On the Learnability of Model Reprogramming: A Reachability and Relabeling Perspective

Zesheng Ye[1] **Pin-Yu Chen**[2] **Feng Liu**[1]
[1]The University of Melbourne [2]IBM Research
{zeshengye.ml, fengliu.ml}@gmail.com, pin-yu.chen@ibm.com

Abstract

Model reprogramming adapts a *frozen* source model to a target task by wrapping it in a *learnable* input transformation from target inputs to the source-model input space and an output mapping from source-model outputs to target predictions. Despite extensive empirical progress, the *finite-sample learnability* of the induced wrapper hypothesis class $\mathcal{F}_{\mathrm{MR}}$ remains underexplored. We address this question for source models that expose hard source labels and analyze it through the wrapper structure: on any target sample, the input transformation selects *reachable* source labels that the output mapping relabels. Through this lens, we answer two questions: ① how many target samples are needed to approach the best predictor in $\mathcal{F}_{\mathrm{MR}}$, and ② what gap remains between the optimum over $\mathcal{F}_{\mathrm{MR}}$ and the target Bayes risk. For ①, the wrapper structure factorizes the sample-wise complexity of $\mathcal{F}_{\mathrm{MR}}$ into a *reachability* term counting reachable source-label patterns and a *conditional relabeling* term counting their relabelings. This yields an agnostic-PAC bound on the *estimation error*, with sample complexity scaling additively in the two terms; this bound becomes closed-form on additive input transformations and is matched by a minimax lower bound on an explicit binary classification construction. For ②, the same structure decomposes the *approximation error* exactly into source-label information loss and output-mapping restriction, together with a source-conditional upper bound on this gap. We extend the approximation analysis to logit-exposing source models, where the upper bound is empirically estimable, and verify both the estimation and approximation mechanisms in controlled synthetic and pretrained-MNIST experiments. The wrapper-side analysis localizes both errors to specific components, giving a structural view of reprogramming practice. Code: .

1 Introduction

As pre-trained machine learning models grow in scale, much of their value lies in capabilities that are difficult or expensive to reproduce by retraining [1–6]. Meanwhile, a rising fraction of them is accessible only through restricted interfaces, e.g., released checkpoints or even deployment-time inference APIs, making adaptation methods that do not modify source weights, and can operate under restricted access, increasingly important [7–9]. *Model reprogramming* (MR) [10–12] stands out as an active model-reuse paradigm for this problem. It adapts a *frozen* pre-trained model (i.e., a source model) to a downstream task (i.e., a target task) by wrapping it in a composite predictor[1] with a *learnable input transformation* that maps target inputs into the source-model input space, and a *learnable output mapping* that converts source-model outputs (e.g., source labels) into target predictions (Figure 1). This wrapper architecture is considered valuable in frozen-source-model reuse scenarios where target-task data are limited [12] and has succeeded across image [10, 13–15], text [11], time-series [16, 17], protein sequences [18, 19], and cross-modal [20] adaptation tasks.

[1]This paper focuses on the settings where the reprogrammed model is a *classifier*.

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

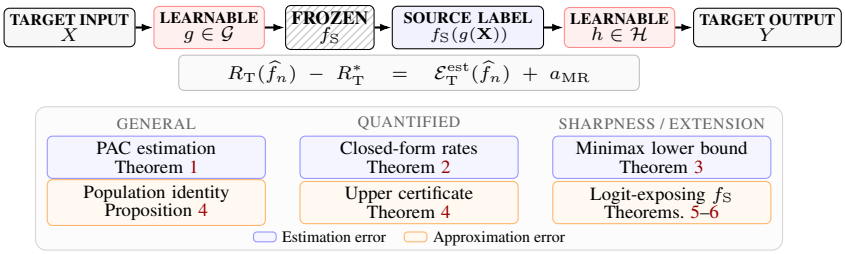


Figure 1: We present a *finite-sample learnability* analysis for hard-label MR through the MR wrapper structure, organized by the *estimation error* (Sections 3 to 5) and *approximation error* (Section 6) perspectives, respectively.

Gap. The empirical progress of MR, however, runs ahead of the theory: particularly, *finite-sample* understanding of *what controls the behavior of such a wrapper*, the regime that limited-target-data use cases demand, remains underexplored. Existing reprogramming theory has not yet addressed this directly, but relies on structural assumptions on the source model, including source–target alignment [17], specific source-model architectures [21], and spectral regularity of the linearized source model [22]. Such routes characterize which source models MR may succeed on, but they do not isolate which properties of the *reprogramming wrapper* itself control its finite-sample behavior.

According to [23, 24], understanding that behavior on a target distribution amounts to controlling the wrapper's target excess risk [23], split into two complementary parts, namely the *estimation error*, which measures the gap from the learned predictor to the best predictor in the MR-induced wrapper hypothesis class $\mathcal{F}_{\mathrm{MR}}$, and the *approximation error*, which measures the gap from that in-class optimum to the target Bayes risk [25]. Each part gives rise to a corresponding question naturally:

① How many target samples does the learned predictor need to approach the best predictor in $\mathcal{F}_{\mathrm{MR}}$?
② What residual gap remains between the optimum over $\mathcal{F}_{\mathrm{MR}}$ and the target Bayes risk, even with unlimited target samples?

We address this for source models that expose hard source labels, i.e., *hard-label* MR [10, 13, 16], grounded in the wrapper structure (defined in Section 2.1): when the source model is fixed, only the input transformation and the output mapping are learnable, and they act sequentially. Thus, on each target sample, the input transformation determines which source-model output (i.e., a source label) reaches the output mapping, which then relabels those reachable patterns into target labels. This MR-specific structure governs both questions, and our answers below trace its consequences.

Answer to ①. We discover that sample-wise complexity of $\mathcal{F}_{\mathrm{MR}}$ factorizes into a *reachability* term and a *conditional relabeling* term, each recording the realizable behavior of one wrapper component on the target sample. This factorization yields an agnostic *probably approximately correct* (PAC) bound [23] on the *estimation error* of approximate *empirical risk minimization* (ERM) over $\mathcal{F}_{\mathrm{MR}}$ (Theorem 1), with sample complexity scaling additively in the growth of these two terms. On additive input transformations [10, 13, 16, 26, 27], both terms admit closed-form rates under structural conditions on the source model (Theorem 2), verified for affine and finite-depth ReLU source-score maps (Corollaries 3-4). We further prove the necessity of this sample-complexity dependence via a matching minimax lower bound [28] on an explicit binary classification construction, whose Bayes-optimal predictor lies in $\mathcal{F}_{\mathrm{MR}}$, so that the lower bound applies to the estimation error (Theorem 3).

Answer to ②. We establish that the *approximation error* of $\mathcal{F}_{\mathrm{MR}}$ admits an exact population identity, decomposing it into the information loss when source labels do not preserve enough target structure for any relabeling to recover the target Bayes rule, and the additional loss when the output-mapping class does not contain that Bayes relabeling (Proposition 4). Substituting Bayes selectors of the source and target distributions into the Proposition 4 gives three source-conditional quantities, whose sum upper-bounds the residual gap (Theorem 4). We extend the same decomposition to logit-exposing source models, with similar population identities (Propositions 5 and 6) and empirically estimable upper bounds on the residual (Theorems 5 and 6 and Corollary 6).

Contribution. We give the first *wrapper-side* finite-sample learnability analysis for hard-label MR. By analyzing the reprogramming wrapper itself, our framework isolates reachability and conditional relabeling, the two structural components that control finite-sample behavior of MR, and we verify both mechanisms in controlled synthetic and pretrained-MNIST experiments (Section 7). Beyond the

bounds themselves, our results yield a theoretical perspective to understand the design choices and failure cases of existing MR practices (Section 8). The wrapper-side viewpoint opens finite-sample MR analysis to source models accessed only through their input–output interface, complementing existing MR theory that focuses on source-model properties [17, 21, 22].

2 Hard-Label Model Reprogramming

We formalize the hard-label MR setup and isolate the two combinatorial objects on which our analysis rests. Section 2.1 fixes the MR-induced hypothesis class and the approximate-ERM learner; Section 2.2 sets up the estimation-approximation decomposition; Section 2.3 introduces the sample-wise growth functions whose factorization drives Section 3. We summarize the notations in Appendix A.

2.1 Formulation of Model Reprogramming

Let $\mathcal{X}_{\mathrm{T}}$ and $\mathcal{X}_{\mathrm{S}}$ denote measurable target and source input spaces with finite label alphabets $[K_{\mathrm{T}}]$ and $[K_{\mathrm{S}}]$, respectively, where $[K] := \{1, \dots, K\}$. Throughout the main text, $f_{\mathrm{S}} : \mathcal{X}_{\mathrm{S}} \to [K_{\mathrm{S}}]$ is a fixed measurable hard-label source classifier.

Model reprogramming (MR) [12] repurposes f_{S} for a target task governed by a distribution P_{T} over $\mathcal{X}_{\mathrm{T}} \times [K_{\mathrm{T}}]$ by composing f_{S} with two *learnable* wrapper stages: an *input transformation* $g \in \mathcal{G}$ (a nonempty class of measurable maps $\mathcal{X}_{\mathrm{T}} \to \mathcal{X}_{\mathrm{S}}$) maps target inputs into $\mathcal{X}_{\mathrm{S}}$, and an *output mapping* $h \in \mathcal{H} \subseteq [K_{\mathrm{T}}]^{[K_{\mathrm{S}}]}$ (nonempty) converts source labels into target predictions. Each $(g, h) \in \mathcal{G} \times \mathcal{H}$ defines a reprogrammed predictor $f_{g,h} := h \circ f_{\mathrm{S}} \circ g$ (abbreviated f). The MR-induced hypothesis class collects all such compositions:

$$\mathcal{F}_{\mathrm{MR}} := \{f_{g,h} = h \circ f_{\mathrm{S}} \circ g : g \in \mathcal{G}, h \in \mathcal{H}\} \subseteq [K_{\mathrm{T}}]^{\mathcal{X}_{\mathrm{T}}}.$$

Given an i.i.d. target sample $\{(X_i, Y_i)\}_{i=1}^{n_{\mathrm{T}}} \sim P_{\mathrm{T}}^{n_{\mathrm{T}}}$, write $\widehat{R}_n(f) := n_{\mathrm{T}}^{-1} \sum_{i=1}^{n_{\mathrm{T}}} \mathbf{1}\{f(X_i) \neq Y_i\}$ for the empirical 0–1 risk. A measurable predictor $\widehat{f}_n \in \mathcal{F}_{\mathrm{MR}}$ is an *η_n-approximate ERM* if $\widehat{R}_n(\widehat{f}_n) \leq \inf_{f \in \mathcal{F}_{\mathrm{MR}}} \widehat{R}_n(f) + \eta_n$ ($\eta_n \geq 0$). Unless stated otherwise, all results below quantify over every target distribution P_{T} over $\mathcal{X}_{\mathrm{T}} \times [K_{\mathrm{T}}]$, every fixed measurable $f_{\mathrm{S}} : \mathcal{X}_{\mathrm{S}} \to [K_{\mathrm{S}}]$, and every nonempty measurable wrapper pair $(\mathcal{G}, \mathcal{H})$.

2.2 Estimation error and approximation residual

For a target predictor $f : \mathcal{X}_{\mathrm{T}} \to [K_{\mathrm{T}}]$, write $R_{\mathrm{T}}(f) := \mathbb{P}_{(X,Y)\sim P_{\mathrm{T}}}(f(X) \neq Y)$ for its population 0–1 risk. Comparing a learned wrapper $\widehat{f}_n \in \mathcal{F}_{\mathrm{MR}}$ against the in-class optimum $R_{\mathrm{T}}^{\mathcal{F}_{\mathrm{MR}},*} := \inf_{f \in \mathcal{F}_{\mathrm{MR}}} R_{\mathrm{T}}(f)$ isolates finite-sample estimation cost, whereas comparing $R_{\mathrm{T}}^{\mathcal{F}_{\mathrm{MR}},*}$ against the unrestricted Bayes risk $R_{\mathrm{T}}^{*} := \inf_f R_{\mathrm{T}}(f)$ over all measurable $f : \mathcal{X}_{\mathrm{T}} \to [K_{\mathrm{T}}]$ isolates the structural expressivity limit of $\mathcal{F}_{\mathrm{MR}}$.

Definition 1 (Estimation–approximation error decomposition [29])**.** For every measurable predictor $\widehat{f}_n : \mathcal{X}_{\mathrm{T}} \to [K_{\mathrm{T}}]$, the target *estimation error* relative to $\mathcal{F}_{\mathrm{MR}}$ is $\mathcal{E}_{\mathrm{T}}^{\mathrm{est}}(\widehat{f}_n) := R_{\mathrm{T}}(\widehat{f}_n) - R_{\mathrm{T}}^{\mathcal{F}_{\mathrm{MR}},*}$ and the *approximation error* of $\mathcal{F}_{\mathrm{MR}}$ is $a_{\mathrm{MR}} := R_{\mathrm{T}}^{\mathcal{F}_{\mathrm{MR}},*} - R_{\mathrm{T}}^{*}$, so that $R_{\mathrm{T}}(\widehat{f}_n) - R_{\mathrm{T}}^{*} = \mathcal{E}_{\mathrm{T}}^{\mathrm{est}}(\widehat{f}_n) + a_{\mathrm{MR}}$.

The *estimation error* measures the finite-sample learning gap between the predictor obtained from n target examples and the best predictor in $\mathcal{F}_{\mathrm{MR}}$. The *approximation error* measures the residual gap between the best attainable risk obtained by any predictor within $\mathcal{F}_{\mathrm{MR}}$ and the target Bayes risk and cannot be reduced even with unlimited target data. In this paper, we handle both terms through the *two-stage* pipeline of the MR wrapper itself: $x \xrightarrow{g} g(x) \xrightarrow{f_{\mathrm{S}}} f_{\mathrm{S}}(g(x)) \xrightarrow{h} h(f_{\mathrm{S}}(g(x)))$. Namely, on any target sample, g first determines which source labels are exposed by f_{S}, and h then relabels them. We use this structure to both *decompose* the finite-sample complexity $\mathcal{F}_{\mathrm{MR}}$ that controls $\mathcal{E}_{\mathrm{T}}^{\mathrm{est}}$ and the population residual a_{MR}, in each case into matching input-side and output-side contributions. We next introduce the combinatorial objects entering the finite-sample complexity analysis.

2.3 Reachable source-label patterns and conditional relabelings

On a fixed target tuple $x_{1:m} := (x_1, \dots, x_m) \in \mathcal{X}_{\mathrm{T}}^m$, the two-stage pipeline induces two coupled pattern sets: the source-label vectors reachable by varying $g \in \mathcal{G}$ through f_{S}, and, for each such reachable vector z, the target-label vectors realizable by varying $h \in \mathcal{H}$.

Definition 2 (Reachable and realizable label patterns)**.** Let $\mathcal{R} := \{\, r := f_{\mathrm{S}} \circ g : g \in \mathcal{G} \,\}$ denote the reachable source-label predictor class. For every integer $m \geq 1$, every target tuple $x_{1:m} = (x_1, \ldots, x_m) \in \mathcal{X}_{\mathrm{T}}^m$, and every source-label vector $z = (z_1, \ldots, z_m) \in [K_{\mathrm{S}}]^m$, the reachable source-label pattern set on $x_{1:m}$ and the realizable target-label pattern set on z are

$$\mathrm{Rea}_{\mathcal{G}}(x_{1:m}) := \{(r(x_1), \ldots, r(x_m)) : r \in \mathcal{R}\}, \quad \mathrm{Rel}_{\mathcal{H}}(z) := \{(h(z_1), \ldots, h(z_m)) : h \in \mathcal{H}\}.$$

For a multiclass hypothesis class $\mathcal{F} \subseteq [K]^{\mathcal{X}}$, write $\Pi_{\mathcal{F}}(m) := \sup_{x_{1:m} \in \mathcal{X}^m} |\{(f(x_1), \ldots, f(x_m)) : f \in \mathcal{F}\}|$ for its sample-wise growth function [29]. Applying worst-case sample-wise counting to $\mathrm{Rea}_{\mathcal{G}}(x_{1:m})$ and to $\mathrm{Rel}_{\mathcal{H}}(z)$ over reachable $z \in \mathrm{Rea}_{\mathcal{G}}(x_{1:m})$ yields the two complexity measures.

Definition 3 (Reachability and conditional-relabeling growth)**.** For every $m \geq 1$, the *sample-wise reachability growth* and *conditional-relabeling growth* of $\mathcal{F}_{\mathrm{MR}}$ are $\Pi_{\mathcal{R}}(m) := \sup_{x_{1:m} \in \mathcal{X}_{\mathrm{T}}^m} |\mathrm{Rea}_{\mathcal{G}}(x_{1:m})|$ and $\Pi_{\mathcal{H}|\mathcal{R}}(m) := \sup_{x_{1:m} \in \mathcal{X}_{\mathrm{T}}^m} \sup_{z \in \mathrm{Rea}_{\mathcal{G}}(x_{1:m})} |\mathrm{Rel}_{\mathcal{H}}(z)|$.

Remark 1. Because the inner supremum in $\Pi_{\mathcal{H}|\mathcal{R}}(m)$ ranges only over reachable patterns $z \in \mathrm{Rea}_{\mathcal{G}}(x_{1:m})$, relaxing it to all $z \in [K_{\mathrm{S}}]^m$ defines the *structure-oblivious* proxy $\Pi_{\mathcal{H}}^{\mathrm{all}}(m) := \sup_{z \in [K_{\mathrm{S}}]^m} |\mathrm{Rel}_{\mathcal{H}}(z)|$, which can exponentially over-count head complexity. We revisit the distinction in Section 3.3 and show that the gap between $\Pi_{\mathcal{H}|\mathcal{R}}$ and $\Pi_{\mathcal{H}}^{\mathrm{all}}$ can be large in regimes practically relevant to MR.

3 Factorized PAC Bound on the Estimation Error

We now answer question ① and bound the sample complexity of learning a hard-label MR. The MR wrapper's two-stage pipeline factorizes the sample-wise growth of $\mathcal{F}_{\mathrm{MR}}$ into two independent terms, one per wrapper component (Section 3.1), and this factorization induces an additive PAC bound on the estimation error of approximate ERM (Section 3.2). Proofs are deferred to Appendix C.

3.1 Reachability before relabeling

On any fixed $x_{1:m} \in \mathcal{X}_{\mathrm{T}}^m$, each $f_{g,h} = h \circ f_{\mathrm{S}} \circ g \in \mathcal{F}_{\mathrm{MR}}$ first selects a reachable source-label vector $z \in \mathrm{Rea}_{\mathcal{G}}(x_{1:m})$ via g and then relabels z via h. The target-label patterns realized by $\mathcal{F}_{\mathrm{MR}}$ on $x_{1:m}$ therefore lie in $\bigcup_{z \in \mathrm{Rea}_{\mathcal{G}}(x_{1:m})} \mathrm{Rel}_{\mathcal{H}}(z)$. Bounding this union by $|\mathrm{Rea}_{\mathcal{G}}(x_{1:m})| \max_{z \in \mathrm{Rea}_{\mathcal{G}}(x_{1:m})} |\mathrm{Rel}_{\mathcal{H}}(z)|$ and taking the supremum over $x_{1:m}$ (Appendix C) gives:

Proposition 1 (Sample-wise factorization)**.** *For every* $m \geq 1$, $\Pi_{\mathcal{F}_{\mathrm{MR}}}(m) \leq \Pi_{\mathcal{R}}(m) \cdot \Pi_{\mathcal{H}|\mathcal{R}}(m)$.

Proposition 1 is the structural reduction behind the analysis. Rather than simply treating $\mathcal{F}_{\mathrm{MR}}$ as an opaque composition, it charges g only for the source-label patterns it can expose on $x_{1:m}$, and h only for relabelings of those exposed patterns. This enables the sample-complexity rate established below.

3.2 Agnostic-PAC estimation bound

Proposition 1 bounds combinatorial realizations on a fixed tuple, leaving open how that product governs the finite-sample estimation error $\mathcal{E}_{\mathrm{T}}^{\mathrm{est}}(\widehat{f}_n)$ of an η_n-approximate ERM without assuming target realizability. As the 0–1 loss indicator $\mathbf{1}\{f(x) \neq y\}$ depends on $f \in \mathcal{F}_{\mathrm{MR}}$ solely through $f(x)$, the induced loss class inherits growth at most $\Pi_{\mathcal{F}_{\mathrm{MR}}}(2n_{\mathrm{T}}) \leq \Pi_{\mathcal{R}}(2n_{\mathrm{T}})\Pi_{\mathcal{H}|\mathcal{R}}(2n_{\mathrm{T}})$. Applying the two-sided growth-function uniform-deviation bound [24, Corollary 3.9 and Eq. (3.23)] and telescoping empirical risk against $R_{\mathrm{T}}^{\mathcal{F}_{\mathrm{MR}},*}$ (Theorem 7) yields an *agnostic-PAC* bound (Definition 5) on the estimation error as follows.

Theorem 1 (Factorized agnostic-PAC bound)**.** *For every* $n_{\mathrm{T}} \geq 1$, $\delta \in (0,1)$, *and* $\eta_n \geq 0$, *with probability at least* $1-\delta$ *over an i.i.d. sample of size* n_{T} *from* P_{T}, *every measurable* η_n*-approximate ERM* $\widehat{f}_n \in \mathcal{F}_{\mathrm{MR}}$ *satisfies*

$$\mathcal{E}_{\mathrm{T}}^{\mathrm{est}}(\widehat{f}_n) \;\leq\; \eta_n + 4\sqrt{6}\sqrt{\frac{\log \Pi_{\mathcal{R}}(2n_{\mathrm{T}}) + \log \Pi_{\mathcal{H}|\mathcal{R}}(2n_{\mathrm{T}}) + \log(1/\delta)}{n_{\mathrm{T}}}}. \tag{1}$$

Theorem 1 bounds $\mathcal{E}_{\mathrm{T}}^{\mathrm{est}}(\widehat{f}_n) = R_{\mathrm{T}}(\widehat{f}_n) - R_{\mathrm{T}}^{\mathcal{F}_{\mathrm{MR}},*}$ for any measurable f_{S}: $\log \Pi_{\mathcal{R}}(2n_{\mathrm{T}})$ measures input-side reachability capacity and $\log \Pi_{\mathcal{H}|\mathcal{R}}(2n_{\mathrm{T}})$ measures output-side relabeling capacity on

reached patterns, vanishing whenever $\Pi_{\mathcal{R}}(2n_{\mathrm{T}}) = 1$ or $\Pi_{\mathcal{H}|\mathcal{R}}(2n_{\mathrm{T}}) = 1$, respectively. Inverting Eqn. (1) under polynomial growth via Lemma 1 (Corollary 7) gives an additive sample complexity.

Corollary 1 (Polynomial-growth sample complexity)**.** *Suppose there exist constants $d_{\mathcal{R}}, d_{\mathcal{H}} \geq 0$ and $C_{\mathcal{R}}, C_{\mathcal{H}} \geq 1$ such that*

$$\log \Pi_{\mathcal{R}}(m) \leq d_{\mathcal{R}} \log(C_{\mathcal{R}} m) \quad \textit{and} \quad \log \Pi_{\mathcal{H}|\mathcal{R}}(m) \leq d_{\mathcal{H}} \log(C_{\mathcal{H}} m) \qquad \textit{for every } m \geq 1.$$

Then for every $\varepsilon, \delta \in (0,1)$ and $\eta_n \geq 0$, there exists $n_0 = \widetilde{O}\big((d_{\mathcal{R}} + d_{\mathcal{H}} + \log(1/\delta))/\varepsilon^2\big)$ such that whenever $n_{\mathrm{T}} \geq n_0$, every measurable η_n-approximate ERM $\widehat{f}_n \in \mathcal{F}_{\mathrm{MR}}$ satisfies $\mathcal{E}^{\mathrm{est}}_{\mathrm{T}}(\widehat{f}_n) \leq \eta_n + \varepsilon$ with probability at least $1 - \delta$ over $P_{\mathrm{T}}^{n_{\mathrm{T}}}$.

Thus, if reachable source-label patterns and conditional relabelings grow polynomially, then $\mathcal{F}_{\mathrm{MR}}$ is *agnostically PAC learnable* by approximate ERM, with sample complexity controlled additively in the two *independent* growth exponents: $d_{\mathcal{R}}$ measures how distinct source-label patterns the input transformation can expose through f_{S} on a target sample, while $d_{\mathcal{H}}$ measures how richly the output mapping relabels a fixed reachable pattern. Section 4 verifies the polynomial-growth condition for additive input transformations [10, 13, 16, 26, 27] under structural conditions on f_{S}, and Section 5 establishes that the additive $d_{\mathcal{R}} + d_{\mathcal{H}}$ dependence is necessary on at least one explicit family.

3.3 Why conditioning on reachability matters

We next compare the reachability-aware term $\Pi_{\mathcal{H}|\mathcal{R}}$ with the structure-oblivious proxy $\Pi^{\mathrm{all}}_{\mathcal{H}}(m)$ from Section 2.3, which counts relabelings on arbitrary length-m source-label vectors, regardless of whether those vectors can be reached by $f_{\mathrm{S}} \circ \mathcal{G}$ on a target sample.

Proposition 2. *For every $m \geq 1$, $\Pi_{\mathcal{F}_{\mathrm{MR}}}(m) \leq \Pi_{\mathcal{R}}(m) \cdot \Pi_{\mathcal{H}|\mathcal{R}}(m) \leq \Pi_{\mathcal{R}}(m) \cdot \Pi^{\mathrm{all}}_{\mathcal{H}}(m)$.*

Thus, conditioning on reachability never worsens the growth bound and can be strictly sharper when source-label vectors are unreachable, with concrete consequences for Theorem 1.

Example 1 (Strict gap from unreachable source-label patterns)**.** Let $\mathcal{H} = [K_{\mathrm{T}}]^{[K_{\mathrm{S}}]}$ be the full relabeling class. For $z \in [K_{\mathrm{S}}]^m$, write $|z|_{\#} := |\{z_1, \ldots, z_m\}|$ for its alphabet size, and define the *reachable alphabet* $t_{\mathcal{R}}(m) := \sup_{x_{1:m}} \sup_{z \in \mathrm{Rea}_{\mathcal{G}}(x_{1:m})} |z|_{\#} \leq \min\{m, K_{\mathrm{S}}\}$ in general. Then,

$$\Pi^{\mathrm{all}}_{\mathcal{H}}(m) = K_{\mathrm{T}}^{\min\{m, K_{\mathrm{S}}\}}, \qquad \Pi_{\mathcal{H}|\mathcal{R}}(m) = K_{\mathrm{T}}^{t_{\mathcal{R}}(m)},$$

The gap $\Pi^{\mathrm{all}}_{\mathcal{H}} / \Pi_{\mathcal{H}|\mathcal{R}}$ can be exponential: when $K_{\mathrm{S}} \geq 2n_{\mathrm{T}}$ and $t_{\mathcal{R}}(2n_{\mathrm{T}}) = O(1)$, the conditional term contributes only $O(\log K_{\mathrm{T}})$ to Eqn. (1), against $2n_{\mathrm{T}} \log K_{\mathrm{T}}$ for the structure-oblivious proxy.

From factorized growth to explicit rates. Theorem 1 reduces hard-label PAC estimation on $\mathcal{F}_{\mathrm{MR}}$ to two distribution-free sample-wise growth functions: $\Pi_{\mathcal{R}}$, controlled by the input-transformation class together with f_{S}, and $\Pi_{\mathcal{H}|\mathcal{R}}$, controlled by the output-mapping class together with the source labels actually reached on the target sample. The bound is only as explicit as these two growth terms. We next turn it into concrete polynomial rates for additive input transformations.

4 Explicit PAC Rates for Additive Input Transformation

We now instantiate Theorem 1 on *additive* input transformations [10, 13, 16, 26, 27]: under structural conditions on f_{S} and $\mathcal{H}$ (Section 4.1), both growth functions admit explicit polynomial bounds (Section 4.2), verified for affine and fixed-depth ReLU source-score classifiers (Section 4.3).

4.1 Additive input transformations and structural conditions

We consider additive input transformations of the form $g_\theta(x) = x + M\theta$, with learnable parameter $\theta \in \Theta \subseteq \mathbb{R}^p$ and a fixed perturbation matrix $M \in \mathbb{R}^{d \times p}$ specifying the directions available to the wrapper. In visual reprogramming [10, 26], for example, M is a binary pixel mask that determines where on the target image the perturbation θ is placed.

Condition A (Reachability arrangement)**.** *There exist constants $c_{\mathcal{R}} \geq 0$ and $C_{\mathcal{R}} \geq 1$ such that, for every integer $m \geq 1$ and every target tuple $x_{1:m} \in \mathcal{X}_{\mathrm{T}}^m$, the parameter set Θ admits a partition into at most $(C_{\mathcal{R}} m)^{c_{\mathcal{R}} p}$* cells *on each of which the source-label vector is constant in θ.*

Condition B (Output-mapping complexity)**.** *The output-mapping class $\mathcal{H} \subseteq [K_{\mathrm{T}}]^{[K_{\mathrm{S}}]}$ has Natarajan dimension [30, 31] $\mathrm{Ndim}(\mathcal{H}) \leq d_{\mathcal{H}} < \infty$.*

Condition A stipulates that varying $\theta \in \Theta$ induces at most polynomially many source-label patterns on $x_{1:m}$ and concentrates all dependence on f_{S} (Section 4.3). Condition B controls multiclass head capacity independently of f_{S}.

4.2 Sample complexity and reachable-label arrangements

Under Condition A, the cell partition directly bounds $|\mathrm{Rea}_{\mathcal{G}}(x_{1:m})| \le (C_{\mathcal{R}} m)^{c_{\mathcal{R}} p}$; under Condition B, identifying $\mathrm{Rel}_{\mathcal{H}}(z)$ with restrictions $\mathcal{H}|_{T(z)}$ on $|T(z)| \le m$ and applying the multivalued Sauer–Shelah lemma [32, Theorem 9] (Lemma 2; Theorem 8) bounds $\Pi_{\mathcal{H}|\mathcal{R}}(m)$ polynomially in m.

Theorem 2 (Explicit PAC bound for additive MR). *Under Conditions A and B for $\mathcal{G}_p$, set $d_m := d_{\mathcal{H}} \wedge m$ (with $B_m := 0$ when $d_m = 0$ or $K_{\mathrm{T}} = 1$). For every $m \ge 1$,*

$$\log \Pi_{\mathcal{R}}(m) \le \underbrace{c_{\mathcal{R}}\, p \log(C_{\mathcal{R}}\, m)}_{=:\, A_m}, \qquad \log \Pi_{\mathcal{H}|\mathcal{R}}(m) \le \underbrace{d_m \log\!\big(e \tbinom{K_{\mathrm{T}}}{2} m/d_m\big)}_{=:\, B_m}.$$

Consequently, for every $n_{\mathrm{T}} \ge 1$, $\delta \in (0,1)$, and $\eta_n \ge 0$, with probability at least $1-\delta$ over $P_{\mathrm{T}}^{n_{\mathrm{T}}}$, every measurable η_n-approximate ERM $\widehat{f}_n \in \mathcal{F}_{\mathrm{MR}}$ satisfies

$$\mathcal{E}_{\mathrm{T}}^{\mathrm{est}}(\widehat{f}_n) \;\le\; \eta_n + 4\sqrt{6} \cdot \sqrt{n_{\mathrm{T}}^{-1}\,(A_{2n_{\mathrm{T}}} + B_{2n_{\mathrm{T}}} + \log(1/\delta))}. \tag{2}$$

Theorem 2 charges the input transformation by $c_{\mathcal{R}} p$ and the head by its f_{S}-free Natarajan dimension $d_{\mathcal{H}}$ (with $A_{2n_{\mathrm{T}}} = 0$ when $p = 0$ or $c_{\mathcal{R}} = 0$, and $B_{2n_{\mathrm{T}}} = 0$ when $d_{\mathcal{H}} = 0$ or $K_{\mathrm{T}} = 1$).

Corollary 2 (Polynomial sample complexity for additive MR). *Under Conditions A and B for $\mathcal{G}_p$, for every $\varepsilon, \delta \in (0,1)$ and $\eta_n \ge 0$, there exists*

$$n_0 = \widetilde{O}\!\left(\frac{c_{\mathcal{R}}\, p + d_{\mathcal{H}} + \log(1/\delta)}{\varepsilon^2}\right),$$

such that whenever $n_{\mathrm{T}} \ge n_0$, every measurable η_n-approximate ERM $\widehat{f}_n \in \mathcal{F}_{\mathrm{MR}}$ satisfies $\mathcal{E}_{\mathrm{T}}^{\mathrm{est}}(\widehat{f}_n) \le \eta_n + \varepsilon$ (and $R_{\mathrm{T}}(\widehat{f}_n) - R_{\mathrm{T}}^{*} \le a_{\mathrm{MR}} + \eta_n + \varepsilon$) with probability at least $1-\delta$ over $P_{\mathrm{T}}^{n_{\mathrm{T}}}$.

Additive hard-label MR is thus *agnostically PAC learnable* whenever Conditions A-B hold, with sample complexity additive in $c_{\mathcal{R}}\, p$ and $d_{\mathcal{H}}$, isolating architectural dependence on f_{S} inside $(c_{\mathcal{R}}, C_{\mathcal{R}})$.

4.3 Discharging Condition A through source-score structure

Condition A is the only f_{S}-side hypothesis in Theorem 2, whereas Theorem 1 holds for any measurable f_{S}. Thus, obtaining closed-form rates concentrates architecture dependence into bounding constant-label cells over Θ. We verify it for two families sharing a structure that f_{S} is induced by a *score map* $s_{\mathrm{S}} : \mathcal{X}_{\mathrm{S}} \to \mathbb{R}^{K_{\mathrm{S}}}$ via $f_{\mathrm{S}}(u) = \arg\max_{k \in [K_{\mathrm{S}}]} s_{\mathrm{S},k}(u)$ [33], with ties broken by a fixed measurable rule. The verification proceeds through weak sign-pattern counting on score-difference functions [34–36] via Lemma 3 bounds cells through $\Phi_p(N) := \sum_{j=0}^{\min\{p,N\}} 2^j \binom{N}{j} \le (1+2N)^p$, which denotes the number of weak sign patterns in $\{-1,0,+1\}^N$ generated by N affine functions over $\mathbb{R}^p$, for every $N \ge 1$.

Corollary 3 (Affine source-score maps). *If $s_S(u) = Au + b$ for some $A \in \mathbb{R}^{K_{\mathrm{S}} \times d}$ and $b \in \mathbb{R}^{K_{\mathrm{S}}}$, then for every $m \ge 1$,*

$$\log \Pi_{\mathcal{R}}(m) \;\le\; \log \Phi_p\!\big(m \tbinom{K_{\mathrm{S}}}{2}\big) \;\le\; p \log\!\big(1 + 2m \tbinom{K_{\mathrm{S}}}{2}\big).$$

Corollary 4 (Fixed finite-depth ReLU source-score networks). *If s_S is a fixed feedforward ReLU network with L hidden layers of widths $q_1, \ldots, q_L$ (definition in Appendix D), then for every $m \ge 1$,*

$$\log \Pi_{\mathcal{R}}(m) \;\le\; \log \Phi_p\!\big(m \tbinom{K_{\mathrm{S}}}{2}\big) + \sum_{\ell=1}^{L} \log \Phi_p(m\, q_\ell) \;=\; O\big(p(L+1) \log m\big).$$

Remark 2. Theorems 1 and 2 separate the architecture-free generalization bound from per-architecture capacity verification (Condition A), mirroring how threshold and piecewise-linear bounds instantiate

general VC guarantees [37–39]. Corollary 4 refines Θ layer by layer across ReLU pre-activations [40, 41], yielding $\widetilde{O}((p+d_{\mathcal{H}})/\varepsilon^2)$ for affine scores and $\widetilde{O}((p(L+1)+d_{\mathcal{H}})/\varepsilon^2)$ for fixed-depth ReLU scores; extension routes to ConvNets [39, 42, 43] and Transformers [44, 45] are discussed in Section 8.

5 Matching Lower Bound on an Explicit Binary Construction

Sections 3 and 4 upper-bound the sample complexity of hard-label MR by $\widetilde{O}((d_{\mathcal{R}}+d_{\mathcal{H}})/\varepsilon^2)$. We now ask whether the additive $(d_{\mathcal{R}}+d_{\mathcal{H}})$ structure is an artifact of Proposition 1 or intrinsic to MR. We resolve this on an explicit binary family that realizes $\Pi_{\mathcal{R}}(m)=2^{d_{\mathcal{R}}}$ and $\Pi_{\mathcal{H}|\mathcal{R}}(m)=2^{d_{\mathcal{H}}}$ while keeping the target Bayes classifier inside $\mathcal{F}_{\mathrm{MR}}$ (i.e., $a_{\mathrm{MR}}=0$). Proofs are deferred to Appendix E.

5.1 A binary family separating reachability and relabeling

Proposition 3 (Binary MR family; full construction in Appendix E.1)**.** *For all integers $d_{\mathcal{R}}, d_{\mathcal{H}} \geq 1$ and $\gamma \in (0, 1/2]$, set $m := d_{\mathcal{R}}+d_{\mathcal{H}}$, $K_{\mathrm{T}} := 2$, and $K_{\mathrm{S}} := d_{\mathcal{H}}+2$. There exist a canonical tuple $x_{1:m}=(x_1,\dots,x_m)\in\mathcal{X}_{\mathrm{T}}^m$, a fixed $f_{\mathrm{S}}:\mathcal{X}_{\mathrm{S}}\to[K_{\mathrm{S}}]$, and wrapper classes $\mathcal{G}=\{g_\alpha:\alpha\in\{0,1\}^{d_{\mathcal{R}}}\}$ ($|\mathcal{G}|=2^{d_{\mathcal{R}}}$) and $\mathcal{H}=\{h_\beta:\beta\in\{0,1\}^{d_{\mathcal{H}}}\}$ ($|\mathcal{H}|=2^{d_{\mathcal{H}}}$) such that $\mathcal{F}_{\mathrm{MR}}=\{h_\beta\circ f_{\mathrm{S}}\circ g_\alpha\}$ realizes every $\omega=(\alpha,\beta)\in\{0,1\}^m$ on $x_{1:m}$, with $\Pi_{\mathcal{R}}(m')\leq 2^{d_{\mathcal{R}}}$ and $\Pi_{\mathcal{H}|\mathcal{R}}(m')\leq 2^{d_{\mathcal{H}}}$ for all $m'\geq 1$ and equality at $m'=m$. Let P_{T}^{ω} be uniform on $\{x_1,\dots,x_m\}$ with $P_{\mathrm{T}}^{\omega}(Y=\omega_i\mid X=x_i)=(1+\gamma)/2$ for $i\in[m]$, and define $\mathfrak{F}^{\mathrm{bin}}_{d_{\mathcal{R}},d_{\mathcal{H}}}(\gamma):=\{P_{\mathrm{T}}^{\omega}:\omega\in\{0,1\}^m\}$.*

Proposition 3 realizes $d_{\mathcal{R}}$ and $d_{\mathcal{H}}$ as exact reachability and conditional-relabeling log-growths on $x_{1:m}$, while placing the Bayes rule $x_i\mapsto\omega_i$ of each P_{T}^{ω} inside $\mathcal{F}_{\mathrm{MR}}$, so $a_{\mathrm{MR}}=0$ and $R_{\mathrm{T}}(\widehat{f}_n)-R_{\mathrm{T}}^*=\mathcal{E}_{\mathrm{T}}^{\mathrm{est}}(\widehat{f}_n)$ across $\mathfrak{F}^{\mathrm{bin}}_{d_{\mathcal{R}},d_{\mathcal{H}}}(\gamma)$.

5.2 Lower bound and matching consequence

With approximation error eliminated on $\mathfrak{F}^{\mathrm{bin}}_{d_{\mathcal{R}},d_{\mathcal{H}}}(\gamma)$, we ask whether any learning rule, not merely approximate ERM, can achieve smaller estimation error than Theorem 1. The lower bound on $\mathfrak{F}^{\mathrm{bin}}_{d_{\mathcal{R}},d_{\mathcal{H}}}(\gamma)$ follows from Assouad's lemma [28, 46] applied to the binary classification problem induced by Proposition 3, with margin $\gamma\asymp\sqrt{(d_{\mathcal{R}}+d_{\mathcal{H}})/n_{\mathrm{T}}}$.

Theorem 3 (Matching minimax lower bound)**.** *With $c_0=1$, $c_1=1/4$, and $c_2=e^{-1/4}/16$, for all integers $d_{\mathcal{R}},d_{\mathcal{H}}\geq 1$ and $n_{\mathrm{T}}\geq 1$ satisfying $d_{\mathcal{R}}+d_{\mathcal{H}}\leq c_0 n_{\mathrm{T}}$, on $\mathfrak{F}^{\mathrm{bin}}_{d_{\mathcal{R}},d_{\mathcal{H}}}(\gamma)$ with margin $\gamma=c_1\sqrt{(d_{\mathcal{R}}+d_{\mathcal{H}})/n_{\mathrm{T}}}$, every learning rule $\widehat{f}_n$ mapping n_{T} i.i.d. target samples to a measurable predictor satisfies $\sup_{P_{\mathrm{T}}\in\mathfrak{F}^{\mathrm{bin}}_{d_{\mathcal{R}},d_{\mathcal{H}}}(\gamma)}\mathbb{E}_{P_{\mathrm{T}}^{n_{\mathrm{T}}}}[\mathcal{E}_{\mathrm{T}}^{\mathrm{est}}(\widehat{f}_n)]\geq c_2\sqrt{(d_{\mathcal{R}}+d_{\mathcal{H}})/n_{\mathrm{T}}}$.*

Theorem 3 matches Theorem 1 on $\mathfrak{F}^{\mathrm{bin}}_{d_{\mathcal{R}},d_{\mathcal{H}}}(\gamma)$, where substituting $\log\Pi_{\mathcal{R}}(2n_{\mathrm{T}})\leq d_{\mathcal{R}}\log 2$ and $\log\Pi_{\mathcal{H}|\mathcal{R}}(2n_{\mathrm{T}})\leq d_{\mathcal{H}}\log 2$ into Eqn. (1) gives $\mathcal{E}_{\mathrm{T}}^{\mathrm{est}}(\widehat{f}_n)\leq\eta_n+4\sqrt{6}\sqrt{n_{\mathrm{T}}^{-1}((d_{\mathcal{R}}+d_{\mathcal{H}})\log 2+\log(1/\delta))}$. Operationally, $d_{\mathcal{R}}$ independent bits must be learned through reachability and $d_{\mathcal{H}}$ bits through conditional relabeling, with neither component substituting for the other (when $d_{\mathcal{R}}\gg d_{\mathcal{H}}$ or $d_{\mathcal{H}}\gg d_{\mathcal{R}}$, the larger capacity alone governs the minimax rate); note that Theorem 3 establishes tightness on this explicit binary family only.

Corollary 5 (Sample-complexity lower bound)**.** *For all $d_{\mathcal{R}},d_{\mathcal{H}}\geq 1$, every $\varepsilon\in(0,c_2/2)$, any learner $\widehat{f}_n$ achieving $\sup_{P_{\mathrm{T}}\in\mathfrak{F}^{\mathrm{bin}}_{d_{\mathcal{R}},d_{\mathcal{H}}}(\gamma)}\mathbb{E}_{P_{\mathrm{T}}^{n_{\mathrm{T}}}}[\mathcal{E}_{\mathrm{T}}^{\mathrm{est}}(\widehat{f}_n)]\leq\varepsilon$ at margin $\gamma=\min\{1/4,\,c_1\sqrt{(d_{\mathcal{R}}+d_{\mathcal{H}})/n_{\mathrm{T}}}\}$ requires $n_{\mathrm{T}}\geq c_2^2(d_{\mathcal{R}}+d_{\mathcal{H}})/\varepsilon^2=\Omega\big((d_{\mathcal{R}}+d_{\mathcal{H}})/\varepsilon^2\big)$ target samples.*

Corollary 5 matches Corollaries 1 and 2 up to constants and shows that the two growth sources contribute additively as they are *independently identifiable* on $\mathfrak{F}^{\mathrm{bin}}_{d_{\mathcal{R}},d_{\mathcal{H}}}(\gamma)$: $d_{\mathcal{R}}$ bits are learned through reachability and $d_{\mathcal{H}}$ bits through conditional relabeling, with neither side substituting for the other.

Sections 3–5 answer question ① with matching $\widetilde{O}((d_{\mathcal{R}}+d_{\mathcal{H}})/\varepsilon^2)$ and $\Omega((d_{\mathcal{R}}+d_{\mathcal{H}})/\varepsilon^2)$ rates. We next address question ②: when the estimation error vanishes, the wrapper class may still be separated from the target Bayes rule by the population residual a_{MR}, which we now characterize exactly.

6 Approximation Error: Identity and Certificate

We have bounded the estimation error of hard-label MR, but left open whether $\mathcal{F}_{\mathrm{MR}}$ contains a predictor close to the target Bayes optimum. Through the wrapper-side viewpoint, we show that the approximation error $a_{\mathrm{MR}} = R_{\mathrm{T}}^{\mathcal{F}_{\mathrm{MR}},*} - R_{\mathrm{T}}^{*}$ is still governed by the two-stage pipeline of MR, which asks whether the reachable source labels through $g \in \mathcal{G}$ preserve enough target-label information, and whether the allowed output mappings $h \in \mathcal{H}$ can use that information. See Appendix F-G for proofs.

6.1 Exact population decomposition

To isolate what prevents $R_{\mathrm{T}}^{\mathcal{F}_{\mathrm{MR}},*}$ from reaching R_{T}^{*} even with infinite target data, fix $g \in \mathcal{G}$ and write $Z_g := f_{\mathrm{S}}(g(X)) \in [K_{\mathrm{S}}]$ on $(X, Y) \sim P_{\mathrm{T}}$. Given Z_g, the best output mapping in $\mathcal{H}$ achieves $R_{\mathcal{H}}^{*}(g) := \inf_{h \in \mathcal{H}} \mathbb{E}_{P_{\mathrm{T}}}[\mathbf{1}\{h(Z_g) \neq Y\}]$, whereas optimizing over *all* unrestricted relabelings $h : [K_{\mathrm{S}}] \to [K_{\mathrm{T}}]$ achieves the channel Bayes risk $R_{\mathrm{all}}^{*}(g) := \inf_{h:[K_{\mathrm{S}}]\to[K_{\mathrm{T}}]} \mathbb{E}_{P_{\mathrm{T}}}[\mathbf{1}\{h(Z_g) \neq Y\}] = \mathbb{E}_{P_{\mathrm{T}}}[1 - \max_{y \in [K_{\mathrm{T}}]} P_{\mathrm{T}}(Y = y \mid Z_g)]$. Writing $R_{\mathrm{T}}^{\mathcal{F}_{\mathrm{MR}},*} = \inf_{g \in \mathcal{G}} R_{\mathcal{H}}^{*}(g)$, we have:

Proposition 4 (Exact decomposition of a_{MR})**.** *For every P_{T}, fixed measurable f_{S}, nonempty measurable $\mathcal{G}$, and nonempty $\mathcal{H} \subseteq [K_{\mathrm{T}}]^{[K_{\mathrm{S}}]}$,*

$$a_{\mathrm{MR}} \;=\; \inf_{g \in \mathcal{G}} \Big\{ \underbrace{R_{\mathrm{all}}^{*}(g) - R_{\mathrm{T}}^{*}}_{\text{source-label information loss}} \;+\; \underbrace{R_{\mathcal{H}}^{*}(g) - R_{\mathrm{all}}^{*}(g)}_{\text{target-relabeling restriction}} \Big\}, \tag{3}$$

where both underbraced summands are nonnegative for every $g \in \mathcal{G}$.

Proposition 4 attributes a_{MR} to two coupled wrapper bottlenecks evaluated at a shared g: *source-label information loss* arises whenever distinct target Bayes-decision regions collapse onto a single source label of Z_g so no relabeling can separate them, whereas *target-relabeling restriction* measures the cost of restricting the head to $\mathcal{H}$. In the extreme cases, information loss vanishes iff $R_{\mathrm{all}}^{*}(g) = R_{\mathrm{T}}^{*}$, and relabeling restriction vanishes iff $\mathcal{H}$ contains a channel-Bayes relabeling $h_g^{\star}$ satisfying $h_g^{\star}(k) \in \arg\max_{y \in [K_{\mathrm{T}}]} P_{\mathrm{T}}(Y = y \mid Z_g = k)$ on the support of Z_g (identically zero when $\mathcal{H} = [K_{\mathrm{T}}]^{[K_{\mathrm{S}}]}$). Note that both losses share g inside the infimum, minimizing one need not minimize the other.

6.2 Source-conditional upper bound

Although Proposition 4 is exact, evaluating $R_{\mathrm{all}}^{*}(g)$ and R_{T}^{*} requires unknown target conditional posteriors. We replace these with three source-conditional quantities computable from the source and target distributions, yielding an empirically tractable upper bound. For a source distribution P_{S} over $\mathcal{X}_{\mathrm{S}} \times [K_{\mathrm{S}}]$ with source-input marginal $P_{\mathrm{S},X}$, target-input marginal $P_{\mathrm{T},X}$, and measurable Bayes selectors $b_{\mathrm{T}} : \mathcal{X}_{\mathrm{T}} \to [K_{\mathrm{T}}]$ and $b_{\mathrm{S}} : \mathcal{X}_{\mathrm{S}} \to [K_{\mathrm{S}}]$, for $(g, h) \in \mathcal{G} \times \mathcal{H}$, define the target-side Bayes-mismatch term $\mathrm{mis}_{\mathrm{hard}}(g, h)$ and source-side relabeled-classifier error $\epsilon_{\mathrm{hard}}(h)$ as

$$\mathrm{mis}_{\mathrm{hard}}(g, h) := \mathbb{P}_{X \sim P_{\mathrm{T}}}\big(b_{\mathrm{T}}(X) \neq h(b_{\mathrm{S}}(g(X)))\big), \;\; \epsilon_{\mathrm{hard}}(h) := \mathbb{P}_{U \sim P_{S,X}}\big(h(f_{\mathrm{S}}(U)) \neq h(b_{\mathrm{S}}(U))\big).$$

With $\phi_h(u) = \mathbf{1}\{h(f_{\mathrm{S}}(u)) \neq h(b_{\mathrm{S}}(u))\}$, define the head-dependent pushforward discrepancy [47]

$$\mathrm{disc}_{\mathrm{hard}}(g, h) := \big|\mathbb{E}_{X \sim P_{\mathrm{T}}} \phi_h(g(X)) - \mathbb{E}_{U \sim P_{S,X}} \phi_h(U)\big|.$$

Theorem 4 (Population upper certificate for a_{MR})**.** *Suppose P_{T} and P_{S} admit measurable Bayes-optimal classifiers $b_{\mathrm{T}} : \mathcal{X}_{\mathrm{T}} \to [K_{\mathrm{T}}]$ and $b_{\mathrm{S}} : \mathcal{X}_{\mathrm{S}} \to [K_{\mathrm{S}}]$. Then*

$$a_{\mathrm{MR}} \;\leq\; \inf_{(g,h) \in \mathcal{G} \times \mathcal{H}} \Big\{ \underbrace{\mathrm{mis}_{\mathrm{hard}}(g, h)}_{\text{mismatch}} \;+\; \underbrace{\epsilon_{\mathrm{hard}}(h)}_{\text{source error}} \;+\; \underbrace{\mathrm{disc}_{\mathrm{hard}}(g, h)}_{\text{discrepancy}} \Big\}. \tag{4}$$

Theorem 4 upper-bounds a_{MR} through three operational terms, where $\mathrm{mis}_{\mathrm{hard}}(g, h)$ measures whether the relabeled source Bayes rule $h \circ b_{\mathrm{S}} \circ g$ agrees with b_{T} on $P_{\mathrm{T},X}$, $\epsilon_{\mathrm{hard}}(h)$ measures the h-relabeled error of f_{S} against b_{S} on $P_{\mathrm{S},X}$, and $\mathrm{disc}_{\mathrm{hard}}(g, h)$ measures the shift of ϕ_h between $g_{\#}P_{\mathrm{T},X}$ and $P_{\mathrm{S},X}$. When $f_{\mathrm{S}} = b_{\mathrm{S}}$ pointwise on $\mathcal{X}_{\mathrm{S}}$, $\epsilon_{\mathrm{hard}}(h) = \mathrm{disc}_{\mathrm{hard}}(g, h) = 0$ and Eqn. (4) reduces to $\inf_{g,h} \mathrm{mis}_{\mathrm{hard}}(g, h)$. Plugging consistent estimators of b_{S} and b_{T} into Eqn. (4) yields i.i.d.-estimable proxies inheriting their rates [24].

Table 1: Empirical verification under exact ERM.

(a) Synthetic Estimation $\mathcal{E}_{\mathrm{T}}^{\mathrm{est}}(\widehat{f}_n) \downarrow$ (Input $\equiv$ Head)

Sample size	$r=0$	$r=2$	$r=4$	$r=6$	$r=8$
$n_{\mathrm{T}}=8$	0.0000	0.0475	0.0995	0.1340	0.1765
$n_{\mathrm{T}}=32$	0.0000	0.0080	0.0155	0.0255	0.0295
$n_{\mathrm{T}}=64$	0.0000	0.0020	0.0035	0.0045	0.0050
$n_{\mathrm{T}}=512$	0.0000	0.0000	0.0000	0.0000	0.0000

(b) Synthetic Approximation Split at Matched $R_{\mathrm{T}}^{\mathcal{F}_{\mathrm{MR}},*}$

Mechanism (m)	$R_{\mathrm{T}}^{\mathcal{F}_{\mathrm{MR}},*}$	I	H	$\widehat{I}$	$\widehat{H}$
Merge labels (2)	0.2000	0.1000	0.0000	0.0979	0.0000
Restrict head (2)	0.2000	0.0000	0.1000	0.0000	0.1000
Merge labels (8)	0.5000	0.4000	0.0000	0.3923	0.0000
Restrict head (8)	0.5000	0.0000	0.4000	0.0000	0.4000

(c) Pretrained Neural Sources on 90°-Rotated MNIST Parity (Held-Out 0–1 Risk $\downarrow$)

		Frozen MLP Source				**Frozen Minimal-LeNet Source**			
Wrapper $(\mathcal{G},\mathcal{H})$	Regime	n_{T}=50	200	800	3200	n_{T}=50	200	800	3200
$(\mathcal{G}_{\mathrm{id}},\mathcal{H}_{\mathrm{thr}})$	Fixed, threshold	0.4701	0.4515	0.4439	0.4385	0.4206	0.4029	0.3979	0.3960
$(\mathcal{G}_{\mathrm{id}},\mathcal{H}_{\mathrm{full}})$	Fixed, full binary	0.4416	0.4069	0.4006	0.3980	0.4006	0.3634	0.3552	0.3538
$(\mathcal{G}_{\mathrm{rot}},\mathcal{H}_{\mathrm{thr}})$	4-rot., threshold	0.3967	0.3718	0.3636	0.3613	0.3621	0.3460	0.3374	0.3352
$(\mathcal{G}_{\mathrm{rot}},\mathcal{H}_{\mathrm{full}})$	4-rot., full binary	**0.0098**	**0.0098**	**0.0098**	**0.0098**	**0.0058**	**0.0058**	**0.0058**	**0.0058**
Untrained $(\mathcal{G}_{\mathrm{rot}},\mathcal{H}_{\mathrm{full}})$	Random-weight source		–		0.4355		–		0.5075

Remark 3. Proposition 4 is an exact, slack-free identity, whereas Theorem 4 is a one-directional upper certificate (smallness certifies small a_{MR}, while large values are inconclusive). Appendix G constructs a converse: (i) bounding $R_{\mathrm{T}}(f_{g,h}) - R_{\mathrm{T}}^*$ by $\mathbb{P}_{P_{\mathrm{T},X}}(f_{g,h}(X) \neq b_{\mathrm{T}}(X))$ is loose wherever the target posterior is near-uniform without an unassumed low-noise margin condition on P_{T}; and (ii) once b_{S} is fixed or estimated, $\mathrm{disc}_{\mathrm{hard}}(g,h)$ is estimable from unlabeled inputs because ϕ_h is evaluated pointwise at the tested h rather than taking a supremum over $\mathcal{H}\Delta\mathcal{H}$ [48]. As such, Theorem 4 is a wrapper-side upper bound on a_{MR}, not an $\mathcal{H}\Delta\mathcal{H}$-style class-symmetric transfer divergence [48].

Extension to logit-exposing MR. The approximation analysis also transfers to source models exposing real-valued logit vectors $U_g = s_{\mathrm{S}}(g(X)) \in \mathbb{R}^{K_{\mathrm{S}}}$, with discrete relabeling replaced by a learnable linear map $W \in \mathcal{W} \subseteq \mathbb{R}^{K_{\mathrm{T}} \times K_{\mathrm{S}}}$ (identifying each matrix W with $u \mapsto Wu$). Appendix B establishes exact population identities (Propositions 5 and 6), aand a population certificate (Theorem 5), which is empirically estimable over a data-independent candidate set of MR wrappers (Theorem 6).

7 Controlled Empirical Verification

Theorem 1 predicts that finite-sample estimation error responds symmetrically to input-side reachability and output-side relabeling capacity, while Proposition 4 attributes population approximation error to lost source-label distinctions or head restriction. To isolate wrapper structure from optimizer dynamics, we evaluate both predictions under exact ERM ($\eta_n = 0$) on a synthetic benchmark and on frozen MNIST classifiers (Appendix K; code at `https://github.com/zyecs/repgm-slt`).

Capacity symmetry and separable approximation bottlenecks. Consider a uniform binary task over 16 basis vectors in $\mathbb{R}^{16}$, grouped into 8 ordered pairs whose clean labels $(0,1)$ flip independently with probability 0.1 ($R_{\mathrm{T}}^* = 0.1$) under a bijective source $f_{\mathrm{S}}(e_i) = i$ (Appendix K.1). On the first $r \in \{0,2,4,6,8\}$ pairs, independently allowing g to swap each pair's vectors under a fixed parity head ($\log_2 \Pi_{\mathcal{R}} = r$, $\log_2 \Pi_{\mathcal{H}|\mathcal{R}} = 0$) realizes the exact same 2^r predictors ($a_{\mathrm{MR}} = 0$) as fixing $g = \mathrm{id}$ and independently flipping the head's pair assignments ($\log_2 \Pi_{\mathcal{R}} = 0$, $\log_2 \Pi_{\mathcal{H}|\mathcal{R}} = r$). Both parameterizations yield identical $\mathcal{E}_{\mathrm{T}}^{\mathrm{est}}(\widehat{f}_n)$ across all (r, n_{T}) (Table 1a), rising at $n_{\mathrm{T}} = 8$ from 0.0475 ($r=2$) to 0.1765 ($r=8$) before vanishing by $n_{\mathrm{T}} = 512$. Across $m \in \{0,2,4,6,8\}$ affected pairs, two wrappers also share the exact suboptimal risk $R_{\mathrm{T}}^{\mathcal{F}_{\mathrm{MR}},*} = 0.1 + 0.05m$ via opposite terms of Proposition 4 (Table 1b): merging each affected pair onto one source label under an unrestricted head pools opposite Bayes labels into information loss ($I = 0.05m$, $H = 0$), whereas keeping all 16 source labels distinct while fixing the wrong label 1 on each pair's first vector places the same $0.05m$ residual into head restriction ($I = 0$, $H = 0.05m$), tracked by plug-ins $(\widehat{I}, \widehat{H})$.

Joint input and output adaptation on pretrained neural sources. Reprogramming frozen upright-MNIST [49] classifiers (MLP and minimal LeNet exposing only predicted digit IDs in $\{0,\ldots,9\}$) for binary parity on unseen 90°-rotated digits illustrates how estimation decay and approximation floors interact across $n_{\mathrm{T}} \in \{50, 200, 800, 3200\}$ (Appendix K.2; Table 1c). Crossing identity access $\mathcal{G}_{\mathrm{id}}$ versus four rotations $\mathcal{G}_{\mathrm{rot}} = \{\mathrm{rot}_0, \mathrm{rot}_{90}, \mathrm{rot}_{180}, \mathrm{rot}_{270}\}$ with single-threshold heads $\mathcal{H}_{\mathrm{thr}}$ versus

full binary relabelings $\mathcal{H}_{\text{full}} = \{0,1\}^{10}$, only $(\mathcal{G}_{\text{rot}}, \mathcal{H}_{\text{full}})$ reaches sub-1% error, locking onto rot_{270} already at $n_{\text{T}} = 50$. When either stage is restricted, growing n_{T} from 50 to 3200 reduces estimation error ($0.4416 \to 0.3980$ under $(\mathcal{G}_{\text{id}}, \mathcal{H}_{\text{full}})$ and $0.3967 \to 0.3613$ under $(\mathcal{G}_{\text{rot}}, \mathcal{H}_{\text{thr}})$ on MLP) before stalling at high approximation floors from unaligned digits or non-alternating heads. Applying $(\mathcal{G}_{\text{rot}}, \mathcal{H}_{\text{full}})$ to untrained networks stays near chance, confirming that wrapper capacity helps only when the frozen source already separates task-relevant structure.

8 On Understanding the Practice of Model Reprogramming

Our *wrapper-side* analysis turns the MR wrapper into a diagnostic tool: Theorem 1 and Proposition 4 attribute finite-sample estimation cost and the approximation residual to specific wrapper stages.

Two failure modes of MR. Proposition 4 splits a_{MR} into source-label information loss $R^*_{\text{all}}(g) - R^*_{\text{T}}$, governed by $Z_g = f_{\text{S}}(g(X))$, and target-relabeling restriction $R^*_{\mathcal{H}}(g) - R^*_{\text{all}}(g)$, governed by $\mathcal{H}$. Notably, Proposition 4 depends on whether the partition induced by Z_g preserves target Bayes decisions rather than whether $g_{\#}P_{\text{T},X}$ matches $P_{\text{S},X}$ in input space, whereas input alignment enters Theorem 4 via $\text{disc}_{\text{hard}}(g,h)$ only as a sufficient condition for small a_{MR}. Because the two terms require distinct remedies (modifying g vs. enlarging $\mathcal{H}$), separating them is key to diagnosing failures.

Input-side trade-off. Theorem 2 and Proposition 4 expose a wrapper-side bias–variance trade-off [50, 51]: enlarging $\mathcal{G}_p$ via p can make Z_g more informative about Y and reduce $\inf_g R^*_{\text{all}}(g)$, at the cost of sample complexity $\widetilde{O}((c_{\mathcal{R}}p + d_{\mathcal{H}})/\varepsilon^2)$. Recent input-side designs sit at complementary points along this curve: Cai et al. [14] *raised* p with a small ConvNet generating input-dependent masks to cut approximation loss, whereas Jin et al. [52] *lowered* p via low-rank sub-region matrices to curb the statistical exponent without empirically sacrificing approximation power.

Output-side implications. Our results distinguish two subsets of $[K_{\text{S}}]$ on the output side. In finite samples, Theorem 1 charges $\mathcal{H}$ via $\Pi_{\mathcal{H}|\mathcal{R}}$, scaling with the sample-wise reachable alphabet size $t_{\mathcal{R}}(m) \le K_{\text{S}}$ rather than K_{S} (e.g., $K_{\text{S}} = 1000$ on ImageNet [53] versus $K_{\text{T}} = 10$ on CIFAR-10). At the population level, Proposition 4 defines the optimal head given g as a channel-Bayes relabeling on the active support $\mathcal{Z}_g := \{k \in [K_{\text{S}}] : \mathbb{P}_{P_{\text{T}}}(Z_g = k) > 0\}$. This shifts head design from picking a one-to-one injection $[K_{\text{S}}] \to [K_{\text{T}}]$ [13] to ensuring expressivity on $\mathcal{Z}_g$: when every Bayes relabeling on $\mathcal{Z}_g$ is many-to-one, injections incur positive $R^*_{\mathcal{H}}(g) - R^*_{\text{all}}(g)$, whereas many-to-many or Bayesian-guided mappings [15] close this gap at cost governed only by $t_{\mathcal{R}}$ (Example 1).

Limitations and outlook. Three boundaries of our analysis point to natural extensions. First, while Theorem 1 holds for any measurable f_{S}, the closed-form rate in Theorem 2 relies on the polynomial cell-partition condition (Condition A), which we verified for affine score maps and finite-depth ReLU networks (Corollaries 3 and 4). Extending explicit rates to broader architectures depends on how f_{S} transforms wrapper parameters. In frozen piecewise-affine ConvNets [42], convolutional, ReLU, residual, and pooling layers keep pre-activations piecewise affine in θ, motivating computation-graph cell bounds, whereas smooth operations in Transformers [44] (e.g., softmax attention, LayerNorm, and GELU) break piecewise affinity and instead call for margin-based covering-number or Rademacher bounds [39, 43, 45]. Second, the growth factorization in Proposition 1 relies on f_{S} taking values in a finite alphabet $[K_{\text{S}}]$. When the source model instead exposes continuous logits in $\mathbb{R}^{K_{\text{S}}}$, our approximation identities and held-out certificates still hold (Appendix B), but the PAC bound of Theorem 1 does not extend automatically; proving same-sample ERM guarantees there calls for combining capacity bounds on $\{x \mapsto s_{\text{S}}(g(x)) : g \in \mathcal{G}\}$ and norm-constrained heads [24, 54] with vector contraction [55] and calibrated surrogate transfer [56, 57]. Third, Theorem 4 gives a one-directional upper bound for a fixed pair (g,h) (Remark 3), whereas practical methods update (g_t, h_t) iteratively. Analyzing how training dynamics reduce the two terms of Proposition 4, and combining our wrapper-side bounds with source-centered reprogrammability conditions [17, 21, 22], would link wrapper design to both optimization and source selection.

Summary. We developed a finite-sample analysis of hard-label MR: the estimation analysis factorizes $\mathcal{F}_{\text{MR}}$'s sample-wise complexity into reachability and conditional-relabeling growth (Theorem 1), giving closed-form PAC rates on additive transformations (Theorem 2) matched by a minimax lower bound (Theorem 3), while the approximation analysis splits the residual into source-label information loss and output-mapping restriction (Proposition 4 and Theorem 4), with both mechanisms verified in synthetic and pretrained-MNIST studies (Section 7). Our analysis complements source-centered MR theory and offers a wrapper-side viewpoint to guide the design of MR wrappers for a given frozen f_{S}.

Acknowledgments and Disclosure of Funding

The authors thank the anonymous reviewers and the Area Chair for their constructive feedback and insightful suggestions during the review process. This research was partly supported by the University of Melbourne's Research Computing Services and the Petascale Campus Initiative. Feng Liu is supported by the ARC with grant numbers DP230101540, DE240101089 and DP260103996. Pin-Yu Chen is solely funded by IBM Research.

References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," in *NeurIPS*, 2020.

[2] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. Casas, L. A. Hendricks, J. Welbl, A. Clark *et al.*, "Training compute-optimal large language models," in *NeurIPS*, 2022.

[3] C. Zhou, Q. Li, C. Li, J. Yu, Y. Liu, G. Wang, K. Zhang, C. Ji, Q. Yan, L. He, H. Peng, J. Li, J. Wu, Z. