frac{\partial \Phi^{\mathrm{Mao}}}{\partial a_i} = q_i\big(S_{\mathcal{S}} - \mathbf{1}\{i \in \mathcal{S}\}\big).$$

Then

$$\frac{\partial g_i}{\partial a_r} = \frac{\partial q_i}{\partial a_r}\big(S_{\mathcal{S}} - \mathbf{1}\{i \in \mathcal{S}\}\big) + q_i \frac{\partial S_{\mathcal{S}}}{\partial a_r}.$$

The first term comes from differentiating the prefactor $q_i$ and the second term comes from differentiating the acceptable-set mass $S_{\mathcal{S}}$. From the softmax Jacobian,

$$\frac{\partial q_i}{\partial a_r} = q_i(\mathbf{1}\{i = r\} - q_r),$$

and from the previous subsection,

$$\frac{\partial S_{\mathcal{S}}}{\partial a_r} = q_r\big(\mathbf{1}\{r \in \mathcal{S}\} - S_{\mathcal{S}}\big).$$

Substituting both identities into the product rule gives the full Hessian entry:

$$\frac{\partial^2 \Phi^{\mathrm{Mao}}}{\partial a_i\, \partial a_r} = q_i(\mathbf{1}\{i = r\} - q_r)\big(S_{\mathcal{S}} - \mathbf{1}\{i \in \mathcal{S}\}\big) + q_i q_r\big(\mathbf{1}\{r \in \mathcal{S}\} - S_{\mathcal{S}}\big).$$

No additional term is hidden. This is the complete second derivative obtained by differentiating the explicit gradient formula.

It is useful to unpack the main cases.

**Case 1:** $i = r \in \mathcal{S}$**.** Then

$$\frac{\partial^2 \Phi^{\mathrm{Mao}}}{\partial a_i^2} = q_i(1 - q_i)(S_{\mathcal{S}} - 1) + q_i^2(1 - S_{\mathcal{S}}).$$

This depends on $q_i$ and on the common scalar $S_{\mathcal{S}}$, but not on which acceptable expert $i$ represents.

**Case 2:** $i \neq r$ **and** $i, r \in \mathcal{S}$**.** Then

$$\frac{\partial^2 \Phi^{\mathrm{Mao}}}{\partial a_i\, \partial a_r} = -q_i q_r(S_{\mathcal{S}} - 1) + q_i q_r(1 - S_{\mathcal{S}}) = 2q_i q_r(1 - S_{\mathcal{S}}).$$

Again the expression depends only on the current masses and on the shared acceptable-set mass.

**Case 3:** $i \in \mathcal{S}$ **and** $r \notin \mathcal{S}$**.** Then

$$\frac{\partial^2 \Phi^{\mathrm{Mao}}}{\partial a_i\, \partial a_r} = -q_i q_r(S_{\mathcal{S}} - 1) - q_i q_r S_{\mathcal{S}} = q_i q_r(1 - 2S_{\mathcal{S}}).$$

**Case 4:** $i, r \notin \mathcal{S}$**.** Then

$$\frac{\partial^2 \Phi^{\mathrm{Mao}}}{\partial a_i\, \partial a_r} = q_i(\mathbf{1}\{i = r\} - q_r)S_{\mathcal{S}} - q_i q_r S_{\mathcal{S}}.$$

In particular, for $i \neq r$ both outside the acceptable set,

$$\frac{\partial^2 \Phi^{\mathrm{Mao}}}{\partial a_i\, \partial a_r} = -2q_i q_r S_{\mathcal{S}}.$$

This expression contains no term that depends on which acceptable expert is statistically better. Even at second order, the surrogate sees acceptable-set mass and membership, not expert quality. This is the geometric analogue of the first-order statement above: Mao25 distinguishes acceptable from unacceptable actions, but its local geometry contains no utility-sensitive comparison inside the acceptable set.

### F.4. Head-to-Head with the Decoupled Surrogate

**Axis (i): statistical target.** Mao25 is aligned with the Bayes action only at the coarse level of choosing the best augmented action. Its conditional risk is linear in $q$, so a minimizer places all mass on one optimal action, but the surrogate does not recover the full utility vector $(\eta, \alpha)$. The decoupled surrogate does. At the conditional minimum, the decoupled surrogate returns $p^\star(x) = \eta(x)$ and $u_j^\star(x) = \alpha_j(x)$ for every expert. The difference is therefore not whether the two surrogates can identify a good action in principle. The difference is that the decoupled surrogate estimates the underlying utilities themselves, whereas Mao25 only optimizes a coarse action-level objective.

**Axis (ii): local geometry.** Mao25 avoids both additive amplification and winner-take-all starvation. In that sense it is geometrically cleaner than CE and PiCCE. But the geometry is still too coarse for utility estimation. On a sample where several actions are acceptable, the loss depends only on the total acceptable mass and the acceptable/unacceptable partition. It contains no term that says which acceptable expert should be preferred. For the concrete choice $\Psi(u) = 1 - u$, the total attractive gradient mass toward the acceptable set is $S_{\mathcal{S}}(1 - S_{\mathcal{S}})$. This quantity is maximal at $1/4$ and shrinks to zero as $S_{\mathcal{S}} \to 0$, so the corrective signal can be weak precisely when the model currently assigns little mass to the right set. The decoupled surrogate keeps the benign bounded geometry while training each expert head against its own correctness process.

The contrast with the decoupled surrogate is therefore different from the PiCCE contrast. Mao25 does not harm correct experts by suppressing them, and it does not amplify high-overlap samples. Its limitation is coarseness: it treats expert supervision as set membership rather than utility estimation, and under the linear choice $\Psi(u) = 1 - u$ this coarse signal can also be weak away from the optimum. The decoupled surrogate retains the stable geometry while preserving expert-specific targets.

### F.5. Worked Example: Two Correct Experts of Unequal Quality

Consider a defer region in which expert 1 is substantially better overall than expert 2:

$$\alpha_1(x) = 0.9, \qquad \alpha_2(x) = 0.5.$$

Assume the best class utility is smaller, say $\max_k \eta_k(x) = 0.4$, so the Bayes-optimal action is to defer to expert 1.

The conditional Mao25 risk at this $x$ is then

$$\mathcal{C}_{\mathrm{Mao}}(q \mid x) = 1 - \sum_{k=1}^{K} \eta_k(x) q_k - 0.9\, q_{K+1} - 0.5\, q_{K+2}.$$

This conditional objective is minimized by placing all mass on expert 1, which indeed recovers the correct action. But the surrogate does not recover the utility values $0.9$ and $0.5$ themselves; it only prefers the larger coefficient through a linear argmax objective.

Now inspect the samplewise supervision when both experts happen to be correct on a particular draw. Compare two augmented allocations with the same acceptable mass:

$$(q_y, q_{K+1}, q_{K+2}) = (0.10, 0.50, 0.00), \qquad (q_y, q_{K+1}, q_{K+2}) = (0.10, 0.25, 0.25).$$

Both have $S_{\mathcal{S}} = 0.60$, so both incur the same Mao25 loss $\Phi^{\mathrm{Mao}} = 0.40$. The surrogate therefore cannot prefer the first allocation over the second on the basis that expert 1 is the reliable one.

The gradients make the same point. Since both experts are acceptable on this sample, the gradient formula gives

$$\frac{\partial \Phi^{\mathrm{Mao}}}{\partial a_i} = q_i(0.60 - 1) = -0.4\, q_i, \qquad i \in \{y, K+1, K+2\}.$$

For the first allocation,

$$\left( \frac{\partial \Phi^{\mathrm{Mao}}}{\partial a_y}, \frac{\partial \Phi^{\mathrm{Mao}}}{\partial a_{K+1}}, \frac{\partial \Phi^{\mathrm{Mao}}}{\partial a_{K+2}} \right) = (-0.04,\ -0.20,\ 0.00),$$

whereas for the second allocation,

$$\left( \frac{\partial \Phi^{\mathrm{Mao}}}{\partial a_y}, \frac{\partial \Phi^{\mathrm{Mao}}}{\partial a_{K+1}}, \frac{\partial \Phi^{\mathrm{Mao}}}{\partial a_{K+2}} \right) = (-0.04,\ -0.10,\ -0.10).$$

The difference between the two acceptable experts is driven only by their current masses $q_{K+1}$ and $q_{K+2}$, not by the fact that expert 1 is the better expert overall.

The decoupled surrogate behaves differently. Its expert-side conditional risk is

$$-0.9\log u_1 - 0.1\log(1-u_1) - 0.5\log u_2 - 0.5\log(1-u_2),$$

which is uniquely minimized at $u_1 = 0.9$ and $u_2 = 0.5$. The ranking signal is there by construction. Mao25 can recover the correct defer action, but it does not estimate the underlying expert utilities. On the sample itself, it only rewards membership in the acceptable set, and the total reward scale is capped by $S_{\mathcal{S}}(1 - S_{\mathcal{S}})$. The decoupled surrogate does both: it preserves the correct action and recovers the utility gap that distinguishes expert 1 from expert 2.

## G. A-SM (Cao et al., 2024)

A-SM (Cao et al., 2024) is the strongest augmented-action baseline on the population axis. It restores bounded estimates and recovers the correct target. Its remaining weakness is optimization-level: the expert losses still leak into the class gradient through the shared score vector.

This section therefore plays a different role from the earlier baseline sections. For additive CE, PiCCE, and Mao25, the main mismatch is already visible in the conditional target. For A-SM, the target is largely the right one. The point here is to isolate the remaining axis-(ii) weakness cleanly and show that it survives even after boundedness and population alignment are repaired.

### G.1. Formulation

A-SM replaces the symmetric augmented normalization by an asymmetric one that keeps the class probabilities and expert probabilities bounded and semantically valid. We use the multi-expert extension given in Appendix H of Cao et al. (2024). To stay consistent with the main text, we write $\xi_k$ for the class estimates and $\psi_j$ for the expert estimates.

The single-expert A-SM paper is written with one defer coordinate, so its expert denominator is naturally expressed over $K+1$ coordinates: the $K$ class logits and the single defer logit. In the multi-expert extension, this construction is applied separately to each expert. Thus the class head still uses the first $K$ logits only, while expert $j$ is normalized against those same $K$ class logits plus its own expert logit $a_{K+j}$, not against the other expert logits. This point is easy to miss when moving from the single-expert notation to the multi-expert one, so we state it explicitly here.

**Definition 17** (Multi-expert A-SM, Cao et al. (2024))**.** Let $a(x) \in \mathbb{R}^{K+J}$ be an augmented score vector. Define the class estimates as the standard softmax over the first $K$ coordinates,

$$\xi_k(a(x)) := \frac{\exp(a_k(x))}{\sum_{k'=1}^{K}\exp(a_{k'}(x))}, \qquad k \in [K],$$

and the expert estimates, for each $j \in [J]$, as

$$\psi_j(a(x)) := \frac{\exp(a_{K+j}(x))}{\sum_{k'=1}^{K}\exp(a_{k'}(x)) + \exp(a_{K+j}(x)) - \max_{k'\in[K]}\exp(a_{k'}(x))}.$$

The corresponding surrogate is

$$\Phi^{\mathrm{ASM}}(a; x, y, m) := -\log \xi_y(a(x)) - \sum_{j=1}^{J}\Big(\mathbf{1}[m_j = y]\log\psi_j(a(x)) + \mathbf{1}[m_j \neq y]\log\big(1-\psi_j(a(x))\big)\Big). \tag{21}$$

Prediction uses the maxima-preserving property of $(\xi, \psi)$, equivalently the argmax of the underlying score vector $a(x)$. The expert estimate can be written through the equivalent sigmoid identity

$$\psi_j(a(x)) = \sigma\left(a_{K+j}(x) - \log\sum_{k'\neq k^\star}\exp(a_{k'}(x))\right),$$

where $k^\star \in \arg\max_{k\in[K]} a_k(x)$. To see the equivalence directly, write

$$B(a) := \sum_{k=1}^{K} \exp(a_k) - \max_{k\in[K]} \exp(a_k) = \sum_{k\neq k^\star} \exp(a_k).$$

Then the denominator of $\psi_j$ is exactly $B(a) + \exp(a_{K+j})$, so

$$\psi_j(a(x)) = \frac{\exp(a_{K+j}(x))}{\exp(a_{K+j}(x)) + B(a(x))} = \sigma(a_{K+j}(x) - \log B(a(x)))\,.$$

This form is algebraically equivalent to the original max-based denominator and is the convenient form for the derivative calculations below. It also makes clear that the map is only piecewise smooth: when the maximizing class changes, the active comparison set changes as well.

