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

Yannis Montreuil
School of Computing
National University of Singapore
Singapore, 117418
yannis.montreuil@u.nus.edu

Axel Carlier
Fédération ENAC ISAE-SUPAERO ONERA
Université de Toulouse
Toulouse, 31555, France

Lai Xing Ng
Agency for Science, Technology and Research
Institute for Infocomm Research
Singapore, 138634

Wei Tsang Ooi
School of Computing
National University of Singapore
Singapore, 117418

Abstract

A learning-to-defer (L2D) system decides, for each input, whether to predict on its own or defer to one of several experts. A widely used recipe trains the classifier and the router jointly by treating the K classes and J experts as competing actions in one shared $(K+J)$-action geometry. We analyze the population targets and optimization geometry of five representative surrogates and identify target distortion, gradient amplification, winner-take-all starvation, set-mass collapse, and class–expert coupling, which can persist alongside statistical consistency; shared competition and local overlap amplification also persist beyond log-softmax. We depart from the augmented-action formulation with a *decoupled surrogate*, which pairs a softmax classifier head with an independent sigmoid head for each expert. Its population minimizer recovers the class posterior and expert utilities exactly, its coordinatewise expert updates avoid overlap amplification and starvation, and its mixed head Hessian block is zero even with a shared representation. We prove an excess-risk bound with calibration constant $\max\{\sqrt{2}, \sqrt{2J/\lambda}\}$, which does not depend on the number of experts at fixed per-expert weight. In synthetic studies the decoupled surrogate preserves rare specialists, on CIFAR-10, CIFAR-10H, Covertype, and Fashion-MNIST it preserves classifier quality, and Fashion-MNIST ablations favor it over additive cross-entropy in all 36 matched comparisons across head layouts and trunk choices. With disjoint networks, the same objective also supports two-stage training.

1 Introduction

Learning-to-Defer (L2D) augments a classifier with the option of handing an input to an expert (Madras et al., 2018; Mozannar and Sontag, 2020). In the multi-expert setting, this means deciding, for each example, whether to predict directly or defer to one of several experts. The Bayes rule compares the class posterior $\eta_k(x) = \Pr(Y=k \mid X=x)$ with the expert utilities $\alpha_j(x) = \Pr(M_j = Y \mid X=x)$, predicting when the best class posterior exceeds the best expert utility and deferring otherwise. A surrogate should learn both quantities and support this comparison from finite data.

Most existing multi-expert surrogates introduce an augmented action space containing the K class actions and J defer-to-expert actions. They learn a single score vector $a(x) \in \mathbb{R}^{K+J}$ and derive class and expert supervision from these shared scores (Mozannar and Sontag, 2020; Verma et al., 2022; Cao et al., 2024; Mao et al., 2025b; Liu et al., 2026). This route has delivered statistically

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

consistent surrogates, but it places a categorical posterior and expert utilities 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). Consistency secures the Bayes decision, but it does not determine whether expert overlap amplifies updates, whether correct non-winning experts are suppressed, or whether class and expert scores remain coupled; these local effects follow from the loss itself and can change as the expert pool grows. We therefore study each surrogate along two axes: what it identifies at its population optimum, and how it distributes supervision during training.

Our analysis begins with the general additive reduction of Mao et al. (2024d), whose comp-sum instances retain shared competition and local overlap amplification beyond log-softmax (Proposition 4). The log-loss instance has the exact, score-independent factor $1 + |\mathcal{J}(y, m)|$ in its gradient and Hessian, where $\mathcal{J}(y, m)$ is the realized set of correct experts (Corollary 4.1), and its conditional target is $(\eta, \alpha)/(1 + \sum_j \alpha_j)$ (Proposition 14). This common normalization preserves the Bayes ordering but changes the raw probability representation and the relative sample weights.

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. (2025b) 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).

We propose a *decoupled surrogate* that separates class and expert estimation: a softmax class head produces $p(x) \in \Delta^K$, and an independent sigmoid head for each expert produces $u_j(x) \in (0, 1)$. This parameterization matches the Bayes-sufficient pair (η, α) without a shared class–expert normalizer and without assuming that expert correctness events are independent. Every correct expert receives its own attractive gradient regardless of rival scores, and expert overlap does not rescale the classifier score gradient. Separate head parameters remain uncoupled even with a shared representation, and disjoint networks also permit two-stage training of the same objective.

Contributions.

- We analyze five representative augmented-action surrogates, with exact target and gradient/curvature statements showing how their action geometry causes target distortion, gradient amplification, winner-take-all starvation, set-mass collapse, class–expert coupling, or improper class geometry. The comp-sum analysis shows that competition persists beyond log-softmax (Section 4).
- We propose a decoupled surrogate that estimates the two Bayes objects on their native scales: a categorical class posterior and separately parameterized expert utilities. We prove exact recovery of (η, α) at the population optimum, coordinatewise expert gradients without overlap amplification or starvation, and uniformly bounded score curvature. Separate head-parameter blocks remain uncoupled with a learned shared trunk, and disjoint networks support two-stage training of the same objective (Section 5).
- We prove a quantitative consistency bound: decoupled surrogate excess risk controls L2D excess risk with calibration constant $\max\{\sqrt{2}, \sqrt{2J/\lambda}\}$, which is independent of J when the per-expert weight λ/J is fixed. We also establish a dimension-independent transfer for unit-weight OvA (Theorem 10, Proposition 20).
- Synthetic studies isolate the predicted pathologies, and four real-data benchmarks show that classifier quality is preserved and routing remains stable. Controlled Fashion-MNIST ablations favor the decoupled loss in all 36 paired seed–pool–architecture comparisons, including separate trunks. Weight, cost, and schedule studies test robustness and two-stage training (Section 6).

2 Related Work

Learning-to-defer extends selective prediction and classification with a reject option (Chow, 1970; Bartlett and Wegkamp, 2008; Geifman and 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 and 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., 2024d, 2025b; Montreuil et al., 2025b, 2026a,d). Our surrogate-regret analysis complements the restricted-class $\mathcal{H}$-consistency program; the distinction and further references are discussed in Appendix A.

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 and Sontag (2020) introduce the first consistent multiclass surrogate via an augmented action space that scores the class and deferral actions jointly, and subsequent work refines this shared-vector design. Verma et al. (2022) extend it to multiple experts, 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. (2025b) 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, and Liu et al. (2024) propose a training-time mitigation in the same setting. More recently, Liu et al. (2026) show that in the multi-expert setting even redundant experts can induce underfitting through the additive expert reward, and they propose a winner-take-all correction (PiCCE).

Two-stage and utility-based approaches. Two-stage L2D first learns a predictor and then fits a router (Narasimhan et al., 2022; Mao et al., 2023a, 2024f). These methods provide guarantees relative to the frozen predictor, including restricted-class and realizability results. Our disjoint formulation also supports two-stage training, while learning marginal utilities directly; Appendix J.4 compares the underlying objectives. Classification with rejection can likewise be reduced to cost-sensitive classification (Charoenphakdee et al., 2021), and dependence-aware L2D uses further information about human–AI complementarity (Wei et al., 2024). Relative to these lines of work, our contribution is a joint analysis of population targets and optimization geometry, followed by an objective that keeps the categorical class posterior and the expert utilities on their native scales.

3 Preliminaries

Multi-expert learning to defer. Following the modern learning-to-defer literature (Madras et al., 2018; Mozannar and Sontag, 2020; Mao et al., 2024d; Montreuil et al., 2026a), let $(\Omega, \mathcal{F}, \mathbb{P})$ 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 we consider (Mozannar and Sontag, 2020; Verma et al., 2022; Cao et al., 2024; Mao et al., 2024d; Liu et al., 2026), the only question is 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

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

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 best option pointwise in expectation.

Bayes quantities. The Bayes rule is governed by two families of conditional probabilities: the class posterior $\eta_k(x) \coloneqq \Pr(Y = k \mid X = x)$ for $k \in [K]$, and the 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, since 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*

$$r^\star(x) \in \begin{cases} \{0\}, & \text{if } \max_{k\in[K]} \eta_k(x) \geq \max_{j\in[J]} \alpha_j(x), \\ \arg\max_{j\in[J]} \alpha_j(x), & \text{otherwise,} \end{cases} \tag{2}$$

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

Lemma 2 exposes the statistical structure of the problem: the class posterior $\eta(x)$ is a categorical distribution and therefore lies in the simplex Δ^K, whereas the expert utilities $\alpha_1(x), \ldots, \alpha_J(x)$ 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 and Sontag (2020), casts the problem as prediction over one enlarged action space $\mathcal{A} := [K] \cup \{\perp_1, \ldots, \perp_J\}$ and learns one score per 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 and 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., 2025b). 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 analyze the augmented-action surrogates of Definition 3 along two axes: *(i)* the conditional target at the population optimum, and *(ii)* the local gradient and curvature geometry during training. Throughout, $\mathcal{J}(y, m) := \{j \in [J] : m_j = y\}$ denotes the set of correct experts and $q_i(x) := \exp(a_i(x))/\sum_{r=1}^{K+J} \exp(a_r(x))$. Detailed proofs, curvature calculations, and worked examples for the baseline analyses are collected in Appendices D–I.

4.1 General Additive Reductions: Competition and Overlap

The general additive reduction of Mao et al. (2024d) uses a multiclass base loss ℓ and score-independent costs $c_j \in [0, 1]$: $\Phi_\ell = \ell(a, y) + \sum_j (1 - c_j)\ell(a, K + j)$. In our zero-one setting, $c_j = \mathbf{1}\{m_j \neq y\}$, giving one term for the true class and one for each correct expert. These terms share the same augmented scores; the base loss determines how their gradients interact.

Shared competition across the comp-sum family. Consider the comp-sum base losses (Mao et al., 2023b):

$$\ell_\tau(a, i) = \Upsilon_\tau\left(\sum_{r\neq i} e^{a_r - a_i}\right) = \begin{cases} (q_i^{\tau-1} - 1)/(1 - \tau), & \tau \neq 1, \\ -\log q_i, & \tau = 1, \end{cases} \qquad \tau \geq 0.$$

This family includes sum-exponential ($\tau = 0$), log loss ($\tau = 1$), generalized CE ($1 < \tau < 2$), and MAE ($\tau = 2$).

Proposition 4 (Competition and overlap beyond log loss)**.** *Let $d = K + J$, $a \in \mathbb{R}^d$, $q = \text{softmax}(a)$, and let e_i be the standard basis vectors. Set $\mathcal{T} = \{y\} \cup \{K + j : j \in \mathcal{J}(y, m)\}$, $N = |\mathcal{T}|$, and $L_\tau = \sum_{i\in\mathcal{T}} \ell_\tau(a, i)$. Write $C(q) = \text{Diag}(q) - qq^\top$, and define*

$$\bar{q}_\tau = \frac{1}{N}\sum_{i\in\mathcal{T}} q_i^{\tau-1}, \qquad \omega_i = \frac{q_i^{\tau-1}}{N\bar{q}_\tau}, \qquad \bar{e}_\tau = \sum_{i\in\mathcal{T}} \omega_i e_i, \qquad \bar{R}_\tau = \sum_{i\in\mathcal{T}} \omega_i (q - e_i)(q - e_i)^\top.$$

Then, for every $\tau \geq 0$,

$$\nabla_a L_\tau = N\bar{q}_\tau(q - \bar{e}_\tau), \qquad \nabla_a^2 L_\tau = N\bar{q}_\tau[C(q) - (\tau - 1)\bar{R}_\tau]. \tag{3}$$

Every omitted action $k \notin \mathcal{T}$ has $\partial L_\tau / \partial a_k = N\bar{q}_\tau q_k > 0$. At uniform scores this derivative is $Nd^{-\tau}$, and every nonzero zero-sum direction supported on omitted actions is an eigenvector of the Hessian with eigenvalue $Nd^{-\tau}$. Relative to the true-class-only loss at the same scores and dimension, both quantities are amplified exactly by N.

Every added target changes all other coordinate gradients through q, and since the proof holds for any target set $\mathcal{T}$, a correct expert omitted by winner-take-all selection is pushed down throughout the family. The effective weight $N\bar{q}_\tau$ depends on the number of active targets and their current probabilities. For $\tau \leq 1$, the Hessian contributions are positive semidefinite; for $\tau > 1$, the correction $-(\tau - 1)\bar{R}_\tau$ can introduce negative curvature. At $\tau = 2$, $N\bar{q}_2 = \sum_{i \in \mathcal{T}} q_i$ and $L_2 = N - \sum_{i \in \mathcal{T}} q_i$, giving the set-mass geometry studied in Section 4.3. Weighted reductions and arbitrary increasing comp-sum outer functions are analyzed in Appendix J.3.

Log-softmax: global amplification and target distortion. For $\tau = 1$, $\bar{q}_1 = 1$ and the correction term vanishes, giving Mozannar's additive CE surrogate (Mozannar and Sontag, 2020):

$$\Phi^{\mathrm{CE}}(a; x, y, m) = -\log q_y(x) - \sum_{j \in \mathcal{J}(y,m)} \log q_{K+j}(x). \tag{4}$$

Corollary 4.1 (Mozannar's log-softmax instance)**.** *For score-independent weights $v_i \geq 0$, set $V = \sum_i v_i$ and $L_v = -\sum_i v_i \log q_i$. Then*

$$\frac{\partial L_v}{\partial a_i} = V q_i - v_i, \tag{5}$$

$$\nabla_a^2 L_v = V C(q), \qquad \lambda_{\max}(\nabla_a^2 L_v) \leq V/2. \tag{6}$$

For Add. CE (Mozannar and Sontag, 2020), $V = N = 1 + |\mathcal{J}|$.

The conditional log-loss optimum is $q^\star = (\eta, \alpha)/(1 + U)$, where $U(x) = \sum_j \alpha_j(x)$ (Appendix D.2). This target preserves the Bayes ordering but normalizes the class posterior and the expert utilities together, so duplicating an expert with $\alpha_j(x) > 0$ enlarges the denominator without changing the best action; the same normalization underlies the unbounded defer-probability reconstruction (Verma et al., 2022; Cao et al., 2024). Overlap therefore changes both the represented target and the samplewise update weight, consistent with redundant-expert underfitting (Liu et al., 2026).

4.2 Why Winner-Take-All Fails

For log-softmax, Section 4.1 identifies the count factor caused by rewarding every correct expert. PiCCE (Liu et al., 2026) caps this factor by rewarding only one correct expert per sample. We analyze its log-loss instance and show that the same selection creates the omitted-action suppression predicted by Proposition 4. 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

$$\Phi^{\mathrm{PiCCE}}(a;\, x, y, m) \;=\; -\log q_y(x) \;-\; \mathbf{1}\{\mathcal{J} \neq \varnothing\} \log q_{K+j^\star}(x). \tag{7}$$

With at most two log-loss terms active, the total positive gradient mass is capped at 2 however many experts are correct, and with a unique correct-expert winner the Hessian is $\nabla_a^2 \Phi^{\mathrm{PiCCE}} = 2(\mathrm{Diag}(q) - qq^\top)$, so $\lambda_{\max} \leq 1$ uniformly and the multiplicity factor of (6) disappears on axis (ii) (Appendix E.4).

Starvation of non-winning experts (axis (ii)).

Proposition 5 (Winner-take-all starvation in PiCCE)**.** *For $\mathcal{J}(y, m) \neq \varnothing$ and a unique winning correct expert, 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 $j' \in \mathcal{J}$ with $j' \neq j^\star$,*

$$\frac{\partial \Phi^{\mathrm{PiCCE}}}{\partial a_{K+j'}} \;=\; 2q_{K+j'}(x) \;>\; 0. \tag{8}$$

Gradient descent pushes expert j''s logit downward*, even though it is correct. At a tie, the same formula is the derivative of the selected branch.*

The positive term $2q_i$ is intended for wrong classes, but it also suppresses every correct expert $j' \in \mathcal{J} \setminus \{j^\star\}$ solely because a rival expert had a higher score.

Self-reinforcing lock-in and a moving target. 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 can turn a small initial score gap into persistent suppression. The same mechanism also distorts the statistical target on axis (i): PiCCE's target is proportional to $(\eta, \gamma(a, \cdot))$ rather than (η, α), 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, γ_j shifts mass toward surviving winners and entrenches them, so the target *moves* with the scores rather than being fixed by the data distribution.

4.3 Three Refined Surrogates, Three Residual Gaps

We now examine three alternatives to winner selection: an acceptable-set objective, a separate class normalizer, and independent binary losses.

Mao25: correct set, no ranking. The linear member of the family of Mao et al. (2025b) penalizes the total mass assigned outside the acceptable set, $\Phi^{\mathrm{Mao}} = 1 - S_{\mathcal{S}}$ with $S_{\mathcal{S}} = q_y + \sum_{j \in \mathcal{J}} q_{K+j}$. The loss has consistency guarantees (Mao et al., 2025b), but it is *linear* in the acceptable mass $S_{\mathcal{S}}$, which makes it a set-valued analogue of multiclass MAE (Ghosh et al., 2017).

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.*

Because the samplewise supervision cannot express preferences inside the acceptable set, Mao25 does not recover the expert utilities directly: it can identify a Bayes-optimal action, but it does not estimate the full vector (η, α) needed to compare experts at the utility level, a limitation on axis (i). The same MAE-like structure also weakens optimization: the total attractive gradient mass on acceptable actions is $\sum_{i \in \mathcal{S}} -\partial\Phi^{\mathrm{Mao}}/\partial a_i = S_{\mathcal{S}}(1 - S_{\mathcal{S}})$, which is at most $1/4$ and vanishes as $S_{\mathcal{S}} \to 0$. When the model assigns little mass to the acceptable set, its corrective signal is therefore weak, consistent with the optimization behavior reported for MAE-like losses (Zhang and 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. However, ψ_j still depends on the class logits, which couples 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*

$$\frac{\partial \Phi^{\mathrm{ASM}}}{\partial a_r} = \underbrace{(\xi_r - \mathbf{1}\{r = y\})}_{\textit{class term}} - \pi_r \underbrace{\sum_{j=1}^{J} \big(\psi_j - \mathbf{1}\{m_j = y\}\big)}_{\textit{expert leakage}}. \tag{9}$$

The leakage term vanishes in expectation at the optimum but can be of order J on individual samples, and the mixed Hessian block has operator norm at most $\sqrt{J}/4$ (Appendix G.4), a bound that grows with the expert pool. A-SM is thus the strongest augmented baseline on axis (i), and its residual weakness lies on axis (ii).

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

Synthesis. Table 1 summarizes these properties: shared normalization accounts for the pathologies of the coupled softmax surrogates, while OvA avoids it but does not impose the categorical constraint

Table 1: Structural properties of multi-expert L2D surrogates. ✓ = satisfied; ✗ = violated; – = no bound asserted here. Axis column: (I) statistical target, (II) local geometry, (TH) 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 (η, α)	(I)	✗ (14)	✗ (E.3)	✗ (6)	✓ (17)	✓ (19)	✓ (10)
Per-sample gradient and Hessian free of multiplicative $O(\lvert\mathcal{J}\rvert)$ factor	(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)	✓ (13)
Class head is a categorical posterior throughout training	(I)	✗ (3)	✗ (3)	✗ (3)	✓ (17)	✗ (19)	✓ (8)
Quantitative surrogate-to-target excess-risk bound	(TH)	✓†	–	✓‡	–	✓ (20)	✓ (10)

†Comp-sum reduction (Mao et al., 2024d), constant $\sqrt{2(J+1)}$. ‡Linear member of the single-stage family (Mao et al., 2025b): $(K+J)\Delta\mathcal{E}_{\mathrm{Mao}}$ (Appendix I).

on its class block. Two design principles follow, 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).

5 The Decoupled Surrogate

The decoupled surrogate implements the two principles from Section 4: separate class and expert scores, and match each head to its target so that the conditional minimizer recovers (η, α) directly.

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 σ is the logistic sigmoid. The class estimate $p(x) \in \Delta^K$ is a proper categorical distribution, while each expert estimate $u_j(x) \in (0,1)$ is parameterized separately, without assuming that the expert correctness events are independent. These types mirror the Bayes-sufficient pair $(\eta, \alpha) \in \Delta^K \times [0,1]^J$ exactly, and no shared simplex forces classes and experts to compete for probability mass. The corresponding bounded cost-sensitive formulation is given in Appendix C; the main text restricts attention to the classification-specialized zero-one defer case.

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

$$\Phi_\lambda^{\mathrm{dec}}(w, s;\, x, y, m) \;:=\; -\log p_y(x) \;-\; \frac{\lambda}{J}\sum_{j=1}^{J}\Big(t_j \log u_j(x) + (1-t_j)\log\big(1-u_j(x)\big)\Big), \quad (10)$$

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

The first term, Φ_{cls}, is the multiclass log loss and the second, Φ_{exp}, is the λ-weighted mean binary cross-entropy (BCE) over experts. The total expert weight is λ, and each expert coordinate receives the per-expert weight λ/J, which scales its gradient, curvature, and transfer coefficient. Every $\lambda > 0$ yields the same unrestricted conditional minimizer, but with restricted models or a shared representation the weight also balances the two estimation tasks (Appendix C.4). A single $(K+J)$-dimensional output projection can implement both heads, with a softmax on the class block and a sigmoid on the expert block; the decoupling comes from the loss and its separate normalizers rather than from separate output layers.

Why we average expert losses. Suppose each of the J experts is copied $r \geq 1$ times and duplicate heads have identical outputs. Writing B_j for expert j's BCE term, the expert loss at fixed total weight λ satisfies

$$\frac{\lambda}{rJ}\sum_{b=1}^{r}\sum_{j=1}^{J} B_j = \frac{\lambda}{J}\sum_{j=1}^{J} B_j.$$

Since the class term is unchanged, duplication preserves the full sample loss at these matched outputs; in the Fashion-MNIST ablations, the duplicate heads are instead learned freely.

At test time, $h_{\mathrm{dec}}(x) \in \arg\max_k p_k(x)$, and the router compares the largest class and expert estimates:

$$r_{\mathrm{dec}}(x) \in \begin{cases} \{0\}, & \text{if } \max_{k\in[K]} p_k(x) \geq \max_{j\in[J]} u_j(x), \\ \arg\max_{j\in[J]} u_j(x), & \text{otherwise.} \end{cases} \tag{11}$$

5.2 Gradient Structure

The decoupled score derivatives avoid the pathologies of Section 4 (proofs in Appendix C.2).

Proposition 9 (Per-sample gradient of the decoupled surrogate)**.** *For one sample (x, y, m), regardless of how the representation is shared, the output-score derivatives are*

$$\frac{\partial \Phi_\lambda^{\mathrm{dec}}}{\partial w_r} = p_r(x) - \mathbf{1}\{r=y\}, \qquad \frac{\partial \Phi_\lambda^{\mathrm{dec}}}{\partial s_j} = \frac{\lambda}{J}\big(u_j(x) - t_j\big). \tag{12}$$

Moreover, suppose $w = w(\theta, \theta_c)$ and $s = s(\theta, \theta_e)$ are twice differentiable, where θ is a shared trunk and θ_c, θ_e are separate head parameters. Then

$$\nabla^2_{\theta_c\theta_e} \Phi_\lambda^{\mathrm{dec}} = 0, \qquad \nabla^2_\theta \Phi_\lambda^{\mathrm{dec}} = \nabla^2_\theta \Phi_{\mathrm{cls}} + \nabla^2_\theta \Phi_{\mathrm{exp}}.$$

Three properties address the failure modes of Section 4:

- *No amplification.* The gradient for expert j depends only on its prediction u_j and target t_j. At fixed per-expert weight, score, and target, adding other correct experts leaves this derivative unchanged.
- *No starvation.* If expert j is correct, then $\partial\Phi/\partial s_j = (\lambda/J)(u_j - 1) < 0$ whenever $u_j < 1$: its logit is pushed *up* even when a rival has a higher score.
- *No direct coupling.* The classifier score gradient involves only the class probabilities $p(x)$ and the label y, and it does not depend on the expert predictions or targets.

The second-order geometry confirms this decoupling (Appendix C.3). The class and expert terms depend on disjoint score coordinates, so the score Hessian has a *zero* mixed block:

$$\nabla^2_{(w,s)} \Phi_\lambda^{\mathrm{dec}} = \begin{pmatrix} \mathrm{Diag}(p) - pp^\top & 0 \\ 0 & \frac{\lambda}{J}\,\mathrm{Diag}\big(u_j(1-u_j)\big) \end{pmatrix}. \tag{13}$$

The class block is the standard softmax covariance with $\lambda_{\max} \leq 1/2$; the expert block is diagonal with entries bounded by $\lambda/(4J)$. In contrast, the curvature bound of additive CE scales as $(1 + |\mathcal{J}|)/2$ (Corollary 4.1), and the mixed-block bound of A-SM grows as $\sqrt{J}$ (Appendix G.4). OvA also separates the score geometry and recovers the Bayes targets at the unrestricted optimum, but its K class sigmoids need not sum to one during training.

Shared representations and two-stage training. Proposition 9 also separates the head-parameter updates when the shared trunk is learned: expert supervision enters neither the class loss nor its head update directly. Both tasks can still shape the common representation, since

$$\nabla_\theta \Phi_\lambda^{\mathrm{dec}} = (D_\theta w)^\top (p - e_y) + \frac{\lambda}{J}(D_\theta s)^\top (u - t).$$

The separation therefore removes the additional direct coupling of A-SM while retaining joint representation learning (Appendix J.1). With disjoint networks and a product feasible set, the empirical objective separates into class CE plus λ times mean expert BCE, and its global minimizer set, when nonempty, is the product of the component minimizer sets. The method can therefore be trained in two stages without changing its objective or its global solutions, whereas published two-stage softmax routers learn a different, predictor-dependent action representation (Appendix J.4).

5.3 Consistency and Surrogate-Regret Transfer

The conditional risk of (10) is the class CE plus separate expert BCE terms, and its unique minimizer over the closed probability domain is $p^\star(x) = \eta(x)$ and $u_j^\star(x) = \alpha_j(x)$ for all $j \in [J]$ (Appendix C.5). We write $\mathcal{E}_{\mathrm{dec}}$ for the population risk of (10) and use stars for infima.

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* (11)*. For any distribution* $\mathcal{D}$*,*

$$\mathcal{E}_{\perp}(f_{p,u}) - \mathcal{E}_{\perp}^{\star} \le \max\left\{\sqrt{2},\ \sqrt{\frac{2J}{\lambda}}\right\} \sqrt{\mathcal{E}_{\mathrm{dec}}(p,u) - \mathcal{E}_{\mathrm{dec}}^{\star}}\,. \tag{14}$$

Consequently, any sequence whose surrogate excess risk tends to zero also has L2D excess risk tending to zero, which proves Bayes consistency. The coefficient equals $\sqrt{2}$ whenever $\lambda/J \ge 1$ and is independent of J at any fixed positive per-expert weight, whereas at fixed total weight λ it eventually grows as $\sqrt{J}$. Both regimes concern surrogate-regret transfer; neither prescribes an optimal training weight.

Comparison with existing bounds. The Add. CE reduction has transfer coefficient $\sqrt{2(J+1)}$ (Mao et al., 2024d), the linear Mao25 member admits the linear transfer $(K+J)\Delta\mathcal{E}_{\mathrm{Mao}}$, and our proof also yields $\sqrt{2}\sqrt{\Delta\mathcal{E}_{\mathrm{OvA}}}$ for unit-weight OvA (Proposition 20); Appendix I compares these transfer shapes and normalizations.

Consultation costs. For expert costs $c_j \in [0,1]$, replace t_j by $1 - c_j$; the expert optimum becomes $1 - \mathbb{E}[c_j \mid X = x]$. For additive prices $b_j \ge 0$, costs $c_j = b_j + \mathbf{1}\{m_j \ne y\}$ can exceed one. With a common bound $B \ge 1 + \max_j b_j$, train on $1 - c_j/B$ and compare the expert estimates with $(B - 1 + \max_k p_k)/B$, keeping both alternatives on the same cost scale. Appendix J.5 derives this rule and a class-weighted extension.

6 Experiments

We evaluate the decoupled surrogate on four synthetic suites and four real-data benchmarks spanning vision, human responses, and tabular model routing. Controlled Fashion-MNIST studies separate the effects of loss, head layout, representation sharing, and training schedule. Backbone architectures, training budgets, and method-specific settings are reported in the appendix.

Synthetic suites (Appendix L). Each suite isolates one surrogate pathology under a distribution whose Bayes rule is known analytically. *(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): the norm of 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 L.4).

CIFAR-10 with synthetic experts (Appendix M). We train a ResNet-12 end-to-end on CIFAR-10 with J nested redundant experts whose correctness is drawn from a shared latent (Appendix L.1). Table 2 reports system and classifier accuracy: the decoupled surrogate is the only method in our comparison whose system accuracy *beats* the standalone classifier at every J, while Add. CE, A-SM, and OvA degrade as the redundant pool grows. A-SM's classifier accuracy collapses from 0.831 (J=8) to 0.680 (J=32), consistent with the class–expert coupling of Proposition 7.

Table 2: CIFAR-10 test accuracy 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	**.918**$_{\pm.001}$	**.919**$_{\pm.003}$	**.919**$_{\pm.002}$	.909$_{\pm.001}$	.909$_{\pm.002}$	**.908**$_{\pm.003}$
Classif. only	.910$_{\pm.002}$	.911$_{\pm.001}$	.908$_{\pm.002}$	**.910**$_{\pm.002}$	**.911**$_{\pm.001}$	**.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$_{\pm.025}$

CIFAR-10H with human annotators (Appendix N). We freeze a pretrained ResNet-12 and train linear L2D heads on CIFAR-10H (Peterson et al., 2019). For each image, J observed human responses are sampled without replacement; expert coordinates are annotation slots whose worker identity can vary across images. We use three training seeds for each of five response draws; Table 3 reports the fixed draw 0, which isolates training variation, and Appendix P.6 reports the full five-draw study. In this table, the decoupled surrogate has the highest mean system accuracy at each J, with 0.962 throughout. Across all 45 runs, its classifier accuracy remains 0.886–0.894 and it defers less than the other L2D methods in 44 or 45 matched runs per comparison.

Table 3: CIFAR-10H test accuracies for fixed per-image annotation draw 0 (mean ± sample SD over three training seeds). System accuracy uses the routed prediction; classifier accuracy uses the model without deferral. Bold marks the highest mean in each column before rounding.

	System accuracy ↑			Classifier accuracy ↑		
Method	$J = 5$	$J = 10$	$J = 20$	$J = 5$	$J = 10$	$J = 20$
Decoupled	$\mathbf{0.962}_{\pm 0.002}$	$\mathbf{0.962}_{\pm 0.003}$	$\mathbf{0.962}_{\pm 0.003}$	$\mathbf{0.891}_{\pm 0.002}$	$0.891_{\pm 0.004}$	$\mathbf{0.890}_{\pm 0.001}$
OvA	$0.957_{\pm 0.003}$	$0.955_{\pm 0.003}$	$0.960_{\pm 0.004}$	$0.889_{\pm 0.001}$	$0.889_{\pm 0.001}$	$0.888_{\pm 0.002}$
A-SM	$0.960_{\pm 0.002}$	$0.957_{\pm 0.003}$	$0.953_{\pm 0.002}$	$0.890_{\pm 0.001}$	$0.888_{\pm 0.001}$	$0.886_{\pm 0.001}$
Add. CE	$0.956_{\pm 0.003}$	$0.955_{\pm 0.004}$	$0.956_{\pm 0.002}$	$0.888_{\pm 0.002}$	$0.891_{\pm 0.002}$	$0.890_{\pm 0.002}$
PiCCE	$0.957_{\pm 0.003}$	$0.956_{\pm 0.001}$	$0.956_{\pm 0.005}$	$0.889_{\pm 0.002}$	$\mathbf{0.892}_{\pm 0.002}$	$0.889_{\pm 0.001}$
Classifier only	$0.890_{\pm 0.001}$	$0.889_{\pm 0.001}$	$0.889_{\pm 0.002}$	$0.890_{\pm 0.001}$	$0.889_{\pm 0.001}$	$0.889_{\pm 0.002}$
Mao25 (linear)	$0.948_{\pm 0.002}$	$0.948_{\pm 0.005}$	$0.946_{\pm 0.005}$	$0.060_{\pm 0.054}$	$0.072_{\pm 0.036}$	$0.082_{\pm 0.045}$

Covertype with model experts (Appendix O). On Forest CoverType (Blackard and Dean, 1999) with J=5 heterogeneous pretrained model experts (oracle +9.8 pp over the best single model), the decoupled surrogate, *the only L2D method in our comparison whose classifier improves over the standalone classifier* (0.929), reaches system accuracy 0.934 and classifier accuracy 0.941; every augmented-action baseline degrades the classifier, and Mao25 collapses to 100% deferral.

Separating loss and architecture (Appendix P.2). On Fashion-MNIST (Xiao et al., 2017), we exactly duplicate an eight-specialist panel to form $J = 16$ and $J = 32$, which adds neither new information nor a better action. A four-seed ablation crosses the decoupled and Add. CE losses with shared/split, shared/single, and parameter-matched separate-trunk architectures. The decoupled loss wins all 36 paired loss contrasts, and its weakest mean (0.9544) exceeds the strongest Add. CE mean (0.9351). The advantage persists with separate trunks (2.18–2.50 percentage points), which shows that neither final-layer layout nor feature sharing explains it; sharing features adds a further gain. At aligned initialization, its expert trunk gradients are also smaller relative to the class gradients and nearly orthogonal to them, unlike those of Add. CE (Appendix P.2).

Schedules, weights, and consultation costs. With disjoint networks, the mean accuracies of one-stage and two-stage training differ by less than 0.001. The shared one-stage model reaches 0.9585–0.9593, versus 0.9546–0.9557 for disjoint one-stage training and 0.9036–0.9118 for the predictor–rejector of Mao et al. (2024f) under the same protocol (Appendix P.4). In the 96-cell weight sweep, the spread of mean accuracy across pool sizes is at most 0.0016 at fixed total λ (Appendix P.3). With consultation prices β/j, $\beta \in \{0, 0.2, 0.4, 0.6, 0.8\}$, the decoupled surrogate has the lowest mean raw risk at every price; its coverage rises from 0.774 to 0.936, monotonically in each seed, while classifier accuracy stays within 0.915–0.920 (Appendix P.5).

7 Conclusion

The shared competition of augmented-action surrogates is not specific to log-softmax and persists across the comp-sum family. The decoupled surrogate removes it: it recovers (η, α) directly, with coordinatewise expert gradients, zero mixed score and head Hessian blocks, and a J-independent calibration constant at fixed λ/J. In our experiments it preserves rare specialists and classifier quality, and its advantage over Add. CE persists with separate trunks and with a single output projection. With disjoint networks, the same objective and global solution set admit two-stage training.

Acknowledgments

This research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-PhD-2023-01-041-J) and by A*STAR, and is part of the programme DesCartes which is supported by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme.

References

Pranjal Awasthi, Anqi Mao, Mehryar Mohri, and Yutao Zhong. 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.

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

Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Jock A. Blackard and Denis J. Dean. 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.

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

Yuzhou Cao, Hussein Mozannar, Lei Feng, Hongxin Wei, and Bo An. 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.

Nontawat Charoenphakdee, Zhenghang Cui, Yivan Zhang, and Masashi Sugiyama. Classification with rejection based on cost-sensitive classification. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1507–1517. PMLR, 2021. URL https://proceedings.mlr.press/v139/charoenphakdee21a.html.

Mohammad-Amin Charusaie, Hussein Mozannar, David Sontag, and Samira Samadi. Sample efficient learning of predictors that complement humans, 2022.

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

Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Learning with rejection. In Ronald Ortner, Hans Ulrich Simon, and Sandra Zilles, editors, *Algorithmic Learning Theory*, pages 67–82, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46379-7.

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

Corinna Cortes, Anqi Mao, Christopher Mohri, Mehryar Mohri, and Yutao Zhong. Cardinality-aware set prediction and top-k classification. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024b. URL https://openreview.net/forum?id=WAT3qu737X.

Corinna Cortes, Anqi Mao, Mehryar Mohri, and Yutao Zhong. Balancing the scales: A theoretical and algorithmic framework for learning from imbalanced data. In *Forty-second International Conference on Machine Learning*, 2025a.

Corinna Cortes, Mehryar Mohri, and Yutao Zhong. Improved balanced classification with theoretically grounded loss functions. In *Advances in Neural Information Processing Systems*, 2025b.

Corinna Cortes, Anqi Mao, Mehryar Mohri, and Yutao Zhong. Optimized deferral for imbalanced settings. In *Forty-Third International Conference on Machine Learning*, 2026.

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

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *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.

Aritra Ghosh, Himanshu Kumar, and P. S. Sastry. Robust loss functions under label noise for deep neural networks. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 1919–1925. AAAI Press, 2017.

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

Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009. URL https://api.semanticscholar.org/CorpusID:18268744.

Shuqi Liu, Yuzhou Cao, Qiaozhen Zhang, Lei Feng, and Bo An. Mitigating underfitting in learning to defer with consistent losses. In *International Conference on Artificial Intelligence and Statistics*, pages 4816–4824. PMLR, 2024.

Shuqi Liu, Yuzhou Cao, Lei Feng, Bo An, and Luke Ong. When more experts hurt: Underfitting in multi-expert learning to defer. *arXiv preprint arXiv:2602.17144*, 2026.

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

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

Anqi Mao, Christopher Mohri, Mehryar Mohri, and Yutao Zhong. 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.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Cross-entropy loss functions: Theoretical analysis and applications. In *International conference on Machine learning*, pages 23803–23828. PMLR, 2023b.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. h-consistency bounds: Characterization and extensions. *Advances in Neural Information Processing Systems*, 36, 2023c.

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

Anqi Mao, Mehryar Mohri, and Yutao Zhong. H-consistency guarantees for regression. In *Forty-first International Conference on Machine Learning*, 2024b.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Multi-label learning with stronger consistency guarantees. In *Advances in Neural Information Processing Systems*, 2024c.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Principled approaches for learning to defer with multiple experts. In *International Symposium on Artificial Intelligence and Mathematics*, 2024d.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Realizable h-consistent and bayes-consistent loss functions for learning to defer. *Advances in neural information processing systems*, 37:73638–73671, 2024e.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Regression with multi-expert deferral. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024f.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. A universal growth rate for learning with smooth surrogate losses. In *Advances in Neural Information Processing Systems*, 2024g.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Theoretically grounded loss functions and algorithms for score-based multi-class abstention. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 4753–4761. PMLR, 02–04 May 2024h. URL https://proceedings.mlr.press/v238/mao24a.html.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Enhanced h-consistency bounds. In *36th International Conference on Algorithmic Learning Theory*, 2025a. URL https://openreview.net/forum?id=qgnVGFJMJo.

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

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Principled algorithms for optimizing generalized metrics in binary classification. In *Forty-second International Conference on Machine Learning*, 2025c.

Mehryar Mohri and Yutao Zhong. Beyond tsybakov: Model margin noise and H-consistency bounds. In *International Symposium on Artificial Intelligence and Mathematics*, 2026a.

Mehryar Mohri and Yutao Zhong. Linear-core surrogates: Smooth loss functions with linear rates for classification and structured prediction. In *Forty-Third International Conference on Machine Learning*, 2026b.

Mehryar Mohri and Yutao Zhong. Mind the gap: Structure-aware consistency in preference learning. In *Forty-Third International Conference on Machine Learning*, 2026c.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT Press, 2012.

Yannis Montreuil, Axel Carlier, Lai Xing Ng, and Wei Tsang Ooi. Adversarial robustness in two-stage learning-to-defer: Algorithms and guarantees. In *Forty-second International Conference on Machine Learning*, 2025a.

Yannis Montreuil, Yeo Shu Heng, Axel Carlier, Lai Xing Ng, and Wei Tsang Ooi. A two-stage learning-to-defer approach for multi-task learning. In *Forty-second International Conference on Machine Learning*, 2025b.

Yannis Montreuil, Axel Carlier, Lai Xing Ng, and Wei Tsang Ooi. 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.

Yannis Montreuil, Hoang Duy Dang, Maxime Meyer, Lai Xing Ng, Axel Carlier, and Wei Tsang Ooi. 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.

Yannis Montreuil, Yeo Shu Heng, Lai Xing Ng, Axel Carlier, and Wei Tsang Ooi. 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.

Yannis Montreuil, Leïna Montreuil, Axel Carlier, Lai Xing Ng, and Wei Tsang Ooi. Learning-to-defer with expert-conditioned advice. *arXiv preprint arXiv:2603.14324*, 2026d.

Yannis Montreuil, Lai Xing Ng, Axel Carlier, Wei Tsang Ooi, and Yu Letian. Adversarial robustness in one-stage learning-to-defer. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026e. URL https://openreview.net/forum?id=gF8tv1SaR2.

Yannis Montreuil, Letian Yu, Axel Carlier, Lai Xing Ng, and Wei Tsang Ooi. Learning to defer in non-stationary time series via switching state-space models. *arXiv preprint arXiv:2601.22538*, 2026f.

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

Hussein Mozannar, Hunter Lang, Dennis Wei, Prasanna Sattigeri, Subhro Das, and David A. Sontag. 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.

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

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

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

Joshua C Peterson, Ruairidh M Battleday, Thomas L Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9617–9626, 2019.

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

Joshua Strong, Qianhui Men, and J Alison Noble. 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, pages 28413–28421, 2025.

Joshua Strong, Pramit Saha, Yasin Ibrahim, Cheng Ouyang, and Alison Noble. Identity-free deferral for unseen experts. In *The Fourteenth International Conference on Learning Representations*, 2026.

Dharmesh Tailor, Aditya Patra, Rajeev Verma, Putra Manggala, and Eric Nalisnick. Learning to defer to a population: A meta-learning approach. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 3475–3483. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/tailor24a.html.

Ambuj Tewari and Peter L. Bartlett. 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.

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

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

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

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

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

Yutao Zhong. *Fundamental Novel Consistency Theory: 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 and Wegkamp, 2008; Geifman and 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,h; Cortes et al., 2024b). The distinguishing feature of L2D is that abstention is replaced by routing to an external expert.

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 and 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. (2024d) derive quantitative guarantees for comp-sum-style reductions, Mao et al. (2024e, 2025b) 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., 2026e), non-stationary time-series deferral (Montreuil et al., 2026f), 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 and 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. These include 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., 2024f). 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), deferral in expert-imbalanced regimes (Cortes et al., 2026), and extractive question answering (Montreuil et al., 2026c).

The two families therefore differ in what they optimize: 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 and 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., 2023c,b), smooth surrogates (Mao et al., 2024g), regression

(Mao et al., 2024b), multi-label learning (Mao et al., 2024c), abstention (Mao et al., 2024h,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., 2025a), noise-adaptive refinements (Mohri and Zhong, 2026a), linear-rate smooth surrogates (Mohri and Zhong, 2026b), and structure-aware consistency for preference learning (Mohri and Zhong, 2026c); see Zhong (2025); Mao (2025) for unified treatments. Theorem 10 gives such a $\sqrt{\cdot}$-form bound for the decoupled surrogate, whose calibration constant is independent of J at fixed per-expert weight λ/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*

$$r^\star(x) \in \begin{cases} \{0\}, & \text{if } \max_{k\in[K]} \eta_k(x) \geq \max_{j\in[J]} \alpha_j(x), \\ \arg\max_{j\in[J]} \alpha_j(x), & \text{otherwise,} \end{cases} \tag{2}$$

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

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

Substituting the definition of $\ell_\perp$ gives

$$\begin{aligned} \mathcal{C}_\perp(h, r \mid x) &= \mathbb{E}\left[\mathbf{1}\{h(x) \neq Y\}\mathbf{1}\{r(x) = 0\} + \sum_{j=1}^{J} \mathbf{1}\{M_j \neq Y\}\mathbf{1}\{r(x) = j\} \,\middle|\, X = x \right] \\ &= \mathbf{1}\{r(x) = 0\} \Pr\big(h(x) \neq Y \mid X = x\big) + \sum_{j=1}^{J} \mathbf{1}\{r(x) = j\} \Pr\big(M_j \neq Y \mid X = x\big). \end{aligned} \tag{15}$$

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 (15), its conditional risk is

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

Since $\eta_k(x) = \Pr(Y = k \mid X = x)$, taking the complementary event gives

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

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 (15),

$$\mathcal{C}_\perp(h, j \mid x) = \Pr(M_j \neq Y \mid X = x) \qquad \text{for any } h.$$

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

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

It remains to compare the best class action with 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, so combining the two displays above gives

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

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 each baseline analysis that follows is compared against it on the same two axes. 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, we 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, where the dependence on the expert prediction m_j is left implicit, and let

$$\tilde{t}_j(x, y) \coloneqq 1 - c_j(x, y) \in [0, 1]$$

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

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

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

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. Definition 11 is the classification-specialized case of the bounded-cost formulation above, since in the present paper

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

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 this section formalizes the consequences of these two features.

C.2 Per-Sample Gradient Structure

Proposition 9 (Per-sample gradient of the decoupled surrogate)**.** *For one sample (x, y, m), regardless of how the representation is shared, the output-score derivatives are*

$$\frac{\partial \Phi_\lambda^{\mathrm{dec}}}{\partial w_r} \;=\; p_r(x) - \mathbf{1}\{r{=}y\}, \qquad \frac{\partial \Phi_\lambda^{\mathrm{dec}}}{\partial s_j} \;=\; \frac{\lambda}{J}\big(u_j(x) - t_j\big). \tag{12}$$

Moreover, suppose $w = w(\theta, \theta_c)$ and $s = s(\theta, \theta_e)$ are twice differentiable, where θ is a shared trunk and θ_c, θ_e are separate head parameters. Then

$$\nabla^2_{\theta_c\theta_e}\Phi_\lambda^{\mathrm{dec}} = 0, \qquad \nabla^2_\theta \Phi_\lambda^{\mathrm{dec}} = \nabla^2_\theta \Phi_{\mathrm{cls}} + \nabla^2_\theta \Phi_{\mathrm{exp}}.$$

Proof The samplewise decoupled loss is the sum of one multiclass cross-entropy term and J separate Bernoulli cross-entropy terms, so we can differentiate 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:

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

which is the usual multiclass cross-entropy gradient.

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

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

Differentiating with respect to s_j and applying the chain rule gives

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

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 λ/J gives

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

For the parameter-level statement, write the loss as $\Phi_{\mathrm{cls}}(w(\theta,\theta_c)) + \Phi_{\exp}(s(\theta,\theta_e))$, where $\Phi_{\mathrm{cls}} = -\log p_y$ and $\Phi_{\exp}$ is the weighted expert term of (10). Since the first summand does not depend on θ_e and the second does not depend on θ_c, their mixed head derivative $\nabla^2_{\theta_c\theta_e}$ is zero, and differentiating the sum twice in θ gives the stated sum of trunk Hessians. Appendix J.1 displays all parameter blocks.

The classifier residuals and the expert residuals therefore 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, which shows both that the curvature is bounded and 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

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

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

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

and for $j \neq \ell$,

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

because Φ_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:

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

Hence the full Hessian is block diagonal:

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

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, the softmax covariance matrix $\mathrm{Diag}(p) - pp^\top$ has largest eigenvalue at most $1/2$ (Mohri et al., 2012). This is the intrinsic curvature scale of ordinary multiclass cross-entropy, and the decoupled surrogate leaves it unchanged.

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

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

The class-side curvature is therefore bounded by the same constant as ordinary multiclass cross-entropy, while the expert-side bound $\lambda/(4J)$ does not depend on the number of experts at fixed per-expert weight λ/J and decreases as J grows when the total expert weight λ is held fixed. Adding experts thus creates neither the multiplicity factor of additive CE nor the mixed class–expert block of A-SM; the absence of winner-take-all suppression is a first-order property, given by Proposition 9.

The class and expert losses have no mixed block, either in the output scores or between the two sets of head parameters. With a learned shared trunk, both terms still contribute to the trunk through that common representation, as described in Proposition 9.

C.4 The Role of λ/J and How to Choose λ

The coefficient λ multiplies the mean expert BCE, so it is the *total* expert weight, while each expert's score gradient and curvature scale with the per-expert weight λ/J. Fixing λ therefore fixes the total weight while reducing the individual weights as J grows; fixing λ/J instead scales λ linearly with J. Both choices preserve the unrestricted probability targets, since every $\lambda > 0$ has the same conditional minimizer.

Theorem 10 has coefficient $\max\{\sqrt{2}, \sqrt{2J/\lambda}\}$, which is independent of J at any fixed positive per-expert weight and equals $\sqrt{2}$ when $\lambda/J \geq 1$. This coefficient describes the transfer from a given surrogate excess; it does not control the surrogate excess attained after training. With a shared trunk, the total expert weight also controls the balance of representation learning between the two tasks, so the theorem gives no universal recommendation to increase λ with J.

Appendix P.3 compares fixed-total, fixed-per-expert, and validation-selected policies on exactly duplicated Fashion-MNIST expert panels. All three preserve system accuracy in that controlled setting, including fixed-total weights across a 128-fold range. In every policy, weights and checkpoints are selected on the validation deferral loss, so these results describe the robustness of the tested protocol rather than an invariance that holds for every optimizer or shared trunk.

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 (11). For any distribution $\mathcal{D}$,*

$$\mathcal{E}_{\perp}(f_{p,u}) - \mathcal{E}_{\perp}^{\star} \leq \max\left\{\sqrt{2},\ \sqrt{\frac{2J}{\lambda}}\right\} \sqrt{\mathcal{E}_{\mathrm{dec}}(p,u) - \mathcal{E}_{\mathrm{dec}}^{\star}}\,. \tag{14}$$

Proof Throughout this proof we abbreviate the per-expert weight as $\beta := \lambda/J$; this is a local shorthand for the fraction appearing in (10), not a separate hyperparameter. Fix x and define

$$\delta_{\perp}(p,u \mid x) := \mathcal{C}_{\perp}(f_{p,u} \mid x) - \mathcal{C}_{\perp}^{\star}(x)$$

for the conditional L2D regret and

$$\delta_{\mathrm{dec}}(p,u \mid x) := \mathcal{C}_{\mathrm{dec}}(p,u \mid x) - \mathcal{C}_{\mathrm{dec}}^{\star}(x)$$

for the conditional surrogate excess, where $\mathcal{C}_{\perp}^{\star}(x)$ and $\mathcal{C}_{\mathrm{dec}}^{\star}(x)$ denote the corresponding conditional infima.

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

$$\begin{aligned}\mathcal{C}_{\mathrm{dec}}(p,u \mid x) = &-\sum_{k=1}^{K} \eta_k(x) \log p_k(x) \\ &- \beta \sum_{j=1}^{J} \Big(\alpha_j(x) \log u_j(x) + (1-\alpha_j(x)) \log(1-u_j(x))\Big).\end{aligned} \tag{16}$$

The class term depends only on $p(x)$, and the j-th expert term depends only on $u_j(x)$. We write $\mathcal{E}_{\mathrm{dec}}(p,u) := \mathbb{E}_X[\mathcal{C}_{\mathrm{dec}}(p,u \mid X)]$ for the population surrogate risk and $\mathcal{E}_{\mathrm{dec}}^{\star}$ for its infimum over

measurable $p : \mathcal{X} \to \Delta^K$ and $u : \mathcal{X} \to [0,1]^J$; since (η, α) is measurable and attains every conditional infimum (shown below), $\mathcal{E}^\star_{\mathrm{dec}} = \mathbb{E}_X[\mathcal{C}^\star_{\mathrm{dec}}(X)]$.

With the conventions $0 \log 0 = 0$ and $-\log 0 = +\infty$, 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 (16) yields

$$\begin{aligned}\mathcal{C}_{\mathrm{dec}}(p, u \mid x) = H(\eta(x)) + \beta \sum_{j=1}^{J} H_{\mathrm{B}}(\alpha_j(x)) \\ + \mathrm{KL}\big(\eta(x) \,\|\, p(x)\big) + \beta \sum_{j=1}^{J} \mathrm{KL}\Big(\mathrm{Bern}(\alpha_j(x)) \Big\| \mathrm{Bern}(u_j(x))\Big).\end{aligned} \tag{17}$$

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

$$\delta_{\mathrm{dec}}(p, u \mid x) \;=\; \mathrm{KL}\big(\eta(x) \,\|\, p(x)\big) \;+\; \beta \sum_{j=1}^{J} \mathrm{KL}\Big(\mathrm{Bern}(\alpha_j(x)) \Big\| \mathrm{Bern}(u_j(x))\Big). \tag{18}$$

This identity shows that the conditional surrogate excess is exactly the sum of one class estimation error and J expert estimation errors weighted by β, each measured in the 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 $a^\star \geq b^\star$ 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

$$\delta_\perp(p, u \mid x) \le 2 \max\{\varepsilon_{\mathrm{cls}}(x), \varepsilon_{\exp}(x)\}. \tag{19}$$

We next bound these two sup-norm errors by the conditional surrogate excess. Since $p(x)$ and $\eta(x)$ both sum to one, the positive and negative parts of $p(x) - \eta(x)$ each have total mass $\frac{1}{2}\|p(x) - \eta(x)\|_1$, so no coordinate can exceed this value, and Pinsker's inequality for categorical distributions gives

$$\varepsilon_{\mathrm{cls}}(x) \le \tfrac{1}{2}\|p(x) - \eta(x)\|_1 \le \sqrt{\tfrac{1}{2}\,\mathrm{KL}(\eta(x)\|p(x))}.$$

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

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

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

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

Combining these inequalities with (18) gives

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

Substituting into (19),

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

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

$$\delta_\perp(p, u \mid x) \le \Gamma\big(\delta_{\mathrm{dec}}(p, u \mid x)\big). \tag{20}$$

Taking expectations over X and using Jensen's inequality, with the concavity of Γ, 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 Γ and $\beta = \lambda/J$. ■

The argument thus has three steps: 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 Since the unrestricted class contains (η, α), any exact minimizer (p, u) achieves the population infimum, and hence its conditional excess satisfies $\delta_{\mathrm{dec}}(p, u \mid x) = 0$ almost surely. By the KL decomposition (18), 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 and Sontag, 2020)

This section collects the results on the additive cross-entropy baseline in one place. Additive CE is misaligned on both axes at once: it changes the population target, and it rescales the local geometry by the realized number of correct experts.

We follow the order of the main paper's two-axis analysis. We first restate the surrogate and derive what it estimates at the conditional optimum, and we then derive its samplewise gradient and Hessian, making the multiplicity factor $1 + |\mathcal{J}|$ explicit. A comparison with the decoupled surrogate and a worked example close the section and show how the two effects arise on a concrete sample.

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 and Sontag (2020))**.** Let $a(x) \in \mathbb{R}^{K+J}$ be an augmented score vector and define

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

The additive augmented cross-entropy is

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

Prediction uses the augmented argmax of $a(x)$, and the additive reward structure analyzed below is consistent with the redundancy-driven underfitting reported by Liu et al. (2026).

D.2 Conditional Target Distortion

Proposition 14 (Conditional minimizer of additive CE)**.** *The conditional risk of additive CE at x is uniquely minimized over $q \in \Delta^{K+J}$ at*

$$q_y^\star(x) = \frac{\eta_y(x)}{1 + U(x)}, \qquad q_{K+j}^\star(x) = \frac{\alpha_j(x)}{1 + U(x)}, \qquad y \in [K],\ j \in [J]. \tag{21}$$

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 (Competition and overlap beyond log loss)**.** *Let $d = K + J$, $a \in \mathbb{R}^d$, $q = \mathrm{softmax}(a)$, and let e_i be the standard basis vectors. Set $\mathcal{T} = \{y\} \cup \{K + j : j \in \mathcal{J}(y, m)\}$, $N = |\mathcal{T}|$, and $L_\tau = \sum_{i \in \mathcal{T}} \ell_\tau(a, i)$. Write $C(q) = \mathrm{Diag}(q) - qq^\top$, and define*

$$\bar{q}_\tau = \frac{1}{N} \sum_{i \in \mathcal{T}} q_i^{\tau - 1}, \qquad \omega_i = \frac{q_i^{\tau-1}}{N \bar{q}_\tau}, \qquad \bar{e}_\tau = \sum_{i \in \mathcal{T}} \omega_i e_i, \qquad \bar{R}_\tau = \sum_{i \in \mathcal{T}} \omega_i (q - e_i)(q - e_i)^\top.$$

Then, for every $\tau \geq 0$,

$$\nabla_a L_\tau = N\bar{q}_\tau(q - \bar{e}_\tau), \qquad \nabla_a^2 L_\tau = N\bar{q}_\tau[C(q) - (\tau - 1)\bar{R}_\tau]. \tag{3}$$

Every omitted action $k \notin \mathcal{T}$ has $\partial L_\tau / \partial a_k = N\bar{q}_\tau q_k > 0$. At uniform scores this derivative is $Nd^{-\tau}$, and every nonzero zero-sum direction supported on omitted actions is an eigenvector of the Hessian with eigenvalue $Nd^{-\tau}$. Relative to the true-class-only loss at the same scores and dimension, both quantities are amplified exactly by N.

Proof For $\tau \neq 1$, differentiating the base loss with respect to q_i gives $-q_i^{\tau-2}$; this also holds at $\tau = 1$. Since $\nabla_a q_i = q_i(e_i - q)$,

$$\nabla_a \ell_\tau(a, i) = q_i^{\tau-1}(q - e_i).$$

Differentiating this product, using $D_a q = C(q)$ and $\nabla_a q_i^{\tau-1} = -(\tau - 1)q_i^{\tau-1}(q - e_i)$, gives

$$\nabla_a^2 \ell_\tau(a, i) = q_i^{\tau-1}[C(q) - (\tau - 1)(q - e_i)(q - e_i)^\top].$$

Summing over $\mathcal{T}$ and factoring out $\sum_{i\in\mathcal{T}} q_i^{\tau-1} = N\bar{q}_\tau$ proves both identities. If $k \notin \mathcal{T}$, $(\bar{e}_\tau)_k = 0$, so its derivative is positive at finite scores. At uniform scores $q_i = 1/d$, this derivative equals $Nd^{-\tau}$. For a unit vector z supported on omitted coordinates with $\sum_k z_k = 0$, $q^\top z = 0$ and $e_i^\top z = 0$ for every active i. Consequently $C(q)z = z/d$ and $\bar{R}_\tau z = 0$, giving $\nabla_a^2 L_\tau z = Nd^{-\tau} z$. None of these steps uses the specific form of $\mathcal{T}$, so the identities hold for any target set. The true-class-only loss has $N = 1$ and the same equations on these coordinates, proving the stated comparison at fixed d and scores. ■

General target weighting, without a comp-sum assumption. For any multiclass base loss $\ell : \mathbb{R}^d \times [d] \to \mathbb{R}$ and finite, score-independent weights $v_i \geq 0$ with $V = \sum_i v_i > 0$, set $\pi_i = v_i/V$. Then

$$L_{\ell,v}(a) = \sum_i v_i \ell(a, i) = V\,\mathbb{E}_{I\sim\pi}[\ell(a, I)]. \tag{22}$$

If ℓ is twice differentiable, its gradient and Hessian are V times the corresponding π-averaged derivatives. This factorization alone need not imply monotone growth of derivative norms. For the additive reduction, $v = e_y + \sum_{j\in\mathcal{J}} e_{K+j}$, so $V = N$ and π is uniform on $\mathcal{T}$. Conditioning on $X = x$ gives

$$\mathcal{C}_\ell(a \mid x) = \sum_{k=1}^{K} \eta_k(x)\ell(a, k) + \sum_{j=1}^{J} \alpha_j(x)\ell(a, K + j) = (1 + U(x))\mathbb{E}_{I\sim\bar{\pi}(x)}[\ell(a, I)],$$

where $\bar{\pi} = (\eta, \alpha)/(1 + U)$. For bounded costs in Mao et al.'s general construction, replace $\alpha_j(x)$ by $1 - \mathbb{E}[c_j \mid X = x]$. These identities follow from finite sums and conditional expectation; no properness, softmax representation, or expert independence is required. If $V = 0$, the loss and its derivatives vanish.

Corollary 4.1 (Mozannar's log-softmax instance)**.** *For score-independent weights $v_i \geq 0$, set $V = \sum_i v_i$ and $L_v = -\sum_i v_i \log q_i$. Then*

$$\frac{\partial L_v}{\partial a_i} = Vq_i - v_i, \tag{5}$$

$$\nabla_a^2 L_v = VC(q), \qquad \lambda_{\max}(\nabla_a^2 L_v) \leq V/2. \tag{6}$$

For Add. CE (Mozannar and Sontag, 2020), $V = N = 1 + |\mathcal{J}|$.

Proof For $V > 0$, dividing each weight by V gives $L_v = V\,\mathrm{CE}(v/V, q)$ exactly. For the derivatives, write $L_v = -v^\top a + V \log \sum_r e^{a_r}$. Differentiating gives $\nabla L_v = Vq - v$ and $\nabla^2 L_v = V(\mathrm{Diag}(q) - qq^\top)$. For any unit vector z, the covariance quadratic form is the variance of z_A for $A \sim q$. The range of z_A is at most $\sqrt{2}$, since $|z_i - z_j| \leq \sqrt{2(z_i^2 + z_j^2)} \leq \sqrt{2}$, and the variance of a random variable supported on an interval of length $\sqrt{2}$ is at most $(\sqrt{2})^2/4 = 1/2$. This proves the $V/2$ eigenvalue bound, including the case $V = 0$. For additive CE we also give a direct calculation that keeps track of where the factor $1 + |\mathcal{J}|$ enters; since the loss is a scaled log-sum-exp plus linear terms, we derive the gradient first and then the Hessian.

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

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

For completeness, we also verify the eigenvalue bound row by row with Gershgorin's circle theorem, writing $M \coloneqq \mathrm{Diag}(q) - qq^\top$.

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 λ 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

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

■

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 address the same decision problem in different ways. Additive CE folds classes and experts into one augmented softmax, whereas the decoupled surrogate keeps the class posterior and the expert utilities as separate objects throughout training and prediction. The next two paragraphs make the consequences of this difference precise.

Axis (i): statistical target. Additive CE does not recover (η, α) 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 $U(x)$, which 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 does not estimate the native decision quantities on their original scale.

The decoupled surrogate has no such dependence, since 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. In practice, 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.

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 $|\mathcal{J}| = 4$.

In this clean redundancy scenario, the decision problem contains only one useful expert behavior, copied four times. A reasonable surrogate should therefore treat this sample almost the same way as a sample with only one correct expert, and 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 largest class posterior 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, although only the number of redundant copies changed and the underlying decision problem did not.

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, so both quantities are inflated by the redundancy count.

The rescaling matters because it differs across samples. If only expert 1 were present, the same derivation would give total positive mass 2 rather than 5, because the sample would reward one class and one expert, so four exact copies multiply the weight of this sample relative to low-overlap samples by $5/2$, 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 only expert 1 were present ($J = 1$), its gradient would be $\lambda(\frac{1}{2} - 1) = -\frac{\lambda}{2}$, which equals the total expert gradient $4 \cdot (-\frac{\lambda}{8})$ above, so at fixed total weight duplication leaves both the class gradient and the total expert gradient unchanged. The decoupled surrogate updates each expert coordinate because that expert is correct, not because 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. The decoupled surrogate therefore lacks the mechanism through which additive CE can become dominated by high-overlap samples.

E PiCCE (Liu et al., 2026)

PiCCE (Liu et al., 2026) was designed to remove the additive amplification of CE (Mozannar and 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, in which correct experts that fail to win the internal $\arg\max$ receive negative updates.

The section parallels the one on 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: PiCCE improves on additive CE because it removes the overlap-dependent amplification factor, but it does so by forcing correct experts to compete with one another, and we show that this competition causes both starvation and a score-dependent target.

E.1 Formulation

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

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

Prediction again uses the augmented argmax over $a(x)$, as for additive CE.

E.2 Winner-Take-All Starvation

Proposition 5 (Winner-take-all starvation in PiCCE)**.** *For $\mathcal{J}(y, m) \neq \varnothing$ and a unique winning correct expert, 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 $j' \in \mathcal{J}$ with $j' \neq j^\star$,*

$$\frac{\partial\Phi^{\mathrm{PiCCE}}}{\partial a_{K+j'}} \;=\; 2q_{K+j'}(x) \;>\; 0. \tag{8}$$

Gradient descent pushes expert j''s logit downward*, even though it is correct. At a tie, the same formula is the derivative of the selected branch.*

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 unique PiCCE winner. Since $a_{K+j^\star} > a_{K+j}$ for every other $j \in \mathcal{J}(y, m)$, this strict ordering persists on a neighbourhood of a, so the winner is locally constant and, on that neighbourhood,

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

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, and every other coordinate sees only the positive softmax term $2q_i$. At a tie, the deterministic tie-break fixes $j^\star$, and the same computation gives the derivative of the selected branch.

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

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

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

The expert j' 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 Section 4.2. Condition on $X = x$ and take expectations over $(Y, \mathbf{M})$. The PiCCE loss rewards expert $j^\star = \arg\max_{j \in \mathcal{J}(Y,\mathbf{M})} a_{K+j}$ whenever $\mathcal{J} \neq \varnothing$. Define $\gamma_j(a, x) = \Pr(j \in \mathcal{J}(Y, \mathbf{M}) \text{ and } j = j^\star \mid X = x)$. 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, in contrast to additive CE, where all coefficients are fixed by the distribution alone.

For a concrete case, 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, which 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 fixed on the neighbourhood of a on which the winner is unique. 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, and by the additive-CE derivation above every eigenvalue of this matrix lies in $[0, 1/2]$. Multiplying by the prefactor 2 gives

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

PiCCE is therefore well behaved in the second-order sense in which additive CE is not, which isolates the actual failure mode. The problem with PiCCE is not exploding curvature or overlap-dependent scaling but the direction of the updates: the surrogate suppresses correct experts that fail to win the internal comparison, and the 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, which is the first 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, so the data alone determine the

target. Under PiCCE, by contrast, shared correct mass is assigned to whichever expert is currently winning the internal competition.

Because the winner map changes during training, the quantity PiCCE estimates can itself shift as training proceeds, whereas the class and expert subproblems of the decoupled surrogate are proper losses with fixed conditional targets (η, α).

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, so that on a sample where several experts are simultaneously correct, PiCCE turns agreement into competition.

The decoupled surrogate makes the opposite design choice, keeping the per-coordinate updates bounded while reinforcing every correct expert independently. Agreement among correct experts therefore creates neither additive amplification nor winner-take-all suppression. PiCCE and the decoupled surrogate agree on the need to avoid additive amplification, but they differ on whether multiple correct experts should compete with one another or all receive supportive updates.

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. In this situation one would like the training signal to preserve both experts, since the generalist is correct but the specialist is also correct on the very region for which it was meant to be useful; PiCCE instead forces the two experts 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, and the sample acts as supportive evidence for both experts rather than as an internal tournament.

This example is the PiCCE analogue of the redundant-expert example for additive CE. Whenever several experts are correct on the same sample, PiCCE converts their agreement into competition, whereas the decoupled surrogate keeps the bounded geometry without introducing that competition; this is the mechanism targeted by the rare-specialist experiments of Section 6.

F Mao25 (Mao et al., 2025b)

Mao25 moves away from additive rewards and winner-take-all selection by optimizing only the total mass assigned to the acceptable set. The object it fits is therefore not the full vector (η, α) but this single set mass. That choice resolves the two most immediate geometric issues and keeps the local geometry comparatively benign, but it also removes the signal needed to rank experts within the acceptable set, which is why the surrogate cannot express preferences among several acceptable experts.

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

F.1 Formulation

We use the member $\Psi(u) = 1 - u$, which Mao et al. (2025b) also instantiate in their multiple-expert experiments and which is the version in our empirical comparison.

Definition 16 (Mao25 single-stage surrogate with $\Psi(u) = 1 - u$, Mao et al. (2025b))**.** 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 $a(x)$, and the surrogate depends only on the total mass assigned to the acceptable set, which consists of the true class and all correct experts.

To see what this implies at the conditional level, fix an input x and take the expectation over the random pair $(Y, \mathbf{M})$. 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

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

The conditional risk is thus a linear function of the augmented probabilities q, and minimizing it is 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 Δ^{K+J}. A linear objective over a simplex attains its maximum at a vertex, so over the closed simplex the minimum is attained by a point mass on an action whose coefficient is maximal among

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

Finite scores only approach this vertex. If the maximizer is unique, this point mass is the unique minimizer over Δ^{K+J} and sits on the Bayes-optimal action. If several actions tie for the maximum, any convex combination over those maximizers is also optimal.

Mao25 can therefore recover a Bayes-optimal action, but it does not recover the full vector (η, α): 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: we first show that the loss depends on the score vector only through the acceptable-set mass, and then 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 $S_{\mathcal{S}} - 1$, and 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 the multiclass MAE loss up to a constant factor. The same weak-gradient phenomenon is well known for MAE-like losses in noisy-label learning (Zhang and Sabuncu, 2018). ■

F.3 Curvature

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

Write

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

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:

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

This is the complete second derivative, since no further term arises from differentiating the explicit gradient formula. It simplifies in four cases, according to whether i and r belong to the acceptable set.

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,

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

None of these expressions contains a term that depends on which acceptable expert is statistically better, so 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 (η, α). The decoupled surrogate does: at the conditional minimum, it 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, but 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, and in that sense it is geometrically cleaner than additive CE and PiCCE. Its geometry is nevertheless 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, and 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 $S_{\mathcal{S}}(1 - S_{\mathcal{S}})$, which is at most $1/4$ 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 comparison with the decoupled surrogate therefore differs from the PiCCE comparison, because Mao25 neither suppresses correct experts nor amplifies 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 largest class posterior 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.

Under the decoupled surrogate, and up to the positive factor λ/J, the 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 $u_2 = 0.5$, so the ranking signal is present 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 $S_S(1-S_S)$. 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, since it restores bounded estimates and recovers the correct target. Its remaining weakness lies at the 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, whereas for A-SM the conditional target is already correct. Our aim here is to isolate the remaining axis-(ii) weakness 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 ξ_k for the class estimates and ψ_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.

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

$$\Phi^{\mathrm{ASM}}(a;x,y,m) := -\log\xi_y(a(x)) - \sum_{j=1}^{J}\Big(\mathbf{1}\{m_j = y\}\log\psi_j(a(x)) + \mathbf{1}\{m_j \neq y\}\log\big(1-\psi_j(a(x))\big)\Big). \tag{23}$$

Prediction uses the maxima-preserving property of (ξ,ψ), 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 ψ_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.

The sigmoid representation also 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, but that the class logits appear inside that reference level, so class and expert optimization remain 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 property is the reason A-SM is the strongest augmented-action baseline on axis (i). Unlike additive CE, PiCCE, and Mao25, A-SM identifies the right statistical object, but it reaches that object through a shared geometry that still couples class and expert optimization.

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 a correct target is sufficient by itself, and 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*

$$\frac{\partial \Phi^{\mathrm{ASM}}}{\partial a_r} = \underbrace{(\xi_r - \mathbf{1}\{r=y\})}_{\textit{class term}} - \underbrace{\pi_r \sum_{j=1}^{J} \big(\psi_j - \mathbf{1}\{m_j=y\}\big)}_{\textit{expert leakage}} . \tag{9}$$

Proof We work on a region where the maximizing class index $k^\star$ is unique and therefore locally fixed; at a tie, the formula is the derivative of the selected branch. 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 an ordinary Bernoulli logistic loss. 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 π_r.

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

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

Viewed in isolation, the expert block therefore behaves exactly like a standard Bernoulli cross-entropy; the pathology is that the same expert residuals also enter the class gradient.

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

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

and adding the class term $\xi_r - \mathbf{1}\{r = y\}$ yields the stated formula.

If many experts are correct but currently underestimated, then many terms $\psi_j - \mathbf{1}\{m_j = y\}$ are negative, so the leakage term $-\pi_r \sum_j (\psi_j - \mathbf{1}\{m_j = y\})$ adds an extra positive 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, because it also carries the aggregate expert residual. ■

G.4 Curvature

For A-SM the most informative second-order object is the mixed class–expert block, because it captures 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 π_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

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

Rank-one structure. The mixed block is an outer product, so it has rank at most one, and all non-max class coordinates therefore interact with the expert block through the same class-side direction π.

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

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

Now π 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 π_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. Because π is normalized over the non-max classes only, it does not shrink as $\xi_{k^\star} \to 1$, and $\|\pi\|_2$ reaches one when the non-max mass sits on a single class, as it always does for $K = 2$. The mixed coupling is therefore strongest when the non-max mass is concentrated and many experts are in the uncertain middle regime rather than near 0 or 1.

In this regime the contrast with the decoupled surrogate is largest, since the decoupled 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, 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 natural second-order object to compare with the decoupled surrogate, because it is where the class and expert tasks interact most directly.

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, since 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.

The remaining gap between A-SM and the decoupled surrogate therefore lies on axis (ii).

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 leakage is the key residual weakness of A-SM, which has already repaired the target distortion and boundedness problems but has not fully separated the class and expert subproblems. Among the methods compared here, the decoupled surrogate combines this zero mixed block with a categorical class posterior; OvA also has a zero mixed block and aligned unrestricted targets.

The practical consequence is that the two methods react differently during training. At fixed scores under the decoupled loss, an error in expert j changes only its own score derivative, whereas under A-SM the same expert error also perturbs the class update whenever the non-max class normalizer is active. Even though both methods aim for the same conditional optimum, they therefore do not follow the same local route toward it: A-SM reaches the correct target through coupled class and expert updates, whereas the decoupled surrogate reaches it with the two sets of updates kept separate.

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

$$\xi_2 = 0.25, \qquad \pi_2 = 0.50, \qquad \sum_{j=1}^{4} \big(\psi_j - \mathbf{1}\{m_j = y\}\big) = 0.80,$$

so that $\xi_3 = 0.25$ and $\xi_1 = 0.50$, consistent with $k^\star = 1$. Then Proposition 7 gives

$$\frac{\partial \Phi^{\mathrm{ASM}}}{\partial a_2} = (\xi_2 - 1) - \pi_2 \sum_{j=1}^{4} \big(\psi_j - \mathbf{1}\{m_j = y\}\big) = -0.75 - 0.40 = -1.15.$$

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, -0.40, carries no class information: it is leakage from the expert heads, which 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.

If A-SM were fully decoupled, the class-2 logit would receive exactly the ordinary cross-entropy gradient -0.75. Instead, the expert block adds -0.40 to this coordinate and the same -0.40 to the class-3 logit, whose derivative flips from $\xi_3 = 0.25$ to -0.15, so the class gradient moves from

$(0.50, -0.75, 0.25)$ to $(0.50, -1.15, -0.15)$ and changes in both magnitude and direction. 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, so the class update would reflect only the classifier's own error on the sample. The example thus illustrates the axis-(ii) difference between A-SM and the decoupled surrogate numerically: both can target the right optimum, but the decoupled loss keeps the class-score derivative free of expert residuals.

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

The conclusion for A-SM is therefore narrower than for additive CE or PiCCE: fixing the statistical target is not enough, since 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 simplest augmented baseline on the optimization axis. It removes the shared softmax entirely and replaces it by independent binary logistic tasks, so 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 what is lost when the class side is represented not by one categorical distribution in Δ^K, but by K independent Bernoulli scores. At the conditional optimum, OvA still identifies the correct class marginals $\eta_k(x)$ and expert utilities $\alpha_j(x)$, and the gap appears only 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.

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

$$\chi_k(y) \coloneqq 2\mathbf{1}\{y = k\} - 1, \qquad \zeta_j(y, m) \coloneqq 2\mathbf{1}\{m_j = y\} - 1,$$

and

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

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

$$b_k(y) \coloneqq \mathbf{1}\{y = k\}, \qquad t_j \coloneqq \mathbf{1}\{m_j = y\}.$$

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

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

This form makes two features explicit: every coordinate is optimized independently, and the class head is not a softmax over the K classes but 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 $-\infty$ or $+\infty$, respectively.*

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

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

This expression separates completely across coordinates, with one Bernoulli cross-entropy for each class coordinate and one 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

$$\varphi''(v)=\sigma(v)(1-\sigma(v))>0,$$

since $\sigma(v)\in(0,1)$ for every finite v, so φ is strictly convex. 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 $j\in[J]$ whose targets lie in $(0,1)$, with the boundary cases understood as the limits above. ■

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 $\alpha_j(x)$. Its limitation is therefore not that it aims at the wrong population object but 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

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

and 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:

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

so

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

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

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

Any weakness of OvA therefore lies outside its local optimization geometry.

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

Axis (i): statistical target. OvA and the decoupled surrogate agree at the population level: 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 are.

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, so the vector $p(x)$ is 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 Δ^K unless the model happens already to be at the conditional optimum.

Axis (ii): local geometry. On the geometric axis the two methods are similar: both keep the class block separate from the expert block, both have bounded curvature (softmax covariance with $\lambda_{\max} \leq 1/2$ for the decoupled class block, diagonal entries at most $1/4$ for OvA), and both avoid all of the shared-geometry pathologies that affect the augmented-action family. The decoupled surrogate is therefore not preferred because it is more decoupled than OvA, since there is little to improve on that axis.

The decoupled surrogate is preferred because it combines this geometry with a categorical class model. Its routing rule 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.

If the decoupled class head had the same three class logits g, its softmax would give

$$p \approx (0.40, 0.33, 0.27), \qquad u_1 = 0.70,$$

and the comparison $\max_k p_k \approx 0.40 < 0.70$ leads the decoupled surrogate to defer. This does not mean that OvA is inconsistent, because 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, and because routing uses the current scores, this lack of simplex structure can matter in practice.

The example illustrates the remaining gap between OvA and the decoupled surrogate. OvA has already fixed the geometric side of the problem but not the representation of class uncertainty, whereas the decoupled surrogate keeps the class–expert separation 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 those available for the baselines, focusing on how the calibration constant scales with the size of the multi-expert action space. Two 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. (2024d), which covers the Mozannar-style augmented-action construction; it is not the constant for the later Mao25 single-stage surrogate (Mao et al., 2025b). 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

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

Since $\sqrt{2J/\lambda} = \sqrt{2/(\lambda/J)}$, the constant depends on J only through the per-expert weight λ/J; whenever $\lambda/J \geq 1$, it equals $C_{\mathrm{dec}} = \sqrt{2} \approx 1.41$.

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

Mao25 (Mao et al., 2025b). For the single-stage surrogate studied by Mao et al. (2025b), the corresponding excess-risk constant scales with the full augmented action dimension. In the multi-expert setting considered here, this yields

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

which multiplies the surrogate excess itself rather than its square root, so the dependence on the number of experts is linear: $O(J)$ for fixed K.

OvA and the categorical-head distinction. Proposition 20 gives OvA the quantitative coefficient $\sqrt{2}$ for unit per-coordinate Bernoulli weights, so J-independence is not specific to the decoupled surrogate and is shared by this other separated surrogate. For the decoupled loss the expert weights are λ/J; at $\lambda/J = 1$, which matches the unit Bernoulli weights of OvA, both coefficients equal $\sqrt{2}$. The distinction between the two lies in the categorical class constraint, which the independent OvA class sigmoids need not satisfy away from their optimum.

The linear Mao25 instance. For the acceptable-mass loss actually analyzed here, a direct calculation also gives a *linear* transfer with factor $K + J$. Let $\mu = (\eta, \alpha)$, $\mu_\star = \max_i \mu_i$, and let $\hat{\imath} \in \arg\max_i q_i$. Its conditional surrogate excess is $\sum_i (\mu_\star - \mu_i) q_i$, which is at least $(\mu_\star - \mu_{\hat{\imath}}) q_{\hat{\imath}}$. Since $q_{\hat{\imath}} \geq 1/(K+J)$, the conditional defer regret of the plug-in action $\hat{\imath}$ is at most $(K + J)$ times this excess, and taking expectation preserves the linear bound. This statement concerns the MAE-like member, not every member of its surrogate family.

Summary. Write $\Delta\mathcal{E}$ for excess over the unrestricted population infimum of the indicated loss. Because the guarantees have different transfer forms, the table below displays each form together with its loss normalization:

Surrogate	Transfer to $\Delta\mathcal{E}_\perp$	Normalization
Add. CE	$\sqrt{2(J+1)}\sqrt{\Delta\mathcal{E}_{\mathrm{CE}}}$	summed CE
Mao25, linear member	$(K+J)\Delta\mathcal{E}_{\mathrm{Mao}}$	$1 - S_{\mathcal{S}}$
OvA	$\sqrt{2}\sqrt{\Delta\mathcal{E}_{\mathrm{OvA}}}$	unit Bernoulli weights
Decoupled	$\max\{\sqrt{2}, \sqrt{2J/\lambda}\}\sqrt{\Delta\mathcal{E}_{\mathrm{dec}}}$	expert weight λ/J

For PiCCE and A-SM, we use the cited Bayes-consistency results without asserting a new quantitative coefficient. The unrestricted transfer statements do not constitute restricted-class $\mathcal{H}$-consistency bounds; those require the relevant minimizability gaps and hypotheses.

J Further Consequences of the Decoupled Geometry

The following results develop the same two objects used in the main paper, the class posterior and the marginal expert utilities. They distinguish sharing a representation from sharing an action normalizer, and explain why the utility-based objective is compatible with either training schedule. All logarithms are natural, conditional identities hold P_X-almost surely for a version of the conditional probabilities, and ties are broken by a fixed measurable rule. Log losses on closed probability domains use the convention $0 \log 0 = 0$, while a positive target assigned zero estimated probability has infinite loss, and boundary optima need not be attained by finite logits.

J.1 Shared Trunk, Separate Head Parameters

Write $F = \Phi_{\mathrm{cls}}$ and $G = \Phi_{\mathrm{exp}}$, including the expert weight in G. Suppose $w = w(\theta, \theta_c) \in \mathbb{R}^K$ and $s = s(\theta, \theta_e) \in \mathbb{R}^J$ are twice differentiable. The trunk parameters θ are shared, while w does not depend on θ_e and s does not depend on θ_c, so the loss is exactly

$$\Phi(\theta, \theta_c, \theta_e) = F(w(\theta, \theta_c)) + G(s(\theta, \theta_e)).$$

Consequently, its parameter Hessian has the form

$$\nabla^2\Phi = \begin{pmatrix} F_{\theta\theta} + G_{\theta\theta} & F_{\theta\theta_c} & G_{\theta\theta_e} \\ F_{\theta_c\theta} & F_{\theta_c\theta_c} & 0 \\ G_{\theta_e\theta} & 0 & G_{\theta_e\theta_e} \end{pmatrix},$$

where subscripts denote derivatives of the composed scalar losses. This proves the parameter-level part of Proposition 9: differentiating the class loss with respect to the expert-head parameters, or the expert loss with respect to the class-head parameters, gives zero, so the head–head block vanishes even when the trunk is learned.

For the trunk, the chain rule gives

$$\nabla_\theta\Phi = (D_\theta w)^\top(p - e_y) + \frac{\lambda}{J}(D_\theta s)^\top(u - t).$$

Each residual is the same residual its estimation task would supply on its own, and the loss introduces neither a shared class–expert normalizer nor an additional product of class and expert residuals. The two contributions can still have a negative inner product, and a trunk update from either task can change both score blocks. Thus direct loss coupling is removed while joint representation learning is retained. The initialization diagnostic in Appendix P.2 measures the remaining trunk interaction rather than assuming it vanishes.

J.2 OvA Also Has a Dimension-Independent Transfer

Proposition 20 (OvA surrogate-regret transfer)**.** *Let $v : \mathcal{X} \to [0,1]^K$ and $u : \mathcal{X} \to [0,1]^J$ be measurable OvA estimates, trained with unit weight on every class and expert Bernoulli log loss. Select a class maximizing v_k if $\max_k v_k \geq \max_j u_j$, and otherwise an expert maximizing u_j. Then*

$$\mathcal{E}_\perp(f_{v,u}) - \mathcal{E}_\perp^\star \leq \sqrt{2}\sqrt{\mathcal{E}_{\mathrm{OvA}}(v,u) - \mathcal{E}_{\mathrm{OvA}}^\star}.$$

The coefficient is independent of both K and J.

Proof Let $\delta_{\mathrm{OvA}}(x)$ denote the conditional OvA surrogate excess at x. Conditional expectation and the binary cross-entropy identity give

$$\delta_{\mathrm{OvA}}(x) = \sum_{k=1}^{K} \mathrm{KL}\big(\mathrm{Bern}(\eta_k(x)) \| \mathrm{Bern}(v_k(x))\big) + \sum_{j=1}^{J} \mathrm{KL}\big(\mathrm{Bern}(\alpha_j(x)) \| \mathrm{Bern}(u_j(x))\big).$$

Bernoulli Pinsker implies that each coordinate error is at most $\sqrt{\delta_{\mathrm{OvA}}(x)/2}$. The four-case comparison in (19) uses only the maxima of the class and expert estimates; it does not require the class estimates to sum to one. Therefore $\delta_\perp(x) \leq \sqrt{2\delta_{\mathrm{OvA}}(x)}$. Taking expectation and applying Jensen's inequality proves the claim. Infinite surrogate excess makes the inequality vacuous but valid. ■

The bound extends our proof technique to OvA, without asserting that the categorical constraint is necessary for Bayes consistency. OvA recovers the same class and expert probabilities at its unrestricted optimum, while the decoupled class head additionally enforces $\sum_k p_k = 1$ at every parameter value. This is a structural distinction between the function parameterizations, not a theorem that every softmax estimate is better calibrated than every OvA estimate.

J.3 Cross-Entropy Reductions and the Comp-Sum Extension

Proposition 4 gives the comp-sum geometry and its overlap dependence. We extend it here to nonnegative weights and then to arbitrary increasing comp-sum outer functions. Let $d \geq 2$, $z \in \mathbb{R}^d$, $q = \mathrm{softmax}(z)$, and let $v \in [0,\infty)^d$ be finite, score-independent action weights. Set $V = \sum_a v_a$ and $C(q) = \mathrm{Diag}(q) - qq^\top$. For $L_v(z) = -\sum_a v_a \log q_a$,

$$\nabla_z L_v = Vq - v, \qquad \nabla_z^2 L_v = VC(q). \tag{24}$$

Indeed, $L_v = -v^\top z + V \log \sum_b e^{z_b}$, so differentiating proves both identities. If $V > 0$, we also have $L_v = V\,\mathrm{CE}(v/V, q)$, and if $V = 0$, the loss and both derivatives are zero. The calculation applies to

any such weighted-CE surrogate, including bounded soft-utility targets and frozen-predictor routers; it does not assume that the weights are binary.

For Add. CE, define $\mathcal{T} = \{y\} \cup \{K + j : j \in \mathcal{J}(y, m)\}$ and $N = |\mathcal{T}| = 1 + |\mathcal{J}|$. Taking $v_a = \mathbf{1}\{a \in \mathcal{T}\}$ gives $L = N\,\mathrm{CE}(\mathbf{1}_{\mathcal{T}}/N, q)$. The normalized target places $1/N$ on the true class and $|\mathcal{J}|/N$ in total on the correct expert actions, and no incorrect class receives target mass. At the conditional optimum the total class mass is $1/(1 + \sum_j \alpha_j)$, and every raw augmented probability margin is contracted by this denominator. Renormalizing the class block recovers η exactly at the optimum, and the common denominator preserves the Bayes action ordering. Thus this augmented-action smoothing is compatible with Fisher consistency; it concerns the raw probability representation and sample weighting.

Now consider the comp-sum base family (Mao et al., 2023b, 2024d):

$$\ell_\tau(z, a) = \begin{cases} (q_a^{\tau-1} - 1)/(1 - \tau), & \tau \neq 1, \\ -\log q_a, & \tau = 1, \end{cases} \qquad \tau \geq 0.$$

For the same score-independent nonnegative weights, set $L_{\tau,v} = \sum_a v_a \ell_\tau(z, a)$, $b_a = v_a q_a^{\tau-1}$, and $B_\tau = \sum_a b_a$.

Proposition 21 (Exact weighted comp-sum geometry)**.** *For finite logits and every $\tau \geq 0$,*

$$\nabla_z L_{\tau,v} = B_\tau q - \sum_a b_a e_a, \tag{25}$$

$$\nabla_z^2 L_{\tau,v} = B_\tau C(q) - (\tau - 1) \sum_a b_a (q - e_a)(q - e_a)^\top. \tag{26}$$

When $v = \mathbf{1}_{\mathcal{T}}$, define $\bar{q}_\tau = N^{-1} \sum_{a \in \mathcal{T}} q_a^{\tau-1}$, $\omega_a = q_a^{\tau-1}/(N\bar{q}_\tau)$, $\bar{e}_\tau = \sum_{a \in \mathcal{T}} \omega_a e_a$, and $\bar{R}_\tau = \sum_{a \in \mathcal{T}} \omega_a (q - e_a)(q - e_a)^\top$. Then

$$\nabla L_{\tau,v} = N\bar{q}_\tau (q - \bar{e}_\tau), \qquad \nabla^2 L_{\tau,v} = N\bar{q}_\tau [C(q) - (\tau - 1)\bar{R}_\tau].$$

Proof Since $\nabla \log q_a = e_a - q$, differentiating the base loss gives $\nabla \ell_\tau = q_a^{\tau-1}(q - e_a)$, including its continuous limit at $\tau = 1$. Differentiate this product, using $\nabla q = C(q)$ and $\nabla q_a^{\tau-1} = (\tau - 1) q_a^{\tau-1}(e_a - q)$. The result is

$$\nabla^2 \ell_\tau = q_a^{\tau-1} [C(q) - (\tau - 1)(q - e_a)(q - e_a)^\top].$$

Summing with weights v_a proves the first two identities; dividing the active weights by their positive sum proves the factorization. ∎

Every omitted action k has derivative $B_\tau q_k > 0$ when some weight is positive, so for every member of this family an omitted correct expert is suppressed by a winner-take-all selection on every region with a fixed winner (at score ties, these are selected-branch derivatives). At $\tau = 1$, $B_1 = V$ is score-independent and the correction term vanishes, recovering the exact weighted-CE result. For $0 \leq \tau \leq 1$, both Hessian contributions are positive semidefinite, whereas for $\tau > 1$ the correction is negative semidefinite, so positive curvature accumulation is no longer guaranteed. For example, with $d = 2$, $v = e_1$, and $\tau > 1$, the Hessian is

$$q_1^{\tau-1}(1 - q_1)[\tau q_1 - (\tau - 1)] \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}.$$

Its contrast eigenvalue is negative whenever $0 < q_1 < (\tau - 1)/\tau$, explicitly exhibiting negative curvature. At $\tau = 2$ and binary active weights, $L_{2,v} = N - \sum_{a \in \mathcal{T}} q_a$ and $B_2 = \sum_{a \in \mathcal{T}} q_a \leq 1$. Thus writing $B_\tau = N\bar{q}_\tau$ exposes the active-target count without implying global linear amplification for every τ. The MAE endpoint is precisely the set-mass geometry studied for Mao25, up to the score-independent additive constant $N - 1$.

Arbitrary increasing comp-sum outer functions. The competition mechanism also extends beyond the one-parameter family. Let $\ell_\Upsilon(z, i) = \Upsilon(\sum_{r \neq i} e^{z_r - z_i})$, where Υ is differentiable and $\Upsilon'(\nu) > 0$ for $\nu > 0$. Writing $\nu_i = 1/q_i - 1$ gives

$$\nabla_z \ell_\Upsilon(z, i) = \frac{\Upsilon'(\nu_i)}{q_i}(q - e_i).$$

For score-independent nonnegative weights v, define $b_i = v_i \Upsilon'(\nu_i)/q_i$ and $B = \sum_i b_i$. Then $\nabla_z L_{\Phi,v} = Bq - \sum_i b_i e_i$. If at least one weight is positive, an omitted coordinate k has $\partial L_{\Phi,v}/\partial z_k = Bq_k > 0$. Thus suppression of omitted actions follows from increasing comp-sum structure rather than from a logarithmic outer function; we do not make this sign claim for arbitrary multiclass base losses.

J.4 Two-Stage Training and Predictor-Dependent Routers

With disjoint parameters (θ_p, θ_u) in a product feasible set, write the empirical objective as

$$\widehat{\mathcal{E}}_{\mathrm{dec}}(\theta_p, \theta_u) = \widehat{\mathcal{E}}_{\mathrm{cls}}(\theta_p) + \lambda \widehat{\mathcal{E}}_{\mathrm{exp}}(\theta_u),$$

where the expert term is mean BCE over examples and experts. Since $\lambda > 0$, a pair minimizes this sum if and only if each component minimizes its summand, whenever finite component minima exist. Sequential training is therefore an alternative implementation of the same separable learning objective. Parameter constraints or regularizers that couple the blocks would invalidate this product argument.

For any fixed measurable classifier h, let $a_h = \eta_{h(x)}(x)$, $\eta_{\max} = \max_k \eta_k(x)$, and $\alpha_{\max} = \max_j \alpha_j(x)$. Its best achievable conditional defer risk is $1 - \max\{a_h, \alpha_{\max}\}$, so freezing h leaves the exact gap

$$\delta_h(x) = \max\{\eta_{\max}, \alpha_{\max}\} - \max\{a_h, \alpha_{\max}\} \geq 0, \tag{27}$$

which follows by minimizing over the $J+1$ available action risks. The gap is zero when h is Bayes or an expert is at least as accurate as the best class action. It applies to any frozen-predictor router, including a two-stage implementation of our own model; it is not a claim that two-stage learning is necessarily inferior.

The 2023 fixed-score router. In the score-based construction of Mao et al. (2023a), fix $z_0(x) = \max_k w_k^p(x)$ and the induced classifier h_p, and learn $z_1, \ldots, z_J$. For the logistic instance, $q = \mathrm{softmax}(z_0, \ldots, z_J)$ and $L_{23} = -t_0 \log q_0 - \sum_j t_j \log q_j$, where $t_0 = \mathbf{1}\{h_p(x) = y\}$ and $t_j = \mathbf{1}\{m_j = y\}$. Equation (24) gives, for $i, j \in [J]$,

$$\frac{\partial L_{23}}{\partial z_j} = (t_0 + |\mathcal{J}|)q_j - t_j, \qquad \frac{\partial^2 L_{23}}{\partial z_i \partial z_j} = (t_0 + |\mathcal{J}|)(q_i \mathbf{1}\{i = j\} - q_i q_j).$$

With $\beta_0 = a_{h_p}$, $\beta_j = \alpha_j$, and $Z = \sum_{a=0}^{J} \beta_a > 0$, conditional excess is $Z\mathrm{KL}(\beta/Z \| q)$, since conditional expectation of the loss is $-\sum_a \beta_a \log q_a$. The optimum in the closed simplex is $q_a^\star = \beta_a/Z$; zero weights are interpreted by limits. If $Z = 0$, every action has zero correctness utility and all router scores minimize the conditional surrogate. Adding a positive-utility expert changes the normalization of every existing optimum, whereas an existing Bernoulli utility target remains α_j. Shifting every frozen class logit by the same function leaves the class softmax and prediction unchanged but shifts the fixed router coordinate z_0 alone. An unrestricted router can compensate, whereas a restricted router need not; our probability-space comparison is invariant to this shift.

The predictor–rejector formulation used in our experiments. The regression/cost-sensitive construction of Mao et al. (2024f) uses a learned score for every router action. With $\ell_h = \mathbf{1}\{h(x) \neq y\}$, $c_j = \mathbf{1}\{m_j \neq y\}$, and $C = \sum_j c_j$, its logistic action weights are $v_0 = C$ and $v_j = \ell_h + C - c_j$. Their sum is $V = J(\ell_h + C)$, so the same weighted-CE calculation gives gradient $Vq - v$ and Hessian $VC(q)$. At the population level, let $\rho_0 = 1 - a_h$, $\rho_j = 1 - \alpha_j$, and $T = \sum_{a=0}^{J} \rho_a$. When $T > 0$,

$$q_a^\star = \frac{T - \rho_a}{JT}.$$

This optimum orders actions by increasing conditional risk and therefore selects the frozen-predictor Bayes action, although $q_a^\star$ is not the action utility $1 - \rho_a$. For positive optimal weights and $\rho_a < \rho_b$,

$$0 < z_a^\star - z_b^\star = \log \frac{T - \rho_a}{T - \rho_b} \leq \frac{\rho_b - \rho_a}{T - \rho_b},$$

using $\log(1 + r) \leq r$. If $T \geq cJ$ for a fixed $c > 0$ and the action risks lie in $[0, 1]$, this margin is $O(1/J)$ as $J \to \infty$ for a fixed action-risk gap. For two expert actions, the Bernoulli optimum instead preserves the utility difference $\alpha_a - \alpha_b$. When $T = 0$, every action has zero error and the surrogate loss is identically zero. These output-level identities complement the restricted-class and realizable-consistency guarantees of the two-stage literature; our unrestricted theorem does not subsume those guarantees.

J.5 Consultation Costs and Class-Weighted Risk

For any measurable expert cost $c_j \in [0, B]$, with fixed $B \geq 1$, train the expert target $\tilde{t}_j = 1 - c_j/B$. By the same binary cross-entropy decomposition, its conditional optimum is $u_j^\star = 1 - \mathbb{E}[c_j \mid X = x]/B$. Class action k has normalized utility $1-(1-\eta_k)/B$, so routing compares $(B-1+\max_k p_k)/B$ with $\max_j u_j$. For $c_j = b_j + \mathbf{1}\{m_j \neq y\}$ this is, at the optimum, exactly $\max_k \eta_k \geq \max_j(\alpha_j - b_j)$. The bound transfers with an overall factor B to the raw cost risk: normalized action-utility errors satisfy the same four-case bound, the class utility error is at most $\|p - \eta\|_\infty/B$, and the expert error is controlled by its Bernoulli KL. Applying the original bound's inequalities and Jensen gives

$$\mathcal{E}_{\text{cost}} - \mathcal{E}^\star_{\text{cost}} \leq B \max\{\sqrt{2}, \sqrt{2J/\lambda}\}\sqrt{\mathcal{E}_{\text{dec,cs}} - \mathcal{E}^\star_{\text{dec,cs}}}.$$

For $B = 1$, this is the bounded-cost version of the original coefficient.

For class weights $\kappa_y > 0$, set $W = \max_y \kappa_y$ and $g_j(x) = \mathbb{E}[\kappa_Y \mathbf{1}\{M_j = Y\} \mid X = x]$. The weighted error of class k is $\sum_y \kappa_y \eta_y - \kappa_k \eta_k$, while that of expert j is $\sum_y \kappa_y \eta_y - g_j$, and the common first term cancels in the decision. We keep ordinary class CE, train the expert target $(\kappa_y/W)\mathbf{1}\{m_j = y\}$, and compare $\max_k \kappa_k p_k/W$ with $\max_j u_j$, predicting $\arg\max_k \kappa_k p_k$ when the class action wins. The expert optimum is g_j/W. Since $\kappa_k/W \leq 1$, the same coordinate-error argument yields the original coefficient for risk normalized by W, and a factor W for raw weighted risk. Balanced error takes $\kappa_y = (K\pi_y)^{-1}$ with $\pi_y = \Pr(Y = y) > 0$ for every class. The weights are assumed known in this population statement; estimating them introduces an additional error not analyzed here.

For binary labels, encoded as $\{0, 1\}$ in this paragraph, a fitted posterior also supplies predictions $\widehat{y}_a(x) = \mathbf{1}\{p_1(x) \geq a\}$. Comparing $p_{\widehat{y}_a(x)}$ with $\max_j u_j + t$ defines a family of routing policies. Evaluating their realized outputs while varying a at fixed t produces an empirical system precision–recall curve, with coverage allowed to change. This is an evaluation procedure, not a consistency claim for constrained precision–recall objectives.

K Compute Resources

Unless stated otherwise, experiments used NVIDIA A100 GPUs with 40GB of memory, and the Fashion-MNIST studies also used NVIDIA H100 and H200 GPUs. Each training run uses a single GPU, and seed replication repeats the specified configuration. The synthetic suites finish within minutes per seed, and each real-data benchmark (CIFAR-10, CIFAR-10H, Covertype, and Fashion-MNIST) takes about one hour on a single NVIDIA H100 GPU.

L Synthetic Experiments

This appendix describes the 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 (η, α) 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 $\Delta_\perp(f) := \mathcal{E}_\perp(f) - \mathcal{E}^\star_\perp$ 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 $\alpha(x)$ never enter the training losses: they are used *after* training to evaluate the exact Bayes regret and, on the validation split, to select the checkpoint with the lowest true defer loss, a model-selection rule applied identically to every surrogate. All reported numbers are means over 3 to 5 seeds (stated per suite), with one standard deviation in parentheses.

L.1 Nested Redundant Experts

Objective. We vary the overlap multiplicity while keeping the Bayes decision problem fixed, so that any performance degradation is attributable to the surrogate itself rather than to a harder task.

Although the suite is designed around Proposition 4 (Add. CE amplification), the nesting also stresses every other baseline, for surrogate-specific reasons discussed below.

Setting. We use K=16 classes, J=24 experts, n_{train}=900, n_{test}=8,000, and linear models (no hidden layer).

Features and class posteriors. Each point draws a region indicator $R \sim \text{Bernoulli}(0.65)$, and 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 $j \in \{0, \ldots, J-1\}$ and $J \geq 2$, with $\rho_0 := 0$ when $J = 1$ (logarithmic spacing, so quality drops fastest over the first few experts and the weaker experts cluster near the lowest accuracy). On D,

$$\alpha_j(x) \;=\; 0.99 \;-\; (0.99 - 0.75)\,\rho_j, \tag{28}$$

giving α_0=0.99 for the strongest expert down to α_{J-1}=0.75 for the weakest. On D^c, all experts are almost always wrong: $\alpha_j(x) = 0.04 - (0.04 - 0.002)\,\rho_j$.

Nested correctness. A single latent $U \sim \text{Unif}[0,1]$ is drawn per sample, and expert j is correct if and only if $U \leq \alpha_j(x)$. Because $\alpha_0 \geq \alpha_1 \geq \cdots \geq \alpha_{J-1}$, 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 $[K] \setminus \{Y\}$. Since $\alpha_0 = 0.99$ for every J and $\alpha_j < \alpha_0$ for $j \geq 1$, changing J never changes the Bayes action: expert 0 is always Bayes-best on D, classification is always best on D^c, and only the realized overlap $|\mathcal{J}(Y, \mathbf{M})|$ grows with J.

Bayes rule. For every J, the Bayes rule defers to expert 0 on D and classifies on D^c.

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

- **Add. CE** (Proposition 4). The gradient and the Hessian both carry the factor $1+|\mathcal{J}|$. With J=24 nested experts, all with $\alpha_j \geq 0.75$ on D, every expert is correct on 75% of defer-region samples and $\mathbb{E}[|\mathcal{J}| \mid D] = \sum_j \alpha_j \approx 19.6$, so a defer-region sample is weighted about 20× (and 25× when all experts are correct), compared with roughly 1× for 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 and inject noise into the class update. Because the leakage enters the class-logit gradient itself, it perturbs the class weights of the linear model and hence 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 $S_{\mathcal{S}} \to 0$, which slows the learning of the defer region.

Result. Table 4 reports the results at $J{=}24$, and Table 5 summarizes a sweep over $J \in \{16, 24, 32\}$.

Table 4: Nested redundant suite at $J{=}24$. The decoupled surrogate remains essentially Bayes-optimal, while all baselines degrade severely. Means with standard deviations in parentheses over 3 seeds.

Method	Exact defer regret	System accuracy
Decoupled	$\mathbf{0.0002}$ (0.0001)	$\mathbf{0.993}$ (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 5: Summary of the J sweep in the nested redundant suite. The decoupled surrogate stays near Bayes-optimal as the expert pool grows, while the best non-decoupled regret increases monotonically.

	$J = 16$	$J = 24$	$J = 32$
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

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 α_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 at or below 0.013 throughout, consistent with the prediction that the baselines' failures scale with the overlap multiplicity.

L.2 Rare Specialist for PiCCE

Objective. We isolate the winner-take-all starvation of Proposition 5 with a specialist expert that is Bayes-preferred on a rare region R but whose correct events are mostly shared with a generalist that is correct more often overall. PiCCE should suppress the specialist because only the current $\arg\max$ winner receives the reinforcing gradient $2q_{K+j^\star}{-}1$ (negative while $q_{K+j^\star} < 1/2$), while every other correct expert is pushed down by $2q_{K+j} > 0$ (eq. (8)).

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

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, with $\alpha_{\mathrm{gen}}(x){=}0.45$ and $\alpha_{\mathrm{spec}}(x){=}0.15$, so 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. On R, correctness is *not* independent, and we specify the joint law directly:

	Specialist correct	Specialist incorrect
Generalist correct	0.55	0.05
Generalist incorrect	0.20	0.20

This gives α_{spec}=0.55+0.20=0.75 and α_{gen}=0.55+0.05=0.60 as required. When the specialist is correct on R, the generalist is also correct with probability

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

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

Bayes rule. The Bayes rule defers to the specialist on R (15% of the data), since $\alpha_{\text{spec}}{=}0.75 > \alpha_{\text{gen}}{=}0.60 = \eta_{\max}{=}0.60$, and classifies on R^c (85% of the data), since $\eta_{\max}{=}0.90 > \alpha_{\text{gen}}{=}0.45$.

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 through initialization noise, the specialist is suppressed on every shared-correct sample, and the gap compounds across iterations: the generalist wins more $\arg\max$ selections and receives more reinforcing gradient, while the specialist is driven down. On the 27% of specialist-correct events that are specialist-only ($C_{\text{gen}}{=}0$), the specialist does receive a negative gradient, but these events make up only 15% × 20%=3% of the training data, which 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 independently receive the reinforcing gradient $(\lambda/J)(u_j{-}1) < 0$. Since no $\arg\max$ selection occurs, the generalist's logit does not compete with the specialist's, each head targets its own marginal utility, 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, because both experts are right and only the ranking $\alpha_{\text{spec}} > \alpha_{\text{gen}}$ distinguishes them.

Result. Table 6 reports the results (means with standard deviations in parentheses over 5 seeds).

Table 6: Rare-specialist suite. *Coverage* is $\Pr(r(X){=}\text{classify})$, and the Bayes-optimal coverage is 0.85. PiCCE defers but never to the specialist, Mao25 never defers at all, and the decoupled surrogate, OvA, and A-SM recover the specialist almost perfectly. Add. CE reaches 1.000 on both specialist metrics only because it defers on 87% of inputs, so no entry in those columns is bold.

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)

Analysis. The table shows three distinct failure modes, one each for PiCCE, Mao25, and Add. CE.

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 behavior matches the mechanism of Proposition 5: the generalist captures

the $\arg\max$ early, and the positive gradient $2q_{K+\text{spec}} > 0$ keeps the specialist suppressed for the rest of training.

Mao25 (no deferral learned). Mao25 has coverage 1.0, so it *never* defers and its specialist selection is trivially 0. Its system accuracy of 0.852 is close to that 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 of Proposition 6: the corrective gradient on acceptable actions has total magnitude $S_{\mathcal{S}}(1-S_{\mathcal{S}}) \leq 1/4$, which vanishes both as $S_{\mathcal{S}} \to 0$ and as $S_{\mathcal{S}} \to 1$, so Mao25 receives little signal to move away from the classify-everything solution.

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 a plausible cause: the factor $1{+}|\mathcal{J}|$ reaches 3 on shared-correct samples, which is consistent with the over-deferral observed here.

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. With only $J{=}2$ experts, the A-SM coupling (Proposition 7) and the OvA miscalibration are negligible here, which indicates that the PiCCE failure comes from the winner-take-all mechanism rather than from a general difficulty of the task.

L.3 Shared Acceptability and Expert Ranking for Mao25

Objective. We 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 therefore construct a task in which the main difficulty is expert ranking *within* a defer region under heavy shared acceptability.

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

Features and class posteriors. Each point draws a defer-region indicator $D \sim \text{Bernoulli}(0.60)$, and 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, and 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

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

while on D^c all experts have utility $\alpha_j{=}0.05$.

The design has two properties that the test relies on. 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. The Bayes rule defers to the sector-best expert Q on D (60% of the data), since $\alpha_Q{=}0.93 \gg \eta_{\max}{=}0.10$, and classifies on D^c (40% of the data), since $\eta_Y{=}0.998 \gg \alpha_j{=}0.05$. The

resulting Bayes-optimal system accuracy is $0.60 \times 0.93 + 0.40 \times 0.998 = 0.957$, with 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). The samplewise loss therefore 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, so we expect a degraded but non-trivial ranking, which we measure 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 ≈ 0.48 ($0.10 \times \frac{1}{4} + 0.78 \times \frac{1}{2} + 0.05 \times 1 + 0.03 \times 0 + 0.04 \times \frac{1}{4}$), and perfect ranking gives 1.0.

Result. Table 7 reports the results (means with standard deviations in parentheses over 3 seeds).

Table 7: Shared-acceptability ranking suite. Chance-level best-expert selection is 0.25, and uniform selection among correct experts gives 0.48. Mao25 reaches 0.750, well above chance but far below the 1.000 achieved by the decoupled surrogate and three other baselines, consistent with a partial but degraded ranking signal.

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

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 α_j accurately enough that the routing comparison $\max_j \sigma(s_j) \gtrless \max_k p_k$ places both the classify-versus-defer boundary and the expert ranking correctly.

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

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 regrets (0.064 and 0.082) are still 80 to 100 times that of the decoupled surrogate. The ranking is correct, but the classify-versus-defer boundary is less well placed: A-SM coverage is 0.330 against the optimal 0.40, which indicates over-deferral and, for A-SM, is consistent with the class–expert coupling of 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: on the 88% of defer-region samples with shared acceptability, $|\mathcal{J}|$ equals 2 or 4, which inflates the gradient by a factor of 3 or 5.

L.4 Logit-Space Geometry Diagnostics

The task-level suites above measure end-to-end performance; this section complements them with *direct measurements of the gradients and Hessians* 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 that 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 (Corollary 4.1). Corollary 4.1 shows that the Hessian of the Add. CE surrogate Φ^{CE} has largest eigenvalue $\lambda_{\max} \leq (1+|\mathcal{J}|)/2$, an upper bound linear in the correct-expert count $|\mathcal{J}|$. The decoupled surrogate, by contrast, has $\lambda_{\max} \leq 1/2$ for the class block and expert entries at most $\lambda/(4J)$, independently of $|\mathcal{J}|$ (see (13)).

Setup. We use the nested redundant-expert construction (Appendix L.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_{train}=3,500 samples and evaluate on n_{test}=6,000. We restrict to test samples on which at least one expert is correct ($|\mathcal{J}| > 0$), and for each such sample we compute (i) the gradient norm $\|\nabla_a \Phi\|_2$ and (ii) the top Hessian eigenvalue $\lambda_{\max}(\nabla_a^2 \Phi)$ from the full $(K+J) \times (K+J)$ Hessian. Results are pooled over $\sim$900 qualifying samples per (J, method) combination across 5 independent seeds.

Table 8: Per-sample curvature as the pool size J varies (5 seeds, about 900 pooled samples per row). The Add. CE top eigenvalue rises from 0.82 at J=1 to 1.09 at J=4 and then saturates near 1.12, far below the bound $(1+|\mathcal{J}|)/2$ of Corollary 4.1, because the softmax mass spreads over $K+J$ actions. The decoupled surrogate remains flat at $\approx$0.36 for all $|\mathcal{J}|$, consistent with the constant bound $1/2$ from (13).

Method	J	Gradient norm $\|\nabla_a \Phi\|_2$	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)

Reading the table. The Add. CE eigenvalue jumps from 0.82 to 1.09 as J goes from 1 to 4, then saturates near 1.12 for larger J. The saturation is expected: at J=32 the theoretical bound is $(1 + 32)/2 = 16.5$, but the *realized* eigenvalue is much smaller because the softmax probabilities q_i become very small when spread over K+32 actions. The relevant contrast is between this sharp increase and saturation for Add. CE and the decoupled eigenvalue, which stays at about 0.36 for every $|\mathcal{J}|$.

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* even though the expert is 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 satisfies $u_j < 1$.

Setup. We use the PiCCE specialist task (Appendix L.2) with K=2 classes and J=2 experts, where samples routinely have both experts correct. For each method (PiCCE and the decoupled surrogate, λ=1.0) we train on n_{train}=4,500 and evaluate on n_{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, which has no winner selection, we measure the gradient on *all* correct expert

logits. Results pool 784 decoupled-surrogate gradient samples and 392 PiCCE gradient samples across 5 seeds.

Table 9: Gradient on correct expert logits (5 seeds; 784 decoupled-surrogate and 392 PiCCE measurements). For PiCCE, only the suppressed (non-winning) correct expert is measured. 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.

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

Reading the table. The positive-gradient rate is 1.000 for PiCCE and 0.000 for the decoupled surrogate, with zero variance: 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 but an active repulsive signal of the same scale as the decoupled surrogate's attractive one.

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 $H_{\text{mix}} = \partial^2\Phi^{\text{ASM}}/(\partial a_{1:K}\, \partial a_{K+1:K+J})$, whereas the decoupled surrogate's Hessian is block-diagonal by (13), 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 (λ=1.0) on n_{train}=4,500 and evaluate on n_{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_{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, since the class logits w and expert logits s_j enter disjoint terms of $\Phi^{\text{dec}}_\lambda$; we therefore record the value 0 directly rather than computing it numerically.

Table 10: Operator norm of the mixed class–expert Hessian block (5 seeds, 128 samples each). The A-SM coupling grows monotonically from 0.12 at J=1 to 0.48 at J=17, consistent with the $O(\sqrt{J})$ bound of Appendix G.4. The decoupled mixed block is zero for every J by the block-diagonal structure of (13).

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)

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, close to $0.12\sqrt{17} \approx 0.49$, which is consistent with the theoretical $O(\sqrt{J})$ bound derived in Appendix G.4. Adding experts, even *distractor* experts that are never correct, therefore injects progressively more expert-side noise into the class gradient. The mixed block of the decoupled score Hessian remains zero at every pool size, because expert residuals do not enter the class-score derivative. With a learned shared trunk, the parameter interaction is the additive one described in Appendix J.1.

M CIFAR-10 with Synthetic Experts

The synthetic experiments of Appendix L use linear models on low-dimensional Gaussian data, which isolates surrogate pathologies in a setting where the Bayes solution is known. This appendix tests whether the same effects persist on a real vision task.

Setup. We train a ResNet-12 backbone end-to-end, without pre-training, on CIFAR-10 ($K{=}10$ classes) with standard data augmentation (random crop, horizontal flip). All methods share the same backbone architecture and optimizer (AdamW, lr 10^{-3}, 80 epochs). All reported metrics are computed on the 10 000-image CIFAR-10 test set. A held-out validation split (5 000 training samples) is used to select the best checkpoint (lowest validation deferral loss) and, for the decoupled surrogate, to choose the total expert weight $\lambda = 1/2$, so that the per-expert weight is $1/(2J)$. For the decoupled surrogate, the linear head is a *split head* made of 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 structure of the decoupled loss, whereas all other methods use a single $\mathbb{R}^{256} \to \mathbb{R}^{K+J}$ head; the total number of parameters is therefore the same across methods.

Redundant expert suite. We construct J synthetic experts with *nested* correctness (Appendix L.1): a single latent $U \sim \text{Uniform}(0, 1)$ per sample determines all expert outcomes, with expert j correct if and only if $U < \alpha_j$, where $\alpha_j = 0.70 - 0.01\, j$. The first 8 experts in a $J{=}32$ run are therefore *identical* to those in a $J{=}8$ run, so performance changes across J can be attributed to the surrogate's response to the 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).

		Test 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	$\mathbf{0.919}_{\pm.002}$	$\mathbf{0.081}_{\pm.002}$	$\mathbf{0.908}_{\pm.003}$	$0.957_{\pm.005}$
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	$0.859_{\pm.008}$	$0.141_{\pm.008}$	$\underline{0.880}_{\pm.014}$	$0.636_{\pm.027}$
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 whose system accuracy exceeds that of a standalone classifier.* At every J, the decoupled surrogate achieves higher system accuracy than the classifier-only baseline, by $+0.8$ pp at $J{=}8$, $+0.8$ pp at $J{=}16$, and $+1.1$ pp at $J{=}32$. Every other L2D surrogate falls *below* the classifier-only baseline, which is consistent with the redundancy effects of Appendix L.1 persisting beyond linear models.

The decoupled surrogate preserves classifier quality, whereas several baselines degrade it. The decoupled surrogate keeps 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 reaching the class head through the class log-normalizer inside the expert sigmoid (Proposition 7), and OvA also degrades (87.5% → 84.4%). In the decoupled surrogate, the

expert BCE term has no direct derivative with respect to the class-head parameters, although both tasks can still update the learned representation.

For most baselines, deferral lowers accuracy. When system accuracy falls *below* classifier accuracy, the routing decisions cost accuracy relative to never deferring. At J=8, PiCCE achieves 88.8% classifier accuracy but only 85.8% system accuracy (-3.0 pp from deferral), and Add. CE similarly loses 8.7 pp. Both route a substantial share of samples (PiCCE 38.4%, Add. CE 47.1%) to experts with $\approx$70% accuracy, well below that of 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), consistent with the gradient amplification of Proposition 4: the $(1+|\mathcal{J}|)$ factor grows with the redundant pool while coverage falls to 30.5%. A-SM shows a similar pattern ($86.0\% \to 78.3\%$), consistent with the class–expert coupling of Proposition 7.

Mao25 fails to learn a classifier. Across all J, Mao25 ends at 100% deferral and reaches system accuracy $\approx 70.3\%$ from expert answers alone, which is consistent with the $O(1/(K+J))$ gradient magnitude through the softmax (Appendix F.3).

The decoupled surrogate is stable in J. The system accuracy of the decoupled surrogate stays at 91.8–91.9% from J=8 to J=32, a change of only 0.1 pp, compared with 2.1 pp for OvA and 10.3 pp for Add. CE. This stability on the present expert panel is consistent with Proposition 9: each expert derivative depends only on that expert's own score and target, and the classifier score gradient does not depend on the experts at all, so adding redundant experts does not rescale the class update.

N CIFAR-10H with Real Human Annotators

The synthetic-expert experiments above control the expert accuracy profile exactly, but they leave open 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 2,571 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 5×10^{-4}) and then *freeze* it, so that 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 training, 2 000 validation, and 2 000 test images, and all methods share the same pretrained backbone and data split.

Annotator selection. We retain workers with at least 200 annotations and sample J observed responses without replacement for each image. The worker identity assigned to an expert coordinate can therefore vary across images, and no response is imputed. The single-seed runs use training seed 0 for each of five response draws (0–4), and the three-seed study reruns training seeds 0, 1, and 2 on the same draws, so its seed-0 runs are separate runs. Table 12 reports the single-seed runs, while Table 3 in the main text and Appendix P.6 report the three-seed study. All methods use AdamW (lr 10^{-3}, 30 epochs). For the decoupled surrogate, the total expert weight is $\lambda = 2.0$ and the weight decay is 10^{-3}; the other methods use weight decay 10^{-4}. The three-seed study in Appendix P.6 holds these settings fixed.

Analysis. *The decoupled surrogate leads at every pool size.* At J=5 the rounded margin is narrow (0.960 vs. 0.959 for OvA), but the gap to OvA widens as the pool grows: at J=20 the decoupled surrogate reaches 0.961 while OvA drops to 0.956 ($+0.5$ pp). As with the synthetic experts of Table 11, the decoupled surrogate keeps a stable accuracy in these single-seed runs, and its per-expert score derivative has no overlap-count factor, with a magnitude controlled by λ/J.

The decoupled surrogate preserves classifier quality in these runs, whereas several baselines collapse. The classifier accuracy of the decoupled surrogate stays at 0.889–0.892 across all J, matching the standalone classifier baseline (0.888–0.889). A-SM maintains 0.890 at J=5 but collapses to 0.470 at J=20, with a standard deviation of 0.435 that indicates highly unstable training across draws; Add. CE drops from 0.889 to 0.742 over the same range, and OvA degrades to 0.799 (std 0.203) at J=10. These methods compensate by deferring more often (A-SM reaches $\geq$99% deferral at J=20), so their system accuracy is sustained by expert answers rather than by the classifier. The collapses are consistent with the class–expert coupling (Proposition 7) and gradient amplification (Proposition 4)

Table 12: CIFAR-10H with one training seed (seed 0): test metrics, mean ± sample SD over five annotation draws. Table 3 uses three training seeds.

Method	J	System acc.	Defer loss	Classifier acc.	Coverage
Decoupled	5	$0.960_{\pm 0.005}$	$0.041_{\pm 0.005}$	$0.892_{\pm 0.001}$	$0.517_{\pm 0.036}$
OvA	5	$0.959_{\pm 0.005}$	$0.041_{\pm 0.005}$	$0.889_{\pm 0.001}$	$0.375_{\pm 0.016}$
A-SM	5	$0.959_{\pm 0.004}$	$0.042_{\pm 0.004}$	$0.890_{\pm 0.002}$	$0.341_{\pm 0.035}$
Add. CE	5	$0.953_{\pm 0.004}$	$0.047_{\pm 0.004}$	$0.889_{\pm 0.002}$	$0.248_{\pm 0.037}$
PiCCE	5	$0.954_{\pm 0.003}$	$0.046_{\pm 0.003}$	$0.889_{\pm 0.001}$	$0.390_{\pm 0.020}$
Classifier only	5	$0.889_{\pm 0.000}$	$0.111_{\pm 0.000}$	$0.889_{\pm 0.000}$	$1.000_{\pm 0.000}$
Mao25 (linear)	5	$0.950_{\pm 0.005}$	$0.050_{\pm 0.005}$	$0.116_{\pm 0.133}$	$0.000_{\pm 0.000}$
Decoupled	10	$0.958_{\pm 0.003}$	$0.042_{\pm 0.003}$	$0.892_{\pm 0.003}$	$0.517_{\pm 0.038}$
OvA	10	$0.954_{\pm 0.007}$	$0.046_{\pm 0.007}$	$0.799_{\pm 0.203}$	$0.285_{\pm 0.164}$
A-SM	10	$0.952_{\pm 0.007}$	$0.048_{\pm 0.007}$	$0.747_{\pm 0.313}$	$0.170_{\pm 0.134}$
Add. CE	10	$0.952_{\pm 0.004}$	$0.048_{\pm 0.004}$	$0.888_{\pm 0.001}$	$0.248_{\pm 0.004}$
PiCCE	10	$0.957_{\pm 0.004}$	$0.043_{\pm 0.004}$	$0.890_{\pm 0.002}$	$0.400_{\pm 0.006}$
Classifier only	10	$0.889_{\pm 0.000}$	$0.112_{\pm 0.000}$	$0.889_{\pm 0.000}$	$1.000_{\pm 0.000}$
Mao25 (linear)	10	$0.948_{\pm 0.003}$	$0.052_{\pm 0.003}$	$0.077_{\pm 0.024}$	$0.000_{\pm 0.000}$
Decoupled	20	$0.961_{\pm 0.004}$	$0.040_{\pm 0.004}$	$0.889_{\pm 0.001}$	$0.530_{\pm 0.038}$
OvA	20	$0.956_{\pm 0.005}$	$0.044_{\pm 0.005}$	$0.890_{\pm 0.001}$	$0.335_{\pm 0.047}$
A-SM	20	$0.949_{\pm 0.005}$	$0.051_{\pm 0.005}$	$0.470_{\pm 0.435}$	$0.002_{\pm 0.004}$
Add. CE	20	$0.951_{\pm 0.003}$	$0.049_{\pm 0.003}$	$0.742_{\pm 0.331}$	$0.183_{\pm 0.103}$
PiCCE	20	$0.953_{\pm 0.002}$	$0.047_{\pm 0.002}$	$0.890_{\pm 0.001}$	$0.399_{\pm 0.022}$
Classifier only	20	$0.888_{\pm 0.000}$	$0.113_{\pm 0.000}$	$0.888_{\pm 0.000}$	$1.000_{\pm 0.000}$
Mao25 (linear)	20	$0.948_{\pm 0.004}$	$0.053_{\pm 0.004}$	$0.045_{\pm 0.014}$	$0.000_{\pm 0.000}$

seen in the synthetic settings. The replication in Appendix P.6 does not, however, reproduce the Add. CE collapse: across its 45 runs, the Add. CE classifier accuracy stays within 0.8850–0.8925.

The decoupled surrogate keeps a stable coverage. The coverage of the decoupled surrogate stays at ≈ 0.52, meaning that it classifies about half the samples itself and defers the rest; this routing yields a system accuracy above both the standalone classifier and the full-deferral Mao25 system.

Mao25 defers everything. As on CIFAR-10 with synthetic experts, Mao25 converges to 100% deferral and never learns a classifier (mean classifier accuracy at most 0.116). Its system accuracy (≈ 0.950) therefore comes entirely from the selected expert answers, which is consistent with the set-mass pathology of Proposition 6 on real human-annotator data.

O Covertype with Model Experts

The preceding experiments use either synthetic accuracy profiles (Appendix M) or real human annotators (Appendix N). This benchmark replaces both with *pretrained models*: the experts are machine-learning classifiers with different architectures, feature views, and training biases, as in model routing, where a system sends each input to the most reliable model in a pool.

Setup. We use Forest CoverType (Blackard and 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, and only the continuous features are standardized, using training-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), so routing among the experts is nontrivial. The data split and the expert-pool training seed are held fixed, and only the downstream L2D training seed changes across runs.

Protocol. All methods use MLP heads with one 64-unit hidden layer and are trained for 50 epochs with AdamW, cosine learning-rate decay, learning rate 10^{-3}, and weight decay 10^{-4}. The decoupled model uses separate class and expert heads, and it selects the total expert weight λ from $\{0.25, 0.5, 1, 2, 4\}$ by validation deferral loss; the selected total weights for seeds 0–3 are $0.5, 1, 0.25, 1$, respectively. For every method, checkpoints are selected by the lowest validation true deferral loss.

Table 13: Covertype with model experts (J=5, mean±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 whose system accuracy exceeds that of 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 falls below the standalone classifier, so its routing decisions cost accuracy relative to never deferring.

The decoupled surrogate preserves and even improves classifier quality. The decoupled surrogate achieves the highest classifier accuracy (0.941), surpassing the standalone classifier (0.929), which shows that joint utility learning is compatible with a strong class predictor in this experiment. By contrast, the classifier of Mao25 collapses to 0.351 and that of A-SM degrades to 0.892.

The decoupled surrogate has 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 defer more often, so a larger share of their system accuracy comes from the experts rather than from their own classifiers.

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, consistent with the vanishing attractive gradient of Proposition 6.

The ordering extends to model experts. As on CIFAR-10 and in the fixed-draw CIFAR-10H table, the decoupled surrogate ranks first and Mao25 last among L2D methods on this tabular benchmark with pretrained model experts, with the ordering decoupled $>$ PiCCE $\approx$ OvA $>$ A-SM $\approx$ Add. CE $>$ Mao25. These results extend the comparison beyond vision backbones and human annotators to a pool of heterogeneous model experts.

P Controlled Ablations and Replication

The experiments in this appendix separate the loss from its implementation and test the extensions derived in Appendix J. Fashion-MNIST (Xiao et al., 2017) provides a real image-classification task with controlled synthetic experts, while the CIFAR-10H replication separately checks training and annotation variation. Reported uncertainty is the sample standard deviation, with the replicate unit specified in each caption, and none of the comparisons below uses test metrics for checkpoint or hyperparameter selection.

P.1 Fashion-MNIST Protocol

We use the official Fashion-MNIST train/test split, holding out $5{,}000$ training examples for validation. The base panel has eight overlapping specialists, with in-domain accuracy 0.90 and out-of-domain accuracy 0.10; an incorrect prediction is uniform over the other nine labels. The domains, in expert order, are

$$\{0,2,6\},\ \{2,4,6\},\ \{0,3,4,6\},\ \{1,3,8\},\ \{5,7,9\},\ \{5,7,8\},\ \{7,9\},\ \{0,1,8\}.$$

The experts are generated with seed 12345, and within each training seed all methods use the same split and realized expert predictions. Larger pools repeat the realized eight-column panel exactly: $\mathcal{M}_{16}$ contains two copies and $\mathcal{M}_{32}$ four copies. Duplication therefore changes the action count without changing the best available action or adding information.

All Fashion-MNIST experiments use four training seeds (0–3), AdamW with learning rate 10^{-3} and weight decay 10^{-4}, batch size 1024, and 25 epochs. Validation is performed every 54 optimization steps; the checkpoint minimizes true validation deferral loss, or raw deployment cost for the price study. The shared CNN has a 128-dimensional feature representation and a linear output projection. Separate-trunk arms use two narrower CNNs to match the deployed parameter count. At $J = 8, 16, 32$, shared systems have $422{,}770$, $423{,}802$, and $425{,}866$ parameters; separate systems have $423{,}014$, $423{,}702$, and $425{,}078$.

Cells with the same nominal configuration in different studies use one canonical result: the shared/split decoupled arm at total $\lambda = 1$ is the same saved weight-study run in the architecture, schedule, and zero-price tables, so that one stochastic realization of a configuration is not presented as several methods.

P.2 Loss, Head Layout, and Representation Sharing

The factorial ablation compares the decoupled and Add. CE losses with (i) a shared trunk and split projections, (ii) a shared trunk and one $(K+J)$ projection, and (iii) separate parameter-matched trunks. A single projection can implement the decoupled loss by applying a softmax only to its class block and a sigmoid to its expert block, so separation is a property of the loss and normalization rather than of whether a layer is stored as one tensor or two. The split and single shared projections are initialized by copying the corresponding weights exactly.

Table 14: Fashion-MNIST loss and architecture ablation: test system accuracy (mean $\pm$ sample SD, four seeds). Parameter counts differ by less than 0.2%.

Loss and architecture	$J = 8$	$J = 16$	$J = 32$
Dec., shared/split	$0.9593_{\pm 0.0025}$	$0.9585_{\pm 0.0011}$	$0.9585_{\pm 0.0005}$
Dec., shared/single	$0.9589_{\pm 0.0010}$	$0.9577_{\pm 0.0013}$	$0.9593_{\pm 0.0013}$
Dec., separate	$0.9560_{\pm 0.0006}$	$0.9552_{\pm 0.0008}$	$0.9544_{\pm 0.0022}$
Add. CE, shared/split	$0.9333_{\pm 0.0053}$	$0.9275_{\pm 0.0041}$	$0.9289_{\pm 0.0063}$
Add. CE, shared/single	$0.9351_{\pm 0.0040}$	$0.9276_{\pm 0.0032}$	$0.9249_{\pm 0.0063}$
Add. CE, separate	$0.9342_{\pm 0.0028}$	$0.9302_{\pm 0.0029}$	$0.9306_{\pm 0.0031}$

Table 15: Standalone classifier accuracy for the same ablation (mean $\pm$ sample SD, four seeds).

Loss and architecture	$J = 8$	$J = 16$	$J = 32$
Dec., shared/split	$0.9200_{\pm 0.0016}$	$0.9191_{\pm 0.0019}$	$0.9164_{\pm 0.0031}$
Dec., shared/single	$0.9189_{\pm 0.0007}$	$0.9152_{\pm 0.0025}$	$0.9182_{\pm 0.0028}$
Dec., separate	$0.9127_{\pm 0.0043}$	$0.9141_{\pm 0.0022}$	$0.9111_{\pm 0.0019}$
Add. CE, shared/split	$0.9074_{\pm 0.0045}$	$0.8995_{\pm 0.0045}$	$0.9049_{\pm 0.0034}$
Add. CE, shared/single	$0.9098_{\pm 0.0041}$	$0.9038_{\pm 0.0020}$	$0.9022_{\pm 0.0027}$
Add. CE, separate	$0.9060_{\pm 0.0039}$	$0.9042_{\pm 0.0063}$	$0.9052_{\pm 0.0046}$

The decoupled loss has higher system accuracy in all 36 paired seed–pool–architecture contrasts. Its weakest four-seed mean is 0.9544, above Add. CE's strongest mean of 0.9351. With separate trunks, the mean loss advantage remains 2.18–2.50 percentage points, whereas head-layout differences are much smaller and have no consistent sign. Sharing a representation adds a further 0.33–0.41 percentage points to the decoupled split-head system. These controlled comparisons support the loss explanation while also showing the benefit of joint representation learning. Because the paired contrasts share data, we do not treat them as independent samples for a significance test.

Table 16: Shared-trunk gradients at aligned initialization (mean ± sample SD, four seeds). Ratio is expert/class gradient norm; cosine is the normalized inner product.

J	Dec. ratio	Dec. cosine	Add. CE ratio	Add. CE cosine
8	$0.406_{\pm 0.056}$	$-0.001_{\pm 0.027}$	$2.849_{\pm 0.163}$	$-0.721_{\pm 0.027}$
16	$0.305_{\pm 0.061}$	$-0.004_{\pm 0.053}$	$3.123_{\pm 0.080}$	$-0.746_{\pm 0.042}$
32	$0.208_{\pm 0.021}$	$-0.016_{\pm 0.062}$	$3.437_{\pm 0.212}$	$-0.732_{\pm 0.028}$

For the shared-trunk diagnostic, we differentiate the class-target and expert-target loss components separately with respect to the CNN parameters at aligned initialization, using the same diagnostic batch within a seed. The expert/class norm ratio is 0.21–0.41 for the decoupled loss, with mean cosine near zero. For Add. CE the ratio is 2.85–3.44, with mean cosine between -0.75 and -0.72. These quantities measure trunk interaction at initialization only: Proposition 9 establishes zero direct mixed head curvature, and it does not assert orthogonality of shared-trunk gradients throughout training.

P.3 Expert-Weight Sensitivity

We sweep total $\lambda \in \{0.0625, 0.125, 0.25, 0.5, 1, 2, 4, 8\}$ under exact panel duplication, completing 96 seed–pool–weight cells. The initial class parameters and the first eight expert rows are identical across pool sizes within each seed.

Table 17: Fashion-MNIST under exact expert-panel duplication: test system accuracy for fixed total weight λ (mean ± sample SD, four seeds).

Total λ	$J = 8$	$J = 16$	$J = 32$
0.0625	$0.9573_{\pm 0.0014}$	$0.9570_{\pm 0.0015}$	$0.9566_{\pm 0.0006}$
0.125	$0.9579_{\pm 0.0019}$	$0.9574_{\pm 0.0018}$	$0.9571_{\pm 0.0022}$
0.25	$0.9581_{\pm 0.0024}$	$0.9579_{\pm 0.0007}$	$0.9567_{\pm 0.0027}$
0.5	$0.9587_{\pm 0.0017}$	$0.9589_{\pm 0.0011}$	$0.9584_{\pm 0.0013}$
1	$0.9593_{\pm 0.0025}$	$0.9585_{\pm 0.0011}$	$0.9585_{\pm 0.0005}$
2	$0.9591_{\pm 0.0009}$	$0.9594_{\pm 0.0017}$	$0.9607_{\pm 0.0019}$
4	$0.9587_{\pm 0.0013}$	$0.9588_{\pm 0.0017}$	$0.9600_{\pm 0.0009}$
8	$0.9582_{\pm 0.0008}$	$0.9588_{\pm 0.0025}$	$0.9577_{\pm 0.0007}$

Table 18: Three weight choices on the same validation/test split (mean ± sample SD, four seeds); validation selection uses true deferral loss.

Weight choice	$J = 8$	$J = 16$	$J = 32$
Fixed total $\lambda = 1$	$0.9593_{\pm 0.0025}$	$0.9585_{\pm 0.0011}$	$0.9585_{\pm 0.0005}$
Fixed $\lambda/J = 0.125$	$0.9593_{\pm 0.0025}$	$0.9594_{\pm 0.0017}$	$0.9600_{\pm 0.0009}$
Validation-selected	$0.9580_{\pm 0.0013}$	$0.9587_{\pm 0.0015}$	$0.9589_{\pm 0.0007}$

At any fixed total weight in this grid, the spread of mean accuracy across the three pool sizes is at most 0.0016. Fixed total $\lambda = 1$, fixed per-expert weight $\lambda/J = 0.125$, and validation selection all produce stable performance here. Validation selection chooses, for each seed and pool size, the total weight that minimizes validation deferral loss; in the one tied cell ($J = 16$, seed 1), the selected total weight is 4. Fixed total weight keeps the total expert loss budget fixed, while fixed per-expert weight keeps each expert's gradient and the transfer coefficient fixed. Both choices give valid surrogate objectives, and the theoretical coefficient alone does not determine the best empirical trade-off for a learned representation.

P.4 One-Stage and Two-Stage Training

We compare shared one-stage training, disjoint one-stage training, and our two-stage implementation. In the latter, stage one selects a class CNN by validation classification error and freezes it; stage two trains an independent Bernoulli-utility CNN, selecting by true validation deferral loss. The disjoint simultaneous and sequential arms use the same component architectures, and their objectives have the product minimizer structure proved in Appendix J.4.

The fourth arm uses the logistic predictor–rejector surrogate of Mao et al. (2024f), implemented with a learned $(J+1)$-action router and the cost weights derived in that appendix, and it follows

the same frozen-classifier protocol and parameter budget as our two-stage arm. This controlled comparison holds the optimizer fixed, so it measures behavior under that protocol rather than after a method-specific hyperparameter search.

Table 19: Fashion-MNIST training schedules: test system accuracy (mean ± sample SD, four seeds). The last row uses the logistic predictor–rejector surrogate of Mao et al. (2024f); all arms share the same optimization protocol.

Training procedure	$J = 8$	$J = 16$	$J = 32$
Dec., one-stage shared	$0.9593_{\pm 0.0025}$	$0.9585_{\pm 0.0011}$	$0.9585_{\pm 0.0005}$
Dec., one-stage disjoint	$0.9546_{\pm 0.0021}$	$0.9548_{\pm 0.0020}$	$0.9557_{\pm 0.0009}$
Dec., two-stage	$0.9551_{\pm 0.0013}$	$0.9546_{\pm 0.0012}$	$0.9550_{\pm 0.0009}$
Predictor–rejector	$0.9036_{\pm 0.0122}$	$0.9107_{\pm 0.0049}$	$0.9118_{\pm 0.0044}$

Table 20: Standalone classifier accuracy for the schedule comparison (mean ± sample SD, four seeds).

Training procedure	$J = 8$	$J = 16$	$J = 32$
Dec., one-stage shared	$0.9200_{\pm 0.0016}$	$0.9191_{\pm 0.0019}$	$0.9164_{\pm 0.0031}$
Dec., one-stage disjoint	$0.9105_{\pm 0.0090}$	$0.9126_{\pm 0.0032}$	$0.9137_{\pm 0.0043}$
Dec., two-stage	$0.9134_{\pm 0.0050}$	$0.9121_{\pm 0.0030}$	$0.9121_{\pm 0.0042}$
Predictor–rejector	$0.9134_{\pm 0.0050}$	$0.9121_{\pm 0.0030}$	$0.9121_{\pm 0.0042}$

Shared one-stage training exceeds disjoint one-stage training by about 0.5, 0.4, and 0.3 percentage points at $J = 8, 16, 32$, with 11 paired wins and one tie. Disjoint one-stage and our two-stage means differ by less than 0.001, with no consistent direction. The predictor–rejector arm reaches 0.9036–0.9118 under the fixed protocol; since its class predictor is the same as in our two-stage arm, this gap isolates the router comparison. Together, these results show that our formulation supports either schedule, while shared one-stage training can use expert supervision to improve the representation.

P.5 Heterogeneous Consultation Prices

For $J = 8$, expert j has consultation price $b_j = \beta/j$, with $\beta \in \{0, 0.2, 0.4, 0.6, 0.8\}$. We evaluate and select checkpoints using the raw cost

$$\widehat{R}_{\mathrm{raw}} = \frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{1}\{r(x_i)=0\}\mathbf{1}\{h(x_i)\neq y_i\} + \sum_{j=1}^{8}\mathbf{1}\{r(x_i)=j\}\left(\frac{\beta}{j} + \mathbf{1}\{m_j(x_i)\neq y_i\}\right)\right].$$

For bounded-target surrogates, the training utility is $1 - (\beta/j + \mathbf{1}\{m_j \neq y\})/(1+\beta)$, and our class utility is transformed to the same scale, as in Appendix J.5. The classifier-only objective does not depend on the prices, so its single canonical run per seed is reused at all prices. For positive prices, PiCCE-CE is our own cost-weighted extension of the zero-price PiCCE objective, not a published consultation-price formulation.

Table 21: Fashion-MNIST consultation prices: raw deployment risk (mean ± sample SD, four seeds). Expert j has price β/j.

Method	$\beta = 0$	0.2	0.4	0.6	0.8
Decoupled	$0.0407_{\pm 0.0025}$	$0.0598_{\pm 0.0014}$	$0.0710_{\pm 0.0017}$	$0.0774_{\pm 0.0019}$	$0.0818_{\pm 0.0014}$
OvA	$0.0465_{\pm 0.0008}$	$0.0689_{\pm 0.0033}$	$0.0800_{\pm 0.0027}$	$0.0915_{\pm 0.0027}$	$0.0992_{\pm 0.0024}$
A-SM	$0.0514_{\pm 0.0018}$	$0.0725_{\pm 0.0032}$	$0.0869_{\pm 0.0015}$	$0.0960_{\pm 0.0020}$	$0.0996_{\pm 0.0024}$
Add. CE	$0.0701_{\pm 0.0047}$	$0.0937_{\pm 0.0042}$	$0.1182_{\pm 0.0069}$	$0.1317_{\pm 0.0060}$	$0.1399_{\pm 0.0102}$
PiCCE-CE extension	$0.0632_{\pm 0.0026}$	$0.1009_{\pm 0.0102}$	$0.1118_{\pm 0.0092}$	$0.1221_{\pm 0.0034}$	$0.1286_{\pm 0.0043}$
Classifier only	$0.0844_{\pm 0.0014}$	$0.0844_{\pm 0.0014}$	$0.0844_{\pm 0.0014}$	$0.0844_{\pm 0.0014}$	$0.0844_{\pm 0.0014}$
Mao25, linear member	$0.2436_{\pm 0.1150}$	$0.2963_{\pm 0.0755}$	$0.3068_{\pm 0.0391}$	$0.3958_{\pm 0.0853}$	$0.4352_{\pm 0.0910}$

The decoupled surrogate has the lowest mean raw deployment risk at every price. Its mean coverage rises from 0.7739 to 0.9357, increasing monotonically in each seed, while its classifier accuracy remains between 0.9154 and 0.9200. At the highest price, its raw risk is 0.0818, below the classifier-only risk of 0.0844, so the router still identifies useful deferrals. We report raw risk rather than accuracy because, with positive consultation prices, raw cost is no longer one minus system accuracy.

Table 22: Fashion-MNIST consultation prices: classifier coverage (mean ± sample SD, four seeds). Expert j has price β/j.

Method	$\beta = 0$	0.2	0.4	0.6	0.8
Decoupled	$0.7739_{\pm 0.0139}$	$0.8215_{\pm 0.0033}$	$0.8692_{\pm 0.0007}$	$0.9054_{\pm 0.0050}$	$0.9357_{\pm 0.0026}$
OvA	$0.6941_{\pm 0.0040}$	$0.7415_{\pm 0.0213}$	$0.8226_{\pm 0.0232}$	$0.8719_{\pm 0.0077}$	$0.9204_{\pm 0.0054}$
A-SM	$0.6169_{\pm 0.0098}$	$0.7256_{\pm 0.0067}$	$0.7890_{\pm 0.0064}$	$0.8599_{\pm 0.0131}$	$0.9127_{\pm 0.0033}$
Add. CE	$0.4740_{\pm 0.0306}$	$0.5586_{\pm 0.0193}$	$0.6067_{\pm 0.0245}$	$0.6734_{\pm 0.0217}$	$0.7358_{\pm 0.0304}$
PiCCE-CE extension	$0.6094_{\pm 0.0134}$	$0.8862_{\pm 0.0960}$	$0.9018_{\pm 0.0884}$	$0.9557_{\pm 0.0190}$	$0.9646_{\pm 0.0143}$
Classifier only	$1.0000_{\pm 0.0000}$	$1.0000_{\pm 0.0000}$	$1.0000_{\pm 0.0000}$	$1.0000_{\pm 0.0000}$	$1.0000_{\pm 0.0000}$
Mao25, linear member	$0.0000_{\pm 0.0000}$	$0.0000_{\pm 0.0000}$	$0.0000_{\pm 0.0000}$	$0.0000_{\pm 0.0000}$	$0.0236_{\pm 0.0472}$

P.6 CIFAR-10H: Training Seeds and Annotation Draws

We repeat the frozen-feature protocol of Appendix N with three training seeds $(0, 1, 2)$ for each of five annotation draws (0–4) and $J \in \{5, 10, 20\}$, giving 45 runs per method, while the data split, backbone, and method-specific hyperparameters remain fixed. The main-text Table 3 reports draw 0, with uncertainty across training seeds, whereas Table 23 first averages the training seeds within each draw and reports variation across the five draw means.

Table 23: CIFAR-10H: system accuracy over five annotation draws (mean ± sample SD of three-seed means); final column is the classifier-accuracy range over all 45 runs. The draw is the replicate unit in this table.

Method	$J = 5$	$J = 10$	$J = 20$	Classifier range
Decoupled	$0.9606_{\pm 0.0036}$	$0.9596_{\pm 0.0039}$	$0.9599_{\pm 0.0037}$	0.8860–0.8935
OvA	$0.9591_{\pm 0.0033}$	$0.9535_{\pm 0.0046}$	$0.9566_{\pm 0.0034}$	0.4350–0.8920
A-SM	$0.9590_{\pm 0.0042}$	$0.9525_{\pm 0.0047}$	$0.9475_{\pm 0.0033}$	0.0045–0.8930
Add. CE	$0.9550_{\pm 0.0037}$	$0.9526_{\pm 0.0019}$	$0.9538_{\pm 0.0017}$	0.8850–0.8925
PiCCE	$0.9555_{\pm 0.0045}$	$0.9571_{\pm 0.0028}$	$0.9548_{\pm 0.0010}$	0.7265–0.8925
Classifier only	$0.8897_{\pm 0.0000}$	$0.8885_{\pm 0.0000}$	$0.8890_{\pm 0.0000}$	0.8875–0.8915
Mao25 (linear)	$0.9500_{\pm 0.0037}$	$0.9496_{\pm 0.0016}$	$0.9478_{\pm 0.0021}$	0.0025–0.3230

The classifier of the decoupled surrogate ranges from 0.8860 to 0.8935 across all 45 runs, while its system accuracy benefits from routing. Add. CE also preserves classifier quality in this replication, whereas OvA, PiCCE, A-SM, and the linear Mao25 instance exhibit wider classifier ranges. Our method defers less in 44 or 45 of the 45 matched comparisons against each other L2D method; we report these counts descriptively and do not treat the three pool sizes as independent replicates.

Because worker identities vary across images (Appendix N), this experiment tests routing over human responses, whereas Covertype uses a fixed pool of named model experts.

Q Limitations

The analysis assumes labeled training examples with observed expert outputs and compares policies under a fixed deployment distribution. Missing annotations, changes in expert behavior, and availability constraints would require an observation or deployment model beyond this setting. The population result is an unrestricted surrogate-regret transfer, and it does not assert that finite neural-network training reaches the population infimum. Representation sharing can improve estimation, as the ablations show, but the zero mixed head block does not remove interaction through the shared trunk. The controlled studies isolate specific mechanisms, and their results are tied to the reported expert panels and training protocols.

R Broader Impacts

This work contributes theory and algorithms for learning-to-defer. Deferral policies that estimate class posteriors and expert utilities more accurately could improve human–AI systems in settings such as medical triage, by routing difficult cases to qualified experts. As with any deferral system, the risks include over-reliance on imperfect experts, error rates that differ across subgroups, and unclear accountability when a model and a human share a decision.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction state the two-axis analysis (statistical target and optimization geometry) of five augmented-action surrogates, the decoupled surrogate, and the consistency bound whose constant is independent of J at fixed λ/J; each claim is established in Sections 4–6 and the appendix.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Appendix Q discusses the observation and deployment assumptions, the unrestricted nature of the guarantee, interaction through a shared representation, and the experimental scope; broader impacts are discussed in Appendix R.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: All propositions and the main consistency theorem are stated with their assumptions in the main text; complete proofs, together with worked examples, are given in Appendices B–I and Appendix J.

Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: Data splits, expert construction, and seed counts are reported in Appendices L–O; backbones, optimizers, learning rates, schedules, and the per-expert weight λ/J for the real-data benchmarks are reported in Appendices M–O and P. Source code for all experiments is included in the supplementary material.

Guidelines:

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

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: Source code with run scripts is provided in the supplementary material. All datasets used (CIFAR-10, CIFAR-10H, Forest CoverType, Fashion-MNIST) are public; loading and preprocessing instructions are included in the code repository.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Training protocols (optimizer, learning rate, epochs, head architecture, validation-based checkpointing, and expert-weight selection) are described for each benchmark, together with weight decay for CIFAR-10H, Covertype, and Fashion-MNIST and the Fashion-MNIST batch size, in Appendices M–O and P.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

Justification: Tables report means and sample standard deviations: four training seeds for CIFAR-10, Covertype, and Fashion-MNIST; three training seeds for the fixed-draw CIFAR-10H main table. The full CIFAR-10H replication separately reports variation across five three-seed draw means. The synthetic suites report standard deviations over three to five seeds, and single-training-seed CIFAR-10H results over five draws (Table 12) are labeled separately. Repeated pool sizes are not treated as independent replicates for inference.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Appendix K reports the hardware (NVIDIA A100 GPUs with 40GB of memory, and H100/H200 GPUs for the Fashion-MNIST studies), one GPU per run, and the execution time: the synthetic suites finish within minutes per seed, and each real-data benchmark takes about one hour on a single H100 GPU.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The work uses only public benchmark datasets, collects no new human-subject data, and the authors have reviewed and adhere to the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: Appendix R discusses broader impacts, covering both the potential benefits of more accurate human–AI deferral and the risks of over-reliance on imperfect experts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper releases only research code for training small classification and routing models on public benchmarks; it does not release datasets, generative models, or pretrained language models posing high misuse risk.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: CIFAR-10 (Krizhevsky, 2009), CIFAR-10H (Peterson et al., 2019), Forest CoverType (Blackard and Dean, 1999), and Fashion-MNIST (Xiao et al., 2017) are credited where used. Fashion-MNIST is released under the MIT license, CIFAR-10H under CC BY-NC-SA 4.0, and Covertype under CC BY 4.0 (UCI repository); CIFAR-10 is used under its distribution terms. All are used in accordance with their licenses and terms of use.

    Guidelines:

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

13. **New assets**

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

    Answer: [Yes]

    Justification: The supplementary material contains the research code implementing the decoupled surrogate and all baselines, with a README documenting usage, experiment scripts, and expected outputs.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: No new crowdsourcing or human-subjects data collection was conducted; CIFAR-10H labels are used as released by Peterson et al. (2019).

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

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

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

    Answer: [N/A]

    Justification: The paper does not involve research with human subjects; no IRB review is required.

    Guidelines:

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

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: The proposed method does not use an LLM. LLM assistance was used during manuscript revision for exposition, derivation checks, source auditing, and table-generation code; the authors are responsible for validating the final claims and artifacts.

    Guidelines:

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