This representation is also conceptually useful. It shows that each expert head is, by itself, an ordinary sigmoid probability whose reference level is set by the non-max class logits. The expert-side issue in A-SM is therefore not that the expert probability is ill-defined or unbounded. The issue is that the class logits appear inside that reference level, so class and expert optimization are still entangled.

### G.2. Conditional Target

At the population level, A-SM is intended to repair exactly the defect that motivated Cao et al.: the single augmented softmax of additive CE does not directly produce bounded estimates of the class posterior and expert correctness. A-SM changes the parameterization so that the class side remains a categorical distribution over the first $K$ coordinates, while each expert side is a bounded probability in $[0, 1]$.

The multi-expert extension in Appendix H of Cao et al. (2024) states the corresponding optimality property directly: at a conditional optimum,

$$\xi_k^\star(x) = \eta_k(x), \qquad \psi_j^\star(x) = \alpha_j(x),$$

for every class $k \in [K]$ and expert $j \in [J]$. This is the precise reason A-SM is the strongest augmented-action baseline on axis (i). Unlike additive CE, PiCCE, and Mao25, its main limitation is not that it identifies the wrong statistical object. It identifies the right object but reaches it through a shared geometry that still couples class and expert optimization.

That distinction matters for the rest of the paper. Once the target axis is repaired, any remaining weakness must come from the optimization geometry rather than from statistical misspecification. A-SM is therefore the cleanest test case for asking whether "correct target" is sufficient by itself. Our answer is no: even with the right conditional optimum, the local update structure can remain coupled in a way that the decoupled surrogate avoids.

### G.3. Class–Expert Gradient Coupling

**Proposition 7** (Class–expert gradient coupling in A-SM)**.** *Fix $k^\star \in \arg\max_{k\in[K]} a_k$. For any $r \neq k^\star$, let $\pi_r = \exp(a_r)/\sum_{k\neq k^\star}\exp(a_k)$. Then*

$$\frac{\partial \Phi^{\mathrm{ASM}}}{\partial a_r} = \underbrace{(\xi_r - \mathbf{1}\{r{=}y\})}_{\textit{class term}} - \pi_r \underbrace{\sum_{j=1}^{J}\big(\psi_j - \mathbf{1}\{m_j{=}y\}\big)}_{\textit{expert leakage}}\,. \tag{8}$$

**Proof** We work on a region where the maximizing class index $k^\star$ is fixed. On such a region, the A-SM map is smooth and the active comparison set in the expert denominators does not change. The goal of this proof is only the *first-order* coupling statement: we derive the ordinary class term, then show how each expert residual contributes an additional term to every non-max class logit.

**Class term.** The term $-\log \xi_y$ is the standard multiclass cross-entropy over $a_1, \ldots, a_K$. Its gradient with respect to $a_r$ is $\xi_r - \mathbf{1}\{r = y\}$.

**Expert term.** For a fixed expert $j$, define

$$t_j \coloneqq \mathbf{1}\{m_j = y\}, \qquad B(a) \coloneqq \sum_{k \neq k^\star} \exp(a_k).$$

Then

$$\psi_j = \sigma\big(a_{K+j} - \log B(a)\big) = \frac{\exp(a_{K+j})}{\exp(a_{K+j}) + B(a)}.$$

The corresponding expert contribution to the loss is the Bernoulli cross-entropy

$$L_j = -t_j \log \psi_j - (1 - t_j) \log(1 - \psi_j).$$

Differentiating with respect to the sigmoid argument gives the standard identity

$$\frac{\partial L_j}{\partial v_j} = \psi_j - t_j, \qquad v_j \coloneqq a_{K+j} - \log B(a).$$

This already shows that the expert head itself is well behaved: with respect to its own effective score $v_j$, it is just an ordinary Bernoulli logistic loss. So the only remaining task is to differentiate $v_j$ with respect to the class logits and see how the coupling arises.

If $r \neq k^\star$, then

$$\frac{\partial v_j}{\partial a_r} = -\frac{1}{B(a)} \frac{\partial B(a)}{\partial a_r} = -\frac{\exp(a_r)}{\sum_{k \neq k^\star} \exp(a_k)} = -\pi_r,$$

where

$$\pi_r \coloneqq \frac{\exp(a_r)}{\sum_{k \neq k^\star} \exp(a_k)}.$$

If instead $r = k^\star$, then $a_{k^\star}$ does not appear in $B(a)$ on this region, so $\partial v_j / \partial a_{k^\star} = 0$.

Combining the two derivatives gives

$$\frac{\partial L_j}{\partial a_r} = \frac{\partial L_j}{\partial v_j} \frac{\partial v_j}{\partial a_r} = -(\psi_j - t_j)\pi_r, \qquad r \neq k^\star,$$

and

$$\frac{\partial L_j}{\partial a_{k^\star}} = 0.$$

This is the leakage mechanism in its simplest form: every expert residual $\psi_j - t_j$ is injected into every non-max class logit through the same weight $\pi_r$.

It is also useful to record the expert-side own derivative. Since $\partial v_j / \partial a_{K+j} = 1$,

$$\frac{\partial L_j}{\partial a_{K+j}} = \frac{\partial L_j}{\partial v_j} = \psi_j - t_j.$$

So the expert block itself behaves exactly like a standard Bernoulli cross-entropy. The pathology is not on the expert-side update viewed in isolation. The pathology is that the same expert residuals also enter the class gradient.

Summing over $j$ for $r \neq k^\star$ gives

$$\sum_{j=1}^{J} \frac{\partial L_j}{\partial a_r} = -\pi_r \sum_{j=1}^{J} (\psi_j - \mathbf{1}\{m_j = y\}).$$

Adding the class term yields the stated formula.

The sign pattern is now transparent. If many experts are correct but currently underestimated, then many terms $\psi_j - \mathbf{1}\{m_j = y\}$ are negative, and their sum adds an extra negative contribution to each non-max class gradient. If instead many experts are currently overestimated, the same mechanism pushes in the opposite direction. Either way, the class update no longer reflects only class-side error. It is contaminated by the aggregate expert residual. ■

### G.4. Curvature

For A-SM the most informative second-order object is the mixed class–expert block. It captures exactly the feature that distinguishes A-SM from the decoupled surrogate: even though the target is correct, the local quadratic geometry does not separate the class and expert subproblems.

We again work on a region where the maximizing class index $k^\star$ is fixed. From Proposition 7, the gradient on a non-max class logit $a_r$ is

$$\frac{\partial \Phi^{\mathrm{ASM}}}{\partial a_r} = (\xi_r - \mathbf{1}\{r = y\}) - \pi_r \sum_{j=1}^{J} (\psi_j - \mathbf{1}\{m_j = y\}), \qquad r \neq k^\star.$$

To obtain the mixed Hessian entry, differentiate this expression with respect to an expert logit $a_{K+j}$.

The class term $\xi_r - \mathbf{1}\{r = y\}$ depends only on the class logits, so its derivative with respect to $a_{K+j}$ is zero. The coefficient $\pi_r$ also depends only on the non-max class logits on the fixed-$k^\star$ region, so it too is constant with respect to $a_{K+j}$. Thus

$$\frac{\partial^2 \Phi^{\mathrm{ASM}}}{\partial a_r \, \partial a_{K+j}} = -\pi_r \frac{\partial \psi_j}{\partial a_{K+j}}, \qquad r \neq k^\star.$$

Now recall that

$$\psi_j = \sigma(v_j), \qquad v_j = a_{K+j} - \log \sum_{k \neq k^\star} \exp(a_k).$$

Since the second term of $v_j$ involves only class logits, we have

$$\frac{\partial v_j}{\partial a_{K+j}} = 1.$$

Applying the sigmoid derivative then gives

$$\frac{\partial \psi_j}{\partial a_{K+j}} = \psi_j(1 - \psi_j).$$

Substituting this into the previous display yields the full mixed entry:

$$\frac{\partial^2 \Phi^{\mathrm{ASM}}}{\partial a_r \, \partial a_{K+j}} = -\pi_r \, \psi_j(1 - \psi_j), \qquad r \neq k^\star.$$

For $r = k^\star$, the expert-leakage term is absent already at first order, so the mixed derivative is zero. Therefore the non-max-class/expert mixed block is exactly

$$H_{\mathrm{mix}} = -\pi \, d^\top, \qquad \pi_r = \frac{\exp(a_r)}{\sum_{k \neq k^\star} \exp(a_k)}, \qquad d_j = \psi_j(1 - \psi_j).$$

This formula contains several useful facts at once.

**Rank-one structure.** The mixed block is an outer product, so it has rank at most one. The coupling is therefore highly structured, not arbitrary: all non-max class coordinates interact with the expert block through the same class-side direction $\pi$.

**Operator-norm bound.** Because $H_{\mathrm{mix}} = -\pi d^\top$, its operator norm is

$$\|H_{\mathrm{mix}}\|_{\mathrm{op}} = \|\pi\|_2 \, \|d\|_2.$$

Now $\pi$ is a probability vector over the non-max classes, hence $\|\pi\|_2 \leq 1$. Also each entry of $d$ satisfies

$$0 \leq d_j = \psi_j(1 - \psi_j) \leq \frac{1}{4},$$

because the scalar function $u \mapsto u(1 - u)$ is maximized at $u = 1/2$. Hence

$$\|d\|_2 \leq \sqrt{\sum_{j=1}^{J} \left(\frac{1}{4}\right)^2} = \frac{\sqrt{J}}{4},$$

and therefore

$$\|H_{\mathrm{mix}}\|_{\mathrm{op}} \leq \frac{\sqrt{J}}{4}.$$

**Interpretation.** The factor $\pi_r$ measures how active non-max class $r$ is inside the class-side comparison, while $\psi_j(1-\psi_j)$ measures how uncertain expert $j$ currently is. The mixed coupling is strongest when both are non-negligible: there is nontrivial class competition among the non-max classes, and many experts are in the uncertain middle regime rather than near 0 or 1.

This is exactly the regime in which the decoupled surrogate differs most sharply. For the decoupled surrogate, the mixed class–expert block is identically zero. For A-SM, the target is already correct, but the optimization geometry still contains a structured interaction whose norm can grow with the expert pool.

For completeness, note also what is *not* happening. The expert block does not introduce expert–expert coupling through the other expert logits: each expert head still depends only on its own expert coordinate and the class-side reference level. The mixed block is therefore the central second-order object, because it is the unique place where the class and expert tasks remain tied together.

### G.5. Head-to-Head with the Decoupled Surrogate

**Axis (i): statistical target.** A-SM and the decoupled surrogate are aligned on the conditional target. Both recover the Bayes-sufficient quantities at the population optimum: a categorical class posterior together with bounded expert probabilities. The difference between them is therefore not about whether they estimate the right object, but about how they organize the optimization problem that reaches that object.

This point is worth stating carefully because it distinguishes A-SM from the earlier baselines. For additive CE, PiCCE, and Mao25, the criticism already appears at the level of what the surrogate is trying to fit. For A-SM, that is no longer the issue. A-SM repairs the population target. In that sense it is much closer to the decoupled surrogate than the other augmented baselines. The remaining gap is entirely geometric.

**Axis (ii): local geometry.** A-SM still routes the expert objectives through the class scores, so the class gradient contains the expert-leakage term of Proposition 7. The decoupled surrogate removes this term completely. Its class gradient depends only on $(p, y)$ and its mixed Hessian block is zero. This is the key residual weakness of A-SM. It has already repaired the target distortion and boundedness problems, but it has not fully separated the class and expert subproblems. The decoupled surrogate is the only method in our comparison that keeps the same target alignment while making the mixed block identically zero.

The practical consequence is that the two methods react differently during training. Under the decoupled surrogate, an error in expert $j$ changes only expert $j$'s update. Under A-SM, that same expert error also perturbs the class update whenever the non-max class normalizer is active. So even though both methods are aiming for the same conditional optimum, they do not follow the same local route toward that optimum. This is the sharpest way to understand the residual gap between the two methods: A-SM is target-correct but route-coupled, whereas the decoupled surrogate is both target-correct and route-decoupled.

### G.6. Worked Example: Complementary Experts

Consider $K = 3$ classes and $J = 4$ complementary experts. Fix a region where the maximizing class index is $k^\star = 1$, and inspect the gradient on the class-2 logit $a_2$ for a sample with $y = 2$. Suppose

$$\xi_2 = 0.25, \qquad \pi_2 = 0.40, \qquad \sum_{j=1}^{4}\big(\psi_j - \mathbf{1}\{m_j = y\}\big) = 0.80.$$

Then Proposition 7 gives

$$\frac{\partial \Phi^{\mathrm{ASM}}}{\partial a_2} = (\xi_2 - 1) - \pi_2 \sum_{j=1}^{4}\big(\psi_j - \mathbf{1}\{m_j = y\}\big) = -0.75 - 0.32 = -1.07.$$

The first term, $-0.75$, is the ordinary class signal saying that class 2 is the true label and its current probability estimate is too small. The second term, $-0.32$, is not class information at all. It is leakage from the expert heads. It appears only because A-SM routes every expert residual through the non-max class normalizer, and it becomes larger as the aggregate expert residual grows.

This example is deliberately chosen so that the issue is easy to see. If A-SM were fully decoupled, the class-2 logit would receive exactly the ordinary cross-entropy gradient $-0.75$. Instead, the expert block contributes an extra $-0.32$, which changes both the magnitude and the direction of the local class update. The class learner is therefore not reacting purely to class error on this sample; it is also reacting to unresolved expert uncertainty.

Under the decoupled surrogate, the same sample would have class gradient

$$\frac{\partial \Phi^{\mathrm{dec}}}{\partial w_2} = p_2 - 1,$$

with no expert contribution at all. So the class update would reflect only the classifier's own error on the sample. This is the clearest numerical illustration of the axis-(ii) difference between A-SM and the decoupled surrogate: both can target the right optimum, but only the decoupled surrogate keeps the class update free of expert noise.

The example also makes clear why the issue can worsen with many experts. The leakage enters through the aggregate residual $\sum_{j=1}^{J}(\psi_j - \mathbf{1}\{m_j = y\})$. When several experts are still in their uncertain regime, those residuals add before being injected into the class gradient. The decoupled surrogate has no analogous accumulation because each expert head is optimized in isolation.

The conclusion is therefore narrower than for additive CE or PiCCE, but still important. A-SM shows that fixing the statistical target is not enough. One can correct boundedness and recover the right population quantities, yet still retain a shared local geometry that mixes class and expert learning. The decoupled surrogate closes that final gap by keeping the same target alignment while removing the mixed interaction altogether.

## H. OvA (Verma et al., 2022)

The multi-expert one-vs-all construction of Verma et al. (2022) is the cleanest augmented baseline on the optimization axis. It removes the shared softmax entirely and replaces it by independent binary logistic tasks. As a result, none of the earlier geometric pathologies remain: there is no multiplicity amplification, no winner-take-all starvation, and no mixed class–expert curvature.

The remaining question is different. What is lost when the class side is represented not by one categorical distribution in $\Delta^K$, but by $K$ independent Bernoulli scores? At the conditional optimum, OvA still identifies the correct class marginals $\eta_k(x)$ and expert utilities $\alpha_j(x)$. The gap appears away from that optimum. For arbitrary parameters, the class vector produced by OvA need not lie on the simplex, so the largest class score is not automatically the confidence of a coherent categorical model. This section makes that distinction explicit.

### H.1. Formulation

OvA separates the routing problem into $K + J$ binary tasks: one logistic problem for each class and one logistic problem for each expert. The class coordinates ask whether the true label equals a given class, while the expert coordinates ask whether a given expert is correct on the current example.

**Definition 18** (Multi-expert OvA, Verma et al. (2022))**.** Let $g(x) \in \mathbb{R}^K$ denote class scores and $s(x) \in \mathbb{R}^J$ denote expert scores. Define

$$\gamma_k(y) := 2\mathbf{1}[y = k] - 1, \qquad \zeta_j(y, m) := 2\mathbf{1}[m_j = y] - 1.$$

$$\Phi^{\mathrm{OvA}}(g, s;\, x, y, m) := \sum_{k=1}^{K} \log\bigl(1 + \exp(-\gamma_k(y) g_k(x))\bigr) \;+\; \sum_{j=1}^{J} \log\bigl(1 + \exp(-\zeta_j(y, m) s_j(x))\bigr).$$

This is the signed-label form used in Verma et al. (2022). For the analysis below it is convenient to rewrite the same loss using the binary targets

$$b_k(y) := \mathbf{1}\{y = k\}, \qquad d_j(y, m) := \mathbf{1}\{m_j = y\}.$$

Using the standard identity

$$\log\bigl(1 + \exp(-(2b - 1)v)\bigr) = -b \log \sigma(v) - (1 - b) \log(1 - \sigma(v)), \qquad b \in \{0, 1\},$$

the OvA loss can be written equivalently as

$$\Phi^{\mathrm{OvA}}(g, s;\, x, y, m) = \sum_{k=1}^{K}\Big(-b_k(y)\log\sigma(g_k(x)) - (1-b_k(y))\log(1-\sigma(g_k(x)))\Big)$$
$$+\sum_{j=1}^{J}\Big(-d_j(y,m)\log\sigma(s_j(x)) - (1-d_j(y,m))\log(1-\sigma(s_j(x)))\Big).$$

This form makes two features explicit. First, every coordinate is optimized independently. Second, the class head is not a softmax over the $K$ classes; it is a vector of $K$ separate sigmoids.

The prediction rule chooses the classifier if $\max_{k\in[K]} g_k(x) > \max_{j\in[J]} s_j(x)$ and otherwise defers to the largest expert score. Since the sigmoid is strictly increasing, this is equivalent to comparing $\max_k \sigma(g_k(x))$ and $\max_j \sigma(s_j(x))$ on the probability scale.

### H.2. Conditional Minimizer

**Proposition 19** (OvA conditional minimizer)**.** *The conditional OvA risk is a sum of $K+J$ independent Bernoulli cross-entropies. Its optimal probabilities satisfy $\sigma(g_k^\star(x)) = \eta_k(x)$ and $\sigma(s_j^\star(x)) = \alpha_j(x)$ whenever the corresponding targets lie in $(0,1)$. If a target equals $0$ or $1$, the conditional infimum is approached as the corresponding logit tends to $\mp\infty$ or $\pm\infty$, respectively.*

**Proof** Fix an input $x$. Taking conditional expectation over $(Y, M) \mid X = x$ gives

$$\mathcal{C}_{\mathrm{OvA}}(g, s \mid x) = \sum_{k=1}^{K}\Big(-\eta_k(x)\log\sigma(g_k) - (1-\eta_k(x))\log(1-\sigma(g_k))\Big)$$
$$+\sum_{j=1}^{J}\Big(-\alpha_j(x)\log\sigma(s_j) - (1-\alpha_j(x))\log(1-\sigma(s_j))\Big).$$

The key point is that this expression separates completely across coordinates. There is one Bernoulli cross-entropy for each class coordinate and one Bernoulli cross-entropy for each expert coordinate. The class-side target for coordinate $k$ is $\eta_k(x)$, while the expert-side target for coordinate $j$ is $\alpha_j(x)$.

To make the argument explicit, consider the generic Bernoulli cross-entropy with target parameter $\theta \in [0,1]$:

$$\varphi(v) = -\theta\log\sigma(v) - (1-\theta)\log(1-\sigma(v)),$$

where $v$ is the Bernoulli logit. Differentiating once gives

$$\varphi'(v) = \sigma(v) - \theta.$$

Thus any finite stationary point must satisfy $\sigma(v) = \theta$. Differentiating again gives

$$\varphi''(v) = \sigma(v)(1-\sigma(v)) \geq 0.$$

Hence $\varphi$ is convex, and it is strictly convex whenever $\sigma(v) \in (0,1)$. If $\theta \in (0,1)$, the condition $\sigma(v) = \theta$ identifies the unique finite minimizer. If $\theta = 1$, then $\varphi'(v) = \sigma(v) - 1 < 0$ for every finite $v$, so the risk decreases monotonically and its infimum is approached as $v \to +\infty$. Similarly, if $\theta = 0$, then $\varphi'(v) = \sigma(v) > 0$ for every finite $v$, so the infimum is approached as $v \to -\infty$.

Applying this argument coordinatewise yields

$$\sigma(g_k^\star(x)) = \eta_k(x), \qquad \sigma(s_j^\star(x)) = \alpha_j(x),$$

for all $k \in [K]$ and $j \in [J]$. ■

This proposition is the positive side of OvA. Unlike Additive CE or PiCCE, OvA does not distort the conditional target. At the probability level, it identifies the correct class-marginal quantities $\eta_k(x)$ and the correct expert utilities $\alpha_j(x)$. The limitation of OvA is therefore not that it aims at the wrong population object. The limitation is representational: the class head becomes a coherent categorical posterior only at the optimum, not by construction at intermediate parameter values.

### H.3. Curvature

The local geometry of OvA is as simple as its conditional risk. Because each loss term depends on only one score coordinate, the first and second derivatives separate completely.

Consider first a class coordinate $g_k$. Its samplewise contribution is

$$L_k(g_k; y) = -b_k(y) \log \sigma(g_k) - (1 - b_k(y)) \log(1 - \sigma(g_k)).$$

Differentiating with respect to $g_k$ gives

$$\frac{\partial L_k}{\partial g_k} = \sigma(g_k) - b_k(y).$$

Differentiating once more gives

$$\frac{\partial^2 L_k}{\partial g_k^2} = \sigma(g_k)(1 - \sigma(g_k)).$$

Exactly the same calculation applies to an expert coordinate $s_j$:

$$L_j(s_j; y, m) = -d_j(y, m) \log \sigma(s_j) - (1 - d_j(y, m)) \log(1 - \sigma(s_j)),$$

so

$$\frac{\partial L_j}{\partial s_j} = \sigma(s_j) - d_j(y, m), \qquad \frac{\partial^2 L_j}{\partial s_j^2} = \sigma(s_j)(1 - \sigma(s_j)).$$

Now observe that no class term depends on any other class coordinate or on any expert coordinate, and no expert term depends on any class coordinate or any other expert coordinate. Therefore every mixed partial derivative is zero:

$$\frac{\partial^2 \Phi^{\mathrm{OvA}}}{\partial g_k \, \partial g_r} = 0 \quad (k \neq r), \qquad \frac{\partial^2 \Phi^{\mathrm{OvA}}}{\partial s_j \, \partial s_\ell} = 0 \quad (j \neq \ell), \qquad \frac{\partial^2 \Phi^{\mathrm{OvA}}}{\partial g_k \, \partial s_j} = 0.$$

If we concatenate all coordinates into one vector $a = (g_1, \ldots, g_K, s_1, \ldots, s_J) \in \mathbb{R}^{K+J}$, the Hessian is therefore diagonal:

$$\nabla_a^2 \Phi^{\mathrm{OvA}} = \mathrm{Diag}\big(\sigma(a_i)(1 - \sigma(a_i))\big),$$

where the diagonal entries corresponding to class coordinates are $\sigma(g_k)(1 - \sigma(g_k))$ and those corresponding to expert coordinates are $\sigma(s_j)(1 - \sigma(s_j))$.

The eigenvalues of a diagonal matrix are exactly its diagonal entries. Since $0 \leq \sigma(v)(1 - \sigma(v)) \leq 1/4$ for every $v \in \mathbb{R}$, we obtain

$$\lambda_{\max}\big(\nabla_a^2 \Phi^{\mathrm{OvA}}\big) \leq \frac{1}{4}.$$

This is the cleanest geometry among the baselines considered in the paper. There is no class–expert coupling, no winner-take-all effect, and no multiplicity amplification. If OvA has a weakness, it is not a weakness of local optimization geometry.

### H.4. Head-to-Head with the Decoupled Surrogate

**Axis (i): statistical target.** OvA and the decoupled surrogate agree on the most important population statement. At the conditional optimum, both methods recover the class marginals $\eta_k(x)$ and the expert utilities $\alpha_j(x)$. In that sense, OvA is much closer to the decoupled surrogate than the earlier augmented-softmax surrogates.

The difference lies in how these quantities are represented away from the optimum. The decoupled surrogate parameterizes the class side as one categorical distribution $p(x) \in \Delta^K$ for every parameter value. The vector $p(x)$ is therefore a coherent class posterior throughout training. OvA instead parameterizes the class side as $K$ independent sigmoids. Each coordinate can be interpreted as a class marginal, but the full vector need not lie in $\Delta^K$ unless the model happens already to be at the conditional optimum.

**Axis (ii): local geometry.** On the geometric axis the two methods are genuinely similar. Both are coordinatewise, both have bounded curvature, and both avoid all of the shared-geometry pathologies that affect the augmented-action family. The decoupled surrogate is not preferred because it is more decoupled than OvA; there is little to improve on that axis.

The decoupled surrogate is preferred because it combines this clean geometry with a categorical class model. The routing rule in the decoupled surrogate always compares $\max_k p_k(x)$ and $\max_j u_j(x)$ on a coherent probability scale. OvA also compares quantities in $[0, 1]$, but the class side is not forced to behave like one categorical posterior during training. Several class sigmoids can be simultaneously large, and the largest one can overstate classifier confidence relative to the expert utilities.

### H.5. Worked Example: Several Class Sigmoids Above $0.5$

Consider $K = 3$ and suppose that during training the OvA class head outputs

$$\sigma(g_1) = 0.77, \qquad \sigma(g_2) = 0.73, \qquad \sigma(g_3) = 0.69.$$

These values are individually admissible Bernoulli probabilities, but together they cannot form a categorical posterior because they sum to $2.19$. Nothing in the OvA objective penalizes this. Each class coordinate is trained separately, so the model is free to regard several classes as simultaneously likely in the one-vs-all sense.

Now suppose that the best expert estimate is

$$\sigma(s_1) = 0.70.$$

The OvA routing rule compares the largest class score with the best expert score. Since

$$\max_{k\in[K]} \sigma(g_k) = 0.77 > 0.70 = \max_{j\in[J]} \sigma(s_j),$$

OvA predicts rather than defers.

Now contrast the decoupled surrogate. A coherent class posterior for the same ambiguous input could be

$$p = (0.40, 0.35, 0.25), \qquad u_1 = 0.70.$$

Then the Bayes-style comparison is immediate: $\max_k p_k = 0.40 < 0.70$, so the decoupled surrogate defers. This does not mean that OvA is inconsistent. If the model class is rich enough and training finds the conditional optimum, the OvA class sigmoids will match the true class-marginal probabilities and therefore sum to one. The point is instead that the intermediate states visited during training need not have that property. Because routing uses the current scores, this lack of simplex structure can matter in practice.

The example isolates the remaining gap between OvA and the decoupled surrogate. OvA has already fixed the geometric side of the problem. What it has not fixed is the representation of class uncertainty. The decoupled surrogate keeps the same clean optimization geometry while ensuring that the class side remains a categorical posterior at every point along training.

## I. Comparison of Excess-Risk Constants

We compare the surrogate-regret transfer guarantees of the decoupled surrogate with the available baselines, focusing on how the calibration constant scales with the size of the multi-expert action space. Two distinct prior references need to be separated carefully. The constant $\sqrt{2(J+1)}$ comes from the earlier comp-sum reduction analyzed by Mao et al. (2024e), which covers the Mozannar-style augmented-action construction. It is not the constant for the later Mao25 single-stage surrogate (Mao et al., 2025d). For the Mao25 single-stage surrogate used in this paper, the relevant excess-risk factor scales with the augmented action dimension $K + J$.

**The decoupled surrogate (Theorem 10).** The decoupled surrogate satisfies

$$\mathcal{E}_\perp - \mathcal{E}_\perp^\star \;\le\; \underbrace{\max\left\{2\sqrt{2},\ \sqrt{\tfrac{2J}{\lambda}}\right\}}_{C_{\mathrm{dec}}} \sqrt{\mathcal{E}_{\mathrm{dec}} - \mathcal{E}_{\mathrm{dec}}^\star}\,.$$

Whenever the per-expert weight satisfies $\lambda/J \ge 1/4$, the constant $C_{\mathrm{dec}} = 2\sqrt{2} \approx 2.83$ is independent of $J$.

**Add. CE (Mozannar & Sontag, 2020).** For Add. CE, Mao et al. (2024e) derive a transfer constant $C_{\mathrm{CE}} = \sqrt{2(J+1)}$, which grows as $O(\sqrt{J})$. This is the $\sqrt{2(J+1)}$ factor quoted in the main paper for the Add. CE baseline.

**Mao25 (Mao et al., 2025d).** The single-stage surrogate studied by Mao et al. (2025d) should be kept separate from the earlier comp-sum reduction above. Quantitatively, the corresponding excess-risk constant scales with the full augmented action dimension. In the multi-expert setting considered here, this yields

$$C_{\mathrm{Mao25}} = K + J,$$

so the dependence on the number of experts is linear: $O(J)$ for fixed $K$. This is a different regime from the Add. CE bound above: Add. CE inherits the earlier $\sqrt{2(J+1)}$ comp-sum constant, whereas Mao25 scales with the entire augmented action count.

**Other surrogates.** For the remaining baselines, the cited analyses establish Bayes consistency but do not give explicit transfer constants: (Verma et al., 2022; Cao et al., 2024; Liu et al., 2026).

**Summary.**

| Surrogate | Consistency type | Transfer constant | Dependence on $J$ |
|---|---|---|---|
| Add. CE | Bayes | — | — |
| PiCCE | Bayes | — | — |
| OvA | Bayes | — | — |
| A-SM | Bayes | — | — |
| Add. CE | Quantitative | $\sqrt{2(J+1)}$ | $O(\sqrt{J})$ |
| Mao25 | Quantitative | $K+J$ | $O(J)$ for fixed $K$ |
| **Decoupled** | Quantitative | $\max\{2\sqrt{2}, \sqrt{2J/\lambda}\}$ | $O(1)$ for fixed $\lambda/J$ |

Among the multi-expert surrogates with quantitative transfer constants, the decoupled surrogate is the only one whose transfer constant does not grow with $J$ for a fixed per-expert weight $\lambda/J$. When $\lambda/J \geq 1/4$, the decoupled surrogate constant simplifies to $2\sqrt{2}$. Add. CE inherits a $\sqrt{2(J+1)}$ factor from the earlier deferral analysis, while the Mao25 surrogate scales with the full augmented action dimension $K+J$. The difference reflects a structural advantage: the decoupled surrogate estimates $(\eta, \alpha)$ through independent subproblems, so the conditional surrogate excess decomposes into a sum of KL terms that can be converted to estimation error without cross-contamination. The augmented-action surrogates embed classes and experts in a shared geometry whose normalization couples the KL structure, so the quantitative transfer factor worsens as the action space grows.

## J. Compute Resources

All experiments were run on NVIDIA A100 GPUs with 40GB of memory. Each run uses a single GPU; multi-seed results are obtained by repeating the same configuration independently across seeds. The synthetic suites are lightweight and finish within minutes per seed, while the CIFAR-10, CIFAR-10H, and Covertype experiments fit comfortably within the same single-GPU setting.

## K. Synthetic Experiments

This appendix reports the retained synthetic suites used in the paper. Each suite instantiates one mechanism from Section 4 under a distribution for which the Bayes rule is known exactly. In every suite we specify $(\eta, \alpha)$ analytically, so the Bayes-optimal L2D action at each $x$ is determined by the comparison

$$\max_{k\in[K]} \eta_k(x) \qquad \text{versus} \qquad \max_{j\in[J]} \alpha_j(x),$$

and the exact defer regret $\Delta_\perp(f) := \mathcal{E}_\perp(f) - \mathcal{E}_\perp^\star$ can be evaluated directly.

**Protocol.** All suites use linear models (no hidden layer) unless stated otherwise. Every method receives the same feature vector $X$ and training tuple $(X, Y, M)$. The ground-truth quantities $\eta(x)$ and $\alpha(x)$ are used only *after* training to evaluate the

exact Bayes regret, and on the validation split to select the checkpoint with the lowest true defer loss. This model-selection rule is applied uniformly to every surrogate. All reported numbers are means with one standard deviation in parentheses over multiple seeds (3–5, stated per suite).

Each suite is written in the same order: *objective*, *setting*, *Bayes rule and link to the theory*, and *result*.

### K.1. Nested Redundant Experts

**Objective.** Vary the overlap multiplicity while keeping the Bayes decision problem fixed, so that any performance degradation must be attributed to the surrogate itself, not to a harder task. Although the suite is designed around Proposition 4 (additive CE amplification), the nesting also stresses every other baseline for surrogate-specific reasons discussed in the analysis below.

**Setting.** The retained setting uses $K$=16 classes, $J$=24 experts, $n_{\text{train}}$=900, and $n_{\text{test}}$=8,000. All models are linear (no hidden layer).

*Features and class posteriors.* Each point draws a region indicator $R \sim \text{Bernoulli}(0.65)$. The feature vector lies in $\mathbb{R}^K$. On the defer-favorable region $D$ ($R$=1), the label is uniform on $[K]$, $\eta_k(x)$=$1/K$ for all $k$, and $X = 3.5\,\mathbf{1}_K + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, 0.05^2 I_K)$; hence no class is linearly separable. On $D^c$ ($R$=0), the class posterior is near-deterministic, $\eta_Y(x)$=0.998, and the feature for the true label is shifted to 5 (easily separable).

*Expert utilities.* The $J$ experts have monotonically decreasing accuracies on $D$. Define a spacing variable $\rho_j := \log(1{+}j)/\log(J)$ for $j \in \{0, \dots, J{-}1\}$ (logarithmic spacing, so the first few experts are close in quality and the last ones are markedly weaker). On $D$,

$$\alpha_j(x) \;=\; 0.99 \;-\; (0.99 - 0.75)\,\rho_j, \tag{22}$$

giving $\alpha_0$=0.99 for the strongest expert down to $\alpha_{J-1}$=0.75 for the weakest. On $D^c$, all experts are near-random: $\alpha_j(x) = 0.04 - (0.04 - 0.002)\,\rho_j$.

*Nested correctness.* A single latent $U \sim \text{Unif}[0,1]$ is drawn per sample, and expert $j$ is correct iff $U \le \alpha_j(x)$. Because $\alpha_0 \ge \alpha_1 \ge \cdots \ge \alpha_{J-1}$, this implies

$$\{M_{j+1} = Y\} \;\subseteq\; \{M_j = Y\} \qquad \text{for all } j.$$

When expert $j$ is incorrect, its prediction is drawn uniformly from $[K] \setminus \{Y\}$. The nesting guarantees that adding experts never changes the Bayes action: expert 0 is always Bayes-best on $D$, and classification is always best on $D^c$. What grows with $J$ is only the realized overlap $|\mathcal{J}(Y, M)|$.

**Bayes rule.** *On $D$:* defer to expert 0. *On $D^c$:* classify. This is invariant for all $J$.

**Why each baseline degrades.** The propositions in Section 4 predict a different failure mechanism for each surrogate:

- **Add. CE** (Proposition 4). The gradient carries the factor $1{+}|\mathcal{J}|$, and the Hessian scales identically. With $J$=24 nested experts and high $\alpha$ values, a typical defer-region sample has $|\mathcal{J}|{\approx}18$, so each such sample is weighted $\sim$19$\times$ compared to a classify-region sample. This gives high-overlap samples disproportionate gradient and curvature relative to samples near the classify-versus-defer boundary.
- **PiCCE** (Proposition 5). Only the winning correct expert receives a negative gradient; all other correct experts receive the positive push $2q_{K+j}{>}0$. With 24 nested experts, as many as 23 correct experts are suppressed on samples where all experts are correct, and in general $|\mathcal{J}| - 1$ correct experts are suppressed whenever multiple experts are correct. The resulting winner-take-all lock-in concentrates all expert mass on whichever expert had the highest logit at initialization, preventing the model from learning the correct ranking $\alpha_0{>}\cdots{>}\alpha_{J-1}$.
- **A-SM** (Proposition 7). The class-logit gradient contains the expert leakage term $-\pi_r \sum_j (\psi_j - t_j)$. When many experts are correct, the residuals $\psi_j{-}t_j$ accumulate, injecting noise into the class update. In a linear model, class and expert logits share the same parameters through the feature matrix, so this leakage directly corrupts the classify-versus-defer boundary.
- **OvA** (Appendix H.2). The $K$ independent sigmoid class heads are not coupled by a simplex constraint, so the class confidence $\max_k \sigma(g_k)$ can be miscalibrated. With $K$=16 uniform classes on $D$, the population binary targets are $1/16$

for each class, but finite-sample and optimization errors are not constrained to preserve a simplex. The maximum over 16 independently trained sigmoids can therefore be inflated relative to the true $\max_k \eta_k = 1/16$, biasing the routing comparison toward classifying when it should defer.

- **Mao25** (Proposition 6). The gradient factor $(S_{\mathcal{S}}-1)$ is common to all acceptable actions, so the update cannot rank experts. Additionally, the MAE-like structure means $\sum_{i\in\mathcal{S}} -\partial\Phi/\partial a_i = S_{\mathcal{S}}(1-S_{\mathcal{S}}) \leq 1/4$, so the total corrective signal is weak and vanishes as $S_{\mathcal{S}} \to 0$, creating slow convergence toward the defer region.

**Result.** Table 4 reports the retained $J$=24 point. The broader suite sweeps $J \in \{16, 24, 32\}$ (Table 5).

| Method | Exact defer regret | System accuracy |
|---|---|---|
| **Decoupled** | **0.0002** (0.0001) | **0.993** (0.001) |
| PiCCE | 0.238 (0.116) | 0.755 (0.118) |
| OvA | 0.294 (0.020) | 0.699 (0.020) |
| A-SM | 0.304 (0.035) | 0.688 (0.035) |
| Mao25 | 0.323 (0.035) | 0.669 (0.036) |
| Add. CE | 0.348 (0.038) | 0.646 (0.036) |

*Table 4.* Nested redundant suite at $J$=24. The decoupled surrogate remains essentially Bayes-optimal; all baselines degrade severely. Means with std in parentheses over 3 seeds.

| | $J = 16$ | $J = 24$ | $J = 32$ |
|---|---|---|---|
| Decoupled defer regret | 0.013 | 0.0002 | 0.007 |
| Best non-decoupled defer regret | 0.226 | 0.238 | 0.293 |
| Best non-decoupled method | PiCCE | PiCCE | OvA |

*Table 5.* $J$-sweep summary. The decoupled surrogate stays near Bayes-optimal as the expert pool grows, while the best non-decoupled regret increases monotonically.

**Analysis.** Because the Bayes rule is unchanged across the $J$-sweep, the large degradation of every baseline cannot be explained by a harder routing problem; it is attributable to how each surrogate responds to redundant multiplicity. The decoupled surrogate remains near Bayes-optimal ($\Delta_\perp$=0.0002 at $J$=24) because its expert loss decomposes into $J$ independent BCE terms: each $\alpha_j$ is estimated coordinatewise, and the gradient $\partial\Phi^{\mathrm{dec}}/\partial s_j = (\lambda/J)(u_j - t_j)$ carries no cross-expert coupling and no dependence on $|\mathcal{J}|$ (Proposition 9). As the $J$-sweep in Table 5 shows, the best non-decoupled regret *increases* with $J$ (from 0.226 to 0.293), while the decoupled surrogate stays below 0.013 throughout, confirming that the baselines' failures scale with the overlap multiplicity as predicted by the theory.

### K.2. Rare Specialist for PiCCE

**Objective.** Isolate the winner-take-all starvation from Proposition 5. The construction places a specialist expert that is Bayes-preferred on a rare region $R$ but whose correct events are mostly shared with a stronger generalist. PiCCE should suppress the specialist because only the current $\arg\max$ winner receives the reinforcing gradient $2q_{K+j^\star} - 1 < 0$, while every other correct expert is pushed down by $2q_{K+j} > 0$ (eq. (7)).

**Setting.** We use $K$=2 classes, $J$=2 experts (generalist $j$=1, specialist $j$=2), $n_{\mathrm{train}}$=7,000, $n_{\mathrm{test}}$=18,000, and 5 seeds. All models are linear.

*Features and class posteriors.* Draw $Z \sim \mathrm{Unif}([-1,1]^2)$ and an independent rare indicator $R \sim \mathrm{Bernoulli}(0.15)$. The feature vector is $X = (Z_1, Z_2, \mathbf{1}\{R\}) \in \mathbb{R}^3$. The class posterior depends on the region:

$$\max_k \eta_k(x) = \begin{cases} 0.60, & x \in R, \\ 0.90, & x \notin R, \end{cases}$$

with the maximizing class determined by $\mathrm{sgn}(Z_1)$. Classification is therefore easy outside $R$ ($\eta_{\max}$=0.90) and uncertain on $R$ ($\eta_{\max}$=0.60).

*Expert utilities.* Outside $R$, expert correctness is independent: $\alpha_{\mathrm{gen}}(x)$=0.45, $\alpha_{\mathrm{spec}}(x)$=0.15. Neither expert is useful, and the Bayes action is to classify. On $R$, the marginal utilities are:

$$\alpha_{\mathrm{spec}}(x) = 0.75 \;>\; \alpha_{\mathrm{gen}}(x) = 0.60 \;=\; \max_k \eta_k(x) = 0.60,$$

so the Bayes action is to defer to the specialist, with a tie between the generalist and the classifier below the specialist.

*Joint correctness on $R$.* The key design choice is that correctness is *not* independent on $R$. We specify the joint law directly:

| | Specialist correct | Specialist incorrect |
|---|---|---|
| Generalist correct | 0.55 | 0.05 |
| Generalist incorrect | 0.20 | 0.20 |

This gives $\alpha_{\text{spec}}$=0.55+0.20=0.75 and $\alpha_{\text{gen}}$=0.55+0.05=0.60 as required. Critically, when the specialist is correct on $R$, the generalist is also correct with probability

$$\Pr(C_{\text{gen}}{=}1 \mid C_{\text{spec}}{=}1,\, x \in R) \;=\; \frac{0.55}{0.75} \;=\; 0.73.$$

So 73% of the specialist's positive training signal comes from shared-correct samples where both experts are right.

**Bayes rule.** *On $R$* (15% of data): defer to the specialist ($\alpha_{\text{spec}}{=}0.75 > \alpha_{\text{gen}}{=}0.60 = \eta_{\max}{=}0.60$). *On $R^c$* (85% of data): classify ($\eta_{\max}{=}0.90 > \alpha_{\text{gen}}{=}0.45$).

**Why PiCCE fails.** On the 73% of specialist-correct samples on $R$ where the generalist is also correct ($\mathcal{J}{=}\{1,2\}$), PiCCE selects the winner $j^\star{=}\arg\max_{j\in\mathcal{J}} a_{K+j}$. The gradient on the *non-winning* correct expert is $\partial\Phi^{\text{PiCCE}}/\partial a_{K+j} = 2q_{K+j} > 0$ (Proposition 5), which pushes its logit *down*. If the generalist starts with a marginally higher logit (by initialization noise), the specialist is suppressed on every shared-correct sample. The gap compounds across iterations: the generalist wins more $\arg\max$ selections, receives more reinforcing gradient, and the specialist is driven to zero. On the 27% of specialist-correct events that are specialist-only ($C_{\text{gen}}{=}0$), the specialist does receive a negative gradient, but $15\% \times 27\%{=}4\%$ of total training data is not enough to overcome the opposite signal from the dominant shared-correct mass.

**Why the decoupled surrogate succeeds.** The decoupled surrogate trains each expert head with an independent BCE loss: $\partial\Phi^{\text{dec}}/\partial s_j = (\lambda/J)(u_j{-}t_j)$, where $u_j{=}\sigma(s_j)$ and $t_j{=}\mathbf{1}\{M_j{=}Y\}$. On a shared-correct sample, *both* experts receive the reinforcing gradient $u_j{-}1 < 0$ independently. No $\arg\max$ selection occurs, so the generalist's logit does not compete with the specialist's. Both utilities are estimated faithfully, and the routing comparison $\sigma(s_{\text{spec}}) \gtrless \sigma(s_{\text{gen}})$ learns $0.75 > 0.60$ from the data.

**Metrics.** We report two specialist-specific diagnostics alongside the standard defer regret and system accuracy:

- *Specialist selection on $R$*: $\Pr(r(X){=}\text{spec} \mid X \in R)$, computed over all test samples in $R$ (including those that classify; classifying samples count as not selecting the specialist).
- *Shared-correct routing*: $\Pr(r(X){=}\text{spec} \mid X \in R,\, C_{\text{gen}}{=}1,\, C_{\text{spec}}{=}1)$, i.e. specialist selection restricted to the shared-correct subset. This is the hardest case: both experts are right, and only the ranking $\alpha_{\text{spec}} > \alpha_{\text{gen}}$ distinguishes them.

**Result.** Table 6 reports the retained result (means with std in parentheses over 5 seeds).

| Method | Defer regret | System acc. | Coverage | Spec. sel. on $R$ | Shared-correct routing |
|---|---|---|---|---|---|
| **Decoupled** | **0.003** (0.002) | 0.875 (0.002) | 0.849 | **0.993** (0.007) | **0.994** (0.007) |
| OvA | 0.003 (0.001) | 0.875 (0.003) | 0.850 | 0.988 (0.023) | 0.987 (0.026) |
| A-SM | 0.003 (0.002) | 0.875 (0.003) | 0.850 | 0.988 (0.017) | 0.988 (0.017) |
| Mao25 | 0.026 (0.002) | 0.852 (0.004) | 1.000 | 0.000 (0.000) | 0.000 (0.000) |
| PiCCE | 0.035 (0.001) | 0.843 (0.002) | 0.821 | 0.000 (0.000) | 0.000 (0.000) |
| Add. CE | 0.324 (0.067) | 0.554 (0.069) | 0.128 | 1.000 (0.000) | 1.000 (0.000) |

*Table 6.* Rare-specialist suite. *Coverage* is $\Pr(r(X){=}\text{classify})$; Bayes-optimal coverage is 0.85. PiCCE defers but never to the specialist; Mao25 never defers at all; The decoupled surrogate, OvA, and A-SM recover the specialist almost perfectly.

**Analysis.** The table reveals three distinct failure modes.

*PiCCE (winner-take-all starvation).* PiCCE has reasonable coverage (0.821, close to the Bayes-optimal 0.85), so it does learn the classify-versus-defer boundary. However, its specialist selection is exactly 0 on $R$: when it defers, it always picks the generalist. The shared-correct routing rate is also 0, confirming that on the decisive subset where both experts are correct, PiCCE never learns to prefer the specialist. This is the direct empirical realization of Proposition 5: the generalist captures the $\arg\max$ early and the positive-gradient push $2q_{K+\text{spec}} > 0$ suppresses the specialist permanently.

*Mao25 (no deferral learned).* Mao25 has coverage 1.0: it *never* defers and always classifies. Consequently, its specialist selection is trivially 0 (there are no deferral events). Its system accuracy of 0.852 is approximately the accuracy of a Bayes-optimal classifier that ignores experts: $0.85 \times 0.90 + 0.15 \times 0.60 = 0.855$. This is consistent with the MAE-like optimization pathology from Proposition 6: the corrective gradient on acceptable actions has magnitude $S_{\mathcal{S}}(1-S_{\mathcal{S}}) \leq 1/4$, which vanishes when $S_{\mathcal{S}} \approx 0$ early in training. Mao25 never escapes the classify-everything initialization.

*Add. CE (global collapse).* Add. CE has coverage 0.128 (it defers 87% of the time) and defer regret 0.324. Even with only $J{=}2$ experts, the gradient amplification of Proposition 4 is sufficient to destabilize training in this low-$K$ regime: the factor $1{+}|\mathcal{J}|$ reaches 3 on shared-correct samples, and the optimizer overshoots the defer boundary.

*The decoupled surrogate, OvA, and A-SM (correct routing).* All three methods achieve near-Bayes-optimal defer regret (0.003), coverage close to 0.85, and specialist selection $\geq$0.988 on $R$. The task has only $J{=}2$ experts, so the A-SM coupling (Proposition 7) and the OvA miscalibration are negligible here. This confirms that the PiCCE failure is specific to the winner-take-all mechanism and not a general difficulty of the task.

### K.3. Shared Acceptability and Expert Ranking for Mao25

**Objective.** Isolate the ranking limitation of Proposition 6: the Mao25 gradient factor $(S_{\mathcal{S}}{-}1)$ is the same for all acceptable actions, so the samplewise update cannot distinguish the best expert from a merely acceptable one. We construct a task where the main difficulty is expert ranking *within* a defer region under heavy shared acceptability.

**Setting.** We use $K{=}10$ classes, $J{=}4$ experts, $n_{\text{train}}{=}1{,}000$, $n_{\text{test}}{=}12{,}000$, and 3 seeds. All models are linear.

*Features and class posteriors.* Each point draws a defer-region indicator $D \sim \text{Bernoulli}(0.60)$. The feature vector lies in $\mathbb{R}^{K+J} = \mathbb{R}^{14}$. On $D$, a sector $Q \sim \text{Unif}([J])$ is drawn, the label is uniform on $[K]$ ($\eta_k{=}1/10$ for all $k$), and

$$X_{1:K} = 3.1\,\mathbf{1}_K + \varepsilon_{\text{cls}}, \qquad X_{K+Q} = 2.2 + \varepsilon_Q, \qquad \varepsilon \sim \mathcal{N}(0,\, 0.05^2).$$

The first $K$ features are uninformative for the class (all near 3.1); the sector feature $X_{K+Q}$ identifies which expert is locally best. On $D^c$, the class posterior is near-deterministic ($\eta_Y{=}0.998$, feature of the true label shifted to 5).

*Joint expert correctness on $D$.* Inside $D$, the identity of the best expert changes with the sector $Q$. The joint correctness of the $J{=}4$ experts is drawn from the following event model:

| Event | Probability | Who is correct |
|---|---|---|
| All correct | 0.10 | all 4 experts |
| $Q$ + one random other | 0.78 | $Q$ and one uniformly chosen $j{\neq}Q$ |
| $Q$ only | 0.05 | only $Q$ |
| One random non-$Q$ only | 0.03 | one uniformly chosen $j{\neq}Q$ |
| None correct | 0.04 | no expert |

The marginal utilities on $D$ are therefore

$$\alpha_Q(x) \;=\; 0.10 + 0.78 + 0.05 \;=\; 0.93, \qquad \alpha_{j\neq Q}(x) \;=\; 0.10 + \tfrac{0.78}{3} + \tfrac{0.03}{3} \;=\; 0.37.$$

On $D^c$, all experts have utility $\alpha_j{=}0.05$.

Two properties of this design are essential. First, *shared acceptability is high*: on 88% of defer-region samples (0.10+0.78) at least two experts are correct, so the acceptable set $\mathcal{S}(y, m)$ typically contains the true class plus multiple experts. Second, *expert ranking matters*: the Bayes-optimal expert changes with the sector $Q$, and the gap between the best and non-best utilities is large (0.93 versus 0.37).

**Bayes rule.** *On $D$* (60% of data): defer to the sector-best expert $Q$ ($\alpha_Q{=}0.93 \gg \eta_{\max}{=}0.10$). *On $D^c$* (40% of data): classify ($\eta_Y{=}0.998 \gg \alpha_j{=}0.05$).

Bayes-optimal system accuracy: $0.60 \times 0.93 + 0.40 \times 0.998 = 0.957$; Bayes-optimal coverage: $0.40$.

**Why Mao25 fails at ranking.** Proposition 6 shows that the Mao25 gradient on an acceptable action $i \in \mathcal{S}$ is $q_i(S_{\mathcal{S}}-1)$, where $S_{\mathcal{S}} = \sum_{c\in\mathcal{S}} q_c$ is the total softmax mass on the acceptable set. The factor $(S_{\mathcal{S}}-1)$ is the same for every acceptable action: the best expert $Q$ and a non-best expert $j{\neq}Q$ receive the same gradient magnitude (up to the softmax weight $q_i$, which is controlled by the model, not the data). This means the samplewise loss provides no direct signal that $Q$ should be preferred over other acceptable experts.

Across samples, the model can still learn the ranking indirectly because expert $Q$ appears in the acceptable set much more often than the other experts on sector $Q$. However, this signal comes from marginal frequency across samples rather than from an explicit within-sample preference among acceptable experts. The result is degraded but non-trivial ranking ability, as measured by the best-expert selection rate.

**Metrics.** We report *best-expert selection on $D$*: among test samples in $D$ where the model defers, the fraction that selects the sector-best expert $Q$. With $J{=}4$ symmetric sectors, chance level (uniform random among experts) is $0.25$. A model that picks uniformly among *correct* experts would achieve $\approx 0.48$ ($0.10 \times \frac{1}{4} + 0.78 \times \frac{1}{2} + 0.05 \times 1 + 0.03 \times 0 + 0.04 \times \frac{1}{4}$). Perfect ranking gives $1.0$.

**Result.** Table 7 reports the retained result (means with std in parentheses over 3 seeds).

| Method | Defer regret | System acc. | Coverage | Best expert sel. on $D$ |
|---|---|---|---|---|
| **Decoupled** | **0.0008** (0.001) | **0.957** (0.003) | 0.397 | 1.000 |
| A-SM | 0.064 (0.020) | 0.894 (0.018) | 0.330 | 1.000 |
| OvA | 0.082 (0.076) | 0.876 (0.078) | 0.349 | 1.000 |
| PiCCE | 0.129 (0.045) | 0.829 (0.043) | 0.383 | 0.661 |
| Add. CE | 0.151 (0.033) | 0.808 (0.031) | 0.241 | 1.000 |
| Mao25 | 0.154 (0.043) | 0.804 (0.045) | 0.324 | 0.750 |

*Table 7.* Pair-shared ranking suite. Chance-level best-expert selection is $0.25$; uniform-among-correct is $0.48$. Mao25 reaches $0.750$, well above chance but far below the $1.000$ achieved by the decoupled surrogate and three other baselines, confirming that the loss provides partial but degraded ranking signal.

**Analysis.** The decoupled surrogate reaches the Bayes-optimal system accuracy ($0.957$) with coverage $0.397 \approx 0.40$ and perfect best-expert selection. Its independent BCE heads estimate each $\alpha_j$ faithfully, so the routing comparison $\max_j \sigma(s_j) \gtrless \max_k p_k$ is well-calibrated on both the classify/defer boundary and the expert ranking.

*Mao25.* Mao25 achieves best-expert selection of $0.750$. This is well above chance ($0.25$) and above the uniform-among-correct baseline ($0.48$), indicating that the linear model does extract partial ranking from the sector feature. However, it is significantly below the $1.000$ achieved by the decoupled surrogate, A-SM, OvA, and Add. CE, all of which have loss terms that produce per-expert ranking signal. The gap from $0.750$ to $1.000$ is the empirical cost of the $(S_{\mathcal{S}}-1)$ factor being common to all acceptable actions (Proposition 6). Combined with imperfect coverage ($0.324$ vs optimal $0.40$), this yields a defer regret of $0.154$ — nearly $200\times$ the decoupled surrogate regret.

*PiCCE.* PiCCE has even worse ranking ($0.661$) with high variance across seeds (one seed drops to $0.50$), consistent with the winner-take-all lock-in from Proposition 5: which expert "wins" depends on initialization noise, producing unstable ranking.

*A-SM and OvA.* Both achieve perfect ranking ($1.000$) and reasonable coverage, but their defer regret ($0.064$ and $0.082$) is still 80–100$\times$ the decoupled surrogate value. The ranking is correct, but the classify-versus-defer boundary is less well-calibrated: A-SM coverage is $0.330$ (optimal: $0.40$), indicating over-deferral. For A-SM this is consistent with the class–expert coupling (Proposition 7) perturbing the class logits.

*Add. CE.* Add. CE also achieves perfect ranking but has the lowest coverage ($0.241$), reflecting severe over-deferral from the gradient amplification of Proposition 4: with $J{=}4$ experts and shared acceptability on $88\%$ of defer samples, $|\mathcal{J}|$ is typically 2–4, inflating the gradient by factors of 3–5.

### K.4. Logit-Space Geometry Diagnostics

The task-level suites above measure end-to-end performance. This section complements them with *direct measurements of the differential quantities* predicted by Propositions 4, 5, and 7.

**Protocol.** For each diagnostic we train a linear model (no hidden layer) to convergence on a synthetic L2D task, then compute per-sample gradients and Hessians of the surrogate loss on held-out test logits. Using trained (rather than random) logits ensures the measurements reflect the geometry at the operating point the optimizer actually reaches. All quantities are computed on the *augmented* action vector $a = (a_1, \ldots, a_K, a_{K+1}, \ldots, a_{K+J})$, i.e. on the logit space of the surrogate itself. Each diagnostic is repeated across multiple seeds (stated per table); we report means with standard deviations in parentheses across all per-sample measurements pooled over seeds.

**Add. CE curvature inflation (Proposition 4).** Proposition 4 predicts that the Hessian of the additive CE surrogate $\Phi^{\mathrm{CE}}$ has largest eigenvalue $\lambda_{\max} \leq (1 + |\mathcal{J}|)/2$, growing linearly in the number of correct experts $|\mathcal{J}|$. The decoupled surrogate, by contrast, has $\lambda_{\max} \leq 1/2$ for the class block and $\lambda/(4J)$ per expert entry, independently of $|\mathcal{J}|$ ((12)).

*Setup.* We use the nested redundant-expert construction (Section K.1) with $J \in \{1, 4, 8, 16, 32\}$ to control the overlap multiplicity $|\mathcal{J}|$. For each value of $J$, we train an Add. CE and a decoupled surrogate model ($\lambda = 1.0$) on $n_{\text{train}}$=3,500 samples and evaluate on $n_{\text{test}}$=6,000. We restrict to test samples where at least one expert is correct ($|\mathcal{J}| > 0$). For each such sample, we compute (i) the gradient norm $\|\nabla_a \Phi\|_2$ and (ii) the top Hessian eigenvalue $\lambda_{\max}(\nabla_a^2 \Phi)$ via the full $(K{+}J) \times (K{+}J)$. Results are pooled over ∼900 qualifying samples per ($J$, method) combination across 5 independent seeds.

| Method | $\lvert\mathcal{J}\rvert$ | Gradient norm $\lVert\nabla_a \Phi\rVert_2$ | Top eigenvalue $\lambda_{\max}(\nabla_a^2 \Phi)$ |
|---|---|---|---|
| Add. CE | 1 | 0.586 (0.305) | 0.821 (0.084) |
| Add. CE | 4 | 0.609 (0.266) | 1.093 (0.067) |
| Add. CE | 8 | 0.616 (0.292) | 1.101 (0.116) |
| Add. CE | 16 | 0.633 (0.287) | 1.125 (0.186) |
| Add. CE | 32 | 0.613 (0.282) | 1.117 (0.181) |
| Decoupled | 1 | 0.704 (0.222) | 0.368 (0.111) |
| Decoupled | 4 | 0.573 (0.250) | 0.358 (0.137) |
| Decoupled | 8 | 0.551 (0.280) | 0.361 (0.125) |
| Decoupled | 16 | 0.534 (0.290) | 0.360 (0.134) |
| Decoupled | 32 | 0.510 (0.297) | 0.359 (0.132) |

*Table 8.* Per-sample curvature as overlap multiplicity $|\mathcal{J}|$ varies (5 seeds, ∼900 samples each). The Add. CE top eigenvalue grows from 0.82 at $|\mathcal{J}|$=1 to 1.12 at $|\mathcal{J}|$=32, consistent with the multiplicative $(1{+}|\mathcal{J}|)$ factor in Proposition 4, while also reflecting the dependence of the softmax covariance on $q$. The decoupled surrogate remains flat at ≈0.36 for all $|\mathcal{J}|$, consistent with the constant $\leq 1/2$ bound from (12).

*Reading the table.* The CE eigenvalue jumps from 0.82 to 1.09 as $|\mathcal{J}|$ goes from 1 to 4, then saturates near 1.12 for larger $|\mathcal{J}|$. The saturation is expected: at $|\mathcal{J}|$=32 the theoretical bound is $(1 + 32)/2 = 16.5$, but the *realised* eigenvalue is much smaller because the softmax probabilities $q_i$ become very small when spread over $K{+}32$ actions. The key observation is the sharp increase and high-overlap saturation of $\lambda_{\max}$ for Add. CE, versus the *complete independence* from $|\mathcal{J}|$ for the decoupled surrogate.

**PiCCE starvation gradient (Proposition 5).** Proposition 5 predicts that under PiCCE, every correct expert other than the winner $j^\star$ receives a strictly positive gradient $\partial\Phi^{\mathrm{PiCCE}}/\partial a_{K+k} = 2q_{K+k} > 0$, pushing its logit *downward* despite being correct. Under the decoupled surrogate, each expert gradient is $(\lambda/J)(u_j - t_j)$, which is *negative* (pulling toward deferral) whenever the expert is correct ($t_j = 1$) and its sigmoid output $u_j < 1$.

*Setup.* We use the PiCCE specialist task (Section K.2) with $K$=2 classes and $J$=2 experts, where samples routinely have both experts correct. For each method (PiCCE and the decoupled surrogate, $\lambda$=1.0) we train on $n_{\text{train}}$=4,500 and evaluate on $n_{\text{test}}$=7,000. We select test samples where $|\mathcal{J}| > 1$ (multiple correct experts). For PiCCE, the "suppressed expert" is any $k \in \mathcal{J}$ with $k \neq j^\star$, where $j^\star = \arg\max_{j \in \mathcal{J}} a_{K+j}$ is the winning expert (determined by the trained logit magnitudes). For the decoupled surrogate, we measure the gradient on *all* correct expert logits (since the decoupled surrogate has no winner selection). Results pool 784 decoupled-surrogate gradient samples and 392 PiCCE gradient samples across 5 seeds.

*Reading the table.* The positive-gradient rate is 1.000 for PiCCE and 0.000 for the decoupled surrogate, with zero variance:

| Method | Mean gradient $\partial\Phi/\partial a_{K+k}$ | Positive-gradient rate |
|---|---|---|
| Decoupled | $-0.245$ (0.103) | 0.000 |
| PiCCE | $+0.240$ (0.076) | 1.000 |

*Table 9.* Gradient on suppressed correct expert logits (5 seeds; 784 decoupled-surrogate and 392 PiCCE measurements). PiCCE produces a *positive* gradient on 100% of suppressed experts — pushing them away from deferral, exactly as Proposition 5 predicts ($2q_{K+k} > 0$). The decoupled surrogate always produces a *negative* gradient, pulling every correct expert toward deferral.

the predicted sign is not a statistical tendency but an exact per-sample property. The mean PiCCE gradient on the suppressed expert ($+0.240$) is comparable in magnitude to the decoupled surrogate gradient ($-0.245$), so the starvation is not a negligible effect — it is an active repulsive force of the same scale as the decoupled surrogate's attractive signal.

**A-SM class–expert coupling (Proposition 7).** Proposition 7 shows that the A-SM gradient on each class logit $a_r$ ($r \neq k^\star$) includes a leakage term $-\pi_r \sum_j (\psi_j - t_j)$ from the expert estimates. At second order, this coupling produces a non-zero mixed class–expert Hessian block $H_{\text{mix}} = \partial^2 \Phi^{\text{ASM}} / (\partial a_{1:K}\, \partial a_{K+1:K+J})$. The decoupled surrogate's Hessian is block-diagonal ((12)), so its mixed block is *identically* zero.

*Setup.* We vary $J \in \{1, 3, 5, 9, 17\}$ by adding distractor experts (always incorrect) alongside one correct expert. For each $J$, we train an A-SM and a decoupled surrogate model ($\lambda$=1.0) on $n_{\text{train}}$=4,500 and evaluate on $n_{\text{test}}$=7,000. For each test sample we assemble the full $(K+J) \times (K+J)$ Hessian via second-order `autograd`, extract the $K \times J$ upper-right block $H_{\text{mix}}$, and compute its operator norm $\|H_{\text{mix}}\|_{\text{op}}$. We report the mean operator norm over 128 test samples, averaged across 5 seeds. For the decoupled surrogate, the mixed block is analytically zero by construction (class logits $w$ and expert logits $s_j$ enter disjoint terms in $\Phi_\lambda^{\text{dec}}$), so the code assigns 0.0 directly.

| Method | Experts $J$ | Mixed-block norm $\|H_{\text{mix}}\|_{\text{op}}$ |
|---|---|---|
| A-SM | 1 | 0.122 (0.004) |
| A-SM | 3 | 0.207 (0.011) |
| A-SM | 5 | 0.263 (0.009) |
| A-SM | 9 | 0.351 (0.009) |
| A-SM | 17 | 0.475 (0.012) |
| Decoupled | 1 | 0.000 (0.000) |
| Decoupled | 3 | 0.000 (0.000) |
| Decoupled | 5 | 0.000 (0.000) |
| Decoupled | 9 | 0.000 (0.000) |
| Decoupled | 17 | 0.000 (0.000) |

*Table 10.* Mixed class–expert Hessian block operator norm (5 seeds, 128 samples each). A-SM coupling grows monotonically from 0.12 at $J$=1 to 0.48 at $J$=17, consistent with the $O(\sqrt{J})$ scaling from Proposition 7. The decoupled surrogate is identically zero for all $J$, confirming the block-diagonal structure of (12).

*Reading the table.* The A-SM coupling norm grows roughly as $\sqrt{J}$: from 0.12 at $J$=1 to 0.48 at $J$=17 $\approx 0.12\sqrt{17} \approx 0.49$. This scaling matches the theoretical $O(\sqrt{J})$ bound derived in Appendix G.4. The growth means that adding experts — even *distractor* experts that are never correct — injects progressively more expert-side noise into the class gradient. The decoupled surrogate is immune: its zero mixed block guarantees that expert estimation errors cannot perturb the class probability estimates, regardless of $J$.

## L. CIFAR-10 with Synthetic Experts

The synthetic experiments of Section K use linear models on low-dimensional Gaussian data, which isolates surrogate pathologies in a setting where the Bayes solution is known. We now test whether the same effects persist on a real vision task.

**Setup.** We train a ResNet-12 backbone (end-to-end, no pre-training) on CIFAR-10 ($K$=10 classes) with standard data augmentation (random crop, horizontal flip). All methods share the same backbone architecture and optimiser (AdamW, lr $10^{-3}$, 80 epochs). A held-out validation split (5 000 samples) is used to select the best checkpoint (lowest validation deferral loss) and, for the decoupled surrogate, to choose the per-expert weight $\lambda/J = 1/2$. For the decoupled surrogate, the linear head is a *split head*: two independent projections $f_{\text{cls}} \colon \mathbb{R}^{256} \to \mathbb{R}^K$ and $f_{\text{exp}} \colon \mathbb{R}^{256} \to \mathbb{R}^J$, matching the decoupled loss structure. All other methods use a single $\mathbb{R}^{256} \to \mathbb{R}^{K+J}$ head. The total number of parameters is the same across all methods.

**Redundant expert suite.** We construct $J$ synthetic experts with *nested* correctness (Section K.1): a single latent $U \sim \mathrm{Uniform}(0,1)$ per sample determines all expert outcomes, with expert $j$ correct iff $U < \alpha_j$ where $\alpha_j = 0.70 - 0.01\,j$. This ensures that the first 8 experts in a $J{=}32$ run are *identical* to those in a $J{=}8$ run, so performance changes are attributable purely to the surrogate's response to added experts.

*Table 11.* CIFAR-10 redundant expert suite (mean±std over 4 seeds). Best values per $J$-block in **bold**; second-best underlined. "Cls. Acc." = classifier accuracy (ignoring the routing decision). "Cov." = coverage (fraction classified by the model, not deferred).

| | | Validation Metrics | | | |
|---|---|---|---|---|---|
| Method | $J$ | Sys. Acc. ↑ | Defer Loss ↓ | Cls. Acc. | Cov. |
| **Decoupled** | 8 | $\mathbf{0.918}_{\pm.001}$ | $\mathbf{0.082}_{\pm.001}$ | $\underline{0.909}_{\pm.001}$ | $0.961_{\pm.004}$ |
| Classifier only | 8 | $\underline{0.910}_{\pm.002}$ | $\underline{0.090}_{\pm.002}$ | $\mathbf{0.910}_{\pm.002}$ | $1.000_{\pm.000}$ |
| OvA | 8 | $0.901_{\pm.001}$ | $0.099_{\pm.001}$ | $0.875_{\pm.001}$ | $0.857_{\pm.015}$ |
| PiCCE | 8 | $0.858_{\pm.007}$ | $0.142_{\pm.007}$ | $0.888_{\pm.008}$ | $0.616_{\pm.032}$ |
| A-SM | 8 | $0.860_{\pm.002}$ | $0.140_{\pm.002}$ | $0.831_{\pm.002}$ | $0.774_{\pm.014}$ |
| Add. CE | 8 | $0.815_{\pm.002}$ | $0.185_{\pm.002}$ | $0.902_{\pm.004}$ | $0.529_{\pm.006}$ |
| Mao25 | 8 | $0.703_{\pm.000}$ | $0.297_{\pm.000}$ | $0.105_{\pm.024}$ | $0.000_{\pm.000}$ |
| **Decoupled** | 16 | $\mathbf{0.919}_{\pm.003}$ | $\mathbf{0.081}_{\pm.003}$ | $\underline{0.909}_{\pm.002}$ | $0.961_{\pm.004}$ |
| Classifier only | 16 | $\underline{0.911}_{\pm.001}$ | $\underline{0.089}_{\pm.001}$ | $\mathbf{0.911}_{\pm.001}$ | $1.000_{\pm.000}$ |
| OvA | 16 | $0.890_{\pm.002}$ | $0.110_{\pm.002}$ | $0.859_{\pm.004}$ | $0.829_{\pm.014}$ |
| PiCCE | 16 | $0.860_{\pm.011}$ | $0.140_{\pm.011}$ | $0.895_{\pm.006}$ | $0.633_{\pm.064}$ |
| A-SM | 16 | $0.827_{\pm.003}$ | $0.174_{\pm.003}$ | $0.772_{\pm.003}$ | $0.662_{\pm.035}$ |
| Add. CE | 16 | $0.790_{\pm.013}$ | $0.210_{\pm.013}$ | $0.895_{\pm.003}$ | $0.502_{\pm.035}$ |
| Mao25 | 16 | $0.703_{\pm.000}$ | $0.297_{\pm.000}$ | $0.111_{\pm.010}$ | $0.000_{\pm.000}$ |
| **Decoupled** | 32 | $\mathbf{0.919}_{\pm.002}$ | $\mathbf{0.081}_{\pm.002}$ | $\underline{0.908}_{\pm.003}$ | $0.957_{\pm.005}$ |
| Classifier only | 32 | $\underline{0.908}_{\pm.002}$ | $\underline{0.092}_{\pm.002}$ | $\mathbf{0.908}_{\pm.002}$ | $1.000_{\pm.000}$ |
| OvA | 32 | $0.880_{\pm.003}$ | $0.120_{\pm.003}$ | $0.844_{\pm.003}$ | $0.798_{\pm.015}$ |
| PiCCE | 32 | $0.859_{\pm.008}$ | $0.141_{\pm.008}$ | $0.880_{\pm.014}$ | $0.636_{\pm.027}$ |
| A-SM | 32 | $0.783_{\pm.006}$ | $0.218_{\pm.006}$ | $0.680_{\pm.009}$ | $0.484_{\pm.039}$ |
| Add. CE | 32 | $0.712_{\pm.015}$ | $0.288_{\pm.015}$ | $0.872_{\pm.009}$ | $0.305_{\pm.022}$ |
| Mao25 | 32 | $0.703_{\pm.000}$ | $0.297_{\pm.000}$ | $0.105_{\pm.025}$ | $0.000_{\pm.000}$ |

**Analysis.** *The decoupled surrogate is the only method in our comparison that improves on a standalone classifier.* At every $J$, the decoupled surrogate achieves higher system accuracy than the classifier-only baseline: $+0.8$ pp at $J{=}8$, $+0.8$ pp at $J{=}16$, and $+1.1$ pp at $J{=}32$. Every other L2D surrogate *degrades* below the classifier-only baseline, confirming that the surrogate pathologies identified in the synthetic experiments (Sections K.1–K.3) are not artefacts of the linear setting.

*The decoupled surrogate preserves classifier quality; other surrogates corrupt it.* The decoupled surrogate maintains classifier accuracy at 90.8–90.9% across all $J$, matching the classifier-only baseline. By contrast, A-SM drops from 83.1% ($J{=}8$) to 68.0% ($J{=}32$)—a 15 pp collapse consistent with expert gradients contaminating the class head through the shared softmax normalizer (Proposition 7). OvA also degrades (87.5% → 84.4%). This supports the interpretation that the decoupled surrogate's decoupled loss protects the classifier: the BCE expert loss trains only the expert head, leaving the class head free to learn optimal class probabilities.

*For most baselines, deferral actively hurts.* When system accuracy falls *below* classifier accuracy, routing decisions are destructive. At $J{=}8$, PiCCE achieves 88.8% classifier accuracy but only 85.8% system accuracy ($-3.0$ pp from deferral); Add. CE similarly loses 8.7 pp. Both route a substantial share of samples (PiCCE 38.4%, Add. CE 47.1%) to experts with ≈70% accuracy—far worse than the classifier.

*Add. CE and A-SM degrade monotonically with $J$.* Add. CE system accuracy drops from 81.5% ($J{=}8$) to 71.2% ($J{=}32$), matching the gradient amplification of Proposition 4: the $(1{+}|\mathcal{J}|)$ factor grows with the redundant pool, pushing coverage down to 30.5%. A-SM shows a similar pattern (86.0% → 78.3%), driven by the class–expert coupling of Proposition 7.

*Mao25 fails to learn.* Across all $J$, Mao25 converges to 100% deferral within the first few epochs, achieving exactly the expert oracle accuracy (≈ 70.3%). The $O(1/(K{+}J))$ gradient magnitude through the softmax (Section F.3) prevents meaningful parameter updates on CIFAR-10's $K{=}10$ classes.

*The decoupled surrogate is $J$-robust.* The decoupled surrogate system accuracy stays at 91.8–91.9% from $J{=}8$ to $J{=}32$ (a change of only 0.1 pp), compared to 2.1 pp for OvA and 10.3 pp for Add. CE. This is consistent with the theoretical prediction: the decoupled surrogate's $J$-independent gradient structure (Proposition 9) translates directly to scalability on

real image features.

## M. CIFAR-10H with Real Human Annotators

The synthetic-expert experiments above let us control the expert accuracy profile exactly, but they leave open the question of whether the same surrogate pathologies appear when the annotators are real humans with correlated, noisy labels. We therefore evaluate all methods on CIFAR-10H (Peterson et al., 2019), a dataset that pairs every CIFAR-10 test image with soft human annotations collected from $\approx$2,700 Amazon Mechanical Turk workers.

**Setup.** We first pretrain a ResNet-12 backbone on the full CIFAR-10 training set (50 000 images, 80 epochs, SGD, lr 0.05, momentum 0.9, weight decay $5\times10^{-4}$) and then *freeze* it. Each L2D method trains only a linear head on top of the frozen 256-dimensional features. We split the CIFAR-10 test set (10 000 images annotated by CIFAR-10H) into 6 000 train / 2 000 val / 2 000 test. All methods share the same pretrained backbone and data split.

**Annotator selection.** From the CIFAR-10H pool we retain only annotators with at least 200 labelled images. For each run we subsample $J \in \{5, 10, 20\}$ annotators uniformly at random; for images not labelled by a selected annotator, we impute a smoothed class-prior prediction (Laplace smoothing, $\alpha$=1). We repeat each configuration over 5 independent annotator draws (seeds 0–4) and report mean with std in parentheses across draws. All methods use AdamW (lr $10^{-3}$, 30 epochs). For the decoupled surrogate, $\lambda$=2.0 and weight decay $10^{-3}$; all other methods use weight decay $10^{-4}$.

*Table 12.* CIFAR-10H with real human annotators (test metrics, mean with std over 5 annotator draws). Best values per $J$-block in **bold**; second-best underlined. "Cls. Acc." = classifier accuracy (ignoring the routing decision). "Cov." = coverage (fraction classified by the model, not deferred).

| | | Test Metrics | | | |
|---|---|---|---|---|---|
| Method | $J$ | Sys. Acc. ↑ | Defer Loss ↓ | Cls. Acc. | Cov. |
| **Decoupled** | 5 | **0.960** (0.005) | **0.039** (0.005) | **0.892** (0.001) | 0.517 (0.036) |
| OvA | 5 | 0.959 (0.005) | 0.039 (0.004) | 0.889 (0.001) | 0.375 (0.016) |
| A-SM | 5 | 0.959 (0.004) | 0.040 (0.004) | 0.890 (0.002) | 0.341 (0.035) |
| PiCCE | 5 | 0.954 (0.003) | 0.045 (0.003) | 0.889 (0.001) | 0.390 (0.020) |
| Add. CE | 5 | 0.953 (0.004) | 0.045 (0.004) | 0.889 (0.002) | 0.248 (0.037) |
| Mao25 | 5 | 0.950 (0.005) | 0.047 (0.005) | 0.116 (0.133) | 0.000 (0.000) |
| Classifier only | 5 | 0.889 (0.000) | 0.109 (0.001) | 0.889 (0.000) | 1.000 (0.000) |
| **Decoupled** | 10 | **0.958** (0.003) | **0.042** (0.004) | **0.892** (0.003) | 0.517 (0.038) |
| PiCCE | 10 | 0.957 (0.004) | 0.043 (0.004) | 0.890 (0.002) | 0.400 (0.006) |
| OvA | 10 | 0.954 (0.008) | 0.046 (0.008) | 0.799 (0.203) | 0.285 (0.164) |
| A-SM | 10 | 0.952 (0.007) | 0.047 (0.007) | 0.747 (0.313) | 0.170 (0.134) |
| Add. CE | 10 | 0.952 (0.004) | 0.047 (0.004) | 0.888 (0.001) | 0.248 (0.004) |
| Mao25 | 10 | 0.948 (0.003) | 0.052 (0.003) | 0.077 (0.024) | 0.000 (0.000) |
| Classifier only | 10 | 0.889 (0.000) | 0.111 (0.000) | 0.889 (0.000) | 1.000 (0.000) |
| **Decoupled** | 20 | **0.961** (0.004) | **0.039** (0.004) | 0.889 (0.001) | 0.530 (0.038) |
| OvA | 20 | 0.956 (0.006) | 0.043 (0.006) | **0.890** (0.002) | 0.335 (0.047) |
| PiCCE | 20 | 0.953 (0.002) | 0.047 (0.002) | 0.890 (0.001) | 0.399 (0.022) |
| Add. CE | 20 | 0.951 (0.003) | 0.049 (0.003) | 0.742 (0.331) | 0.183 (0.103) |
| A-SM | 20 | 0.949 (0.005) | 0.051 (0.005) | 0.470 (0.435) | 0.002 (0.004) |
| Mao25 | 20 | 0.948 (0.004) | 0.051 (0.004) | 0.045 (0.014) | 0.000 (0.000) |
| Classifier only | 20 | 0.888 (0.000) | 0.112 (0.000) | 0.888 (0.000) | 1.000 (0.000) |

**Analysis.** *The decoupled surrogate leads at every pool size.* At $J$=5 the margin is narrow (0.960 vs. 0.959 for OvA), but the decoupled surrogate's advantage widens as the pool grows: at $J$=20 the decoupled surrogate reaches 0.961 while OvA drops to 0.956 (+0.5 pp). This mirrors the pattern from synthetic experts (Table 11): the decoupled surrogate scales better because its per-expert BCE terms are unaffected by pool size.

*The decoupled surrogate preserves classifier quality; baselines collapse.* The decoupled surrogate's classifier accuracy stays at 0.889–0.892 across all $J$, matching the standalone classifier baseline (0.889). A-SM maintains 0.890 at $J$=5 but collapses to 0.470 at $J$=20 (std 0.435, indicating complete instability); Add. CE similarly drops from 0.889 to 0.742; OvA degrades to 0.799 (std 0.203) at $J$=10. All three methods compensate by over-deferring—A-SM reaches $\geq$99% deferral at $J$=20—so their system accuracy is sustained by expert answers, not by the classifier. This confirms on real

human annotators the same softmax-coupling (Proposition 7) and gradient-amplification (Proposition 4) pathologies seen in synthetic settings.

*The decoupled surrogate maintains healthy coverage.* The decoupled surrogate's coverage is stable at $\approx 0.52$, meaning it classifies about half the samples itself and defers the rest—a balanced routing that produces system accuracy well above either component alone.

*Mao25 defers everything.* As in the synthetic setting, Mao25 converges to $100\%$ deferral and never learns a classifier (classifier accuracy $\leq 0.116$). Its system accuracy ($\approx 0.950$) is entirely the expert oracle accuracy, confirming the vanishing-gradient pathology of Proposition 6 on real human-annotator data.

## N. Covertype with Model Experts

The preceding experiments use either synthetic accuracy profiles (Section L) or real human annotators (Section M). This final benchmark replaces both with *real pretrained models*: the experts are machine-learning classifiers with different architectures, feature views, and training biases. This is the most common deployment scenario for learning to defer—a system that routes each input to the most reliable model in a pool.

**Setup.** We use Forest CoverType (Blackard & Dean, 1999), a 7-class tabular benchmark with $54$ features (continuous topographic variables, binary wilderness indicators, and binary soil-type indicators). A single stratified $60/20/20$ train/validation/test split is made once; only the continuous features are standardized using train-set statistics. The L2D backbone is a shared two-layer MLP ($54 \to 256 \to 128$, ReLU, batch normalization, no dropout).

**Expert pool.** The pool contains $J{=}5$ pretrained models, trained once on the training split and then frozen:

(i) a global histogram-gradient-boosted tree on all features;

(ii) a low-elevation specialist (reweighted toward low elevations and wilderness areas 1–2);

(iii) a high-elevation specialist (reweighted toward high elevations and wilderness areas 3–4);

(iv) a topography-view MLP trained only on the continuous variables;

(v) a context-view random forest trained only on wilderness and soil indicators.

The expert oracle accuracy on the validation split is $0.946$, a $+9.8$ pp gain over the best single expert ($0.848$), confirming that routing is nontrivial. The data split and expert-pool training seed are held fixed; only the downstream L2D training seed changes across runs.

**Protocol.** All methods use AdamW with cosine learning-rate decay. All the approaches use a linear head, lr $10^{-3}$, weight decay $10^{-4}$, and $40$ epochs. We report results for per-expert weight $\lambda/J = 1/2$. Checkpoints are selected by the lowest validation true deferral loss.

*Table 13.* Covertype with model experts ($J{=}5$, mean$\pm$std over 4 seeds). Best values in **bold**; second-best underlined. "Cls. Acc." = classifier accuracy (ignoring routing). "Cov." = coverage (fraction classified, not deferred).

| Method | Sys. Acc. ↑ | Defer Loss ↓ | Cls. Acc. | Cov. |
|---|---|---|---|---|
| **Decoupled** | $\mathbf{0.934}_{\pm.000}$ | $\mathbf{0.066}_{\pm.000}$ | $\mathbf{0.941}_{\pm.001}$ | $0.712_{\pm.040}$ |
| Classifier only | $\underline{0.929}_{\pm.001}$ | $\underline{0.071}_{\pm.001}$ | $\underline{0.929}_{\pm.001}$ | $1.000_{\pm.000}$ |
| PiCCE | $0.908_{\pm.003}$ | $0.092_{\pm.003}$ | $0.911_{\pm.002}$ | $0.449_{\pm.007}$ |
| OvA | $0.907_{\pm.001}$ | $0.093_{\pm.001}$ | $0.906_{\pm.000}$ | $0.437_{\pm.011}$ |
| A-SM | $0.899_{\pm.001}$ | $0.101_{\pm.001}$ | $0.892_{\pm.001}$ | $0.439_{\pm.007}$ |
| Add. CE | $0.899_{\pm.001}$ | $0.101_{\pm.001}$ | $0.907_{\pm.001}$ | $0.098_{\pm.002}$ |
| Mao25 | $0.877_{\pm.001}$ | $0.123_{\pm.001}$ | $0.351_{\pm.023}$ | $0.000_{\pm.000}$ |

**Analysis.** *The decoupled surrogate is the only L2D method in our comparison that improves over the standalone classifier.* The decoupled surrogate reaches $0.934$ system accuracy, $+0.5$ pp above the classifier-only baseline ($0.929$) and $+2.6$ pp above the best baseline (PiCCE at $0.908$). Every other L2D surrogate degrades below the standalone classifier, confirming that their routing decisions are harmful.

*The decoupled surrogate preserves and even improves classifier quality.* The decoupled surrogate achieves the highest classifier accuracy ($0.941$), surpassing even the standalone classifier ($0.929$), showing that the expert BCE loss does not interfere with class learning. By contrast, Mao25's classifier collapses to $0.351$ and A-SM degrades to $0.892$.

*The decoupled surrogate maintains the highest coverage among L2D methods.* The decoupled surrogate classifies $71.2\%$ of samples directly, far more than PiCCE ($44.9\%$), A-SM ($43.9\%$), or Add. CE ($9.8\%$). The baselines over-defer because their class heads are too degraded to be confidently preferred over an expert.

*Mao25 collapses to full deferral.* As in CIFAR-10 and CIFAR-10H, Mao25 converges to $100\%$ deferral (zero coverage) and never learns a useful classifier, matching the vanishing-gradient prediction of Proposition 6.

*All surrogate pathologies transfer to model experts.* The method ranking observed in synthetic and human-annotator experiments holds on tabular data with real model experts: The decoupled surrogate $>$ PiCCE $\approx$ OvA $>$ A-SM $\approx$ Add. CE $>$ Mao25. This demonstrates that the decoupled surrogate's advantages are not specific to vision backbones or to any particular type of expert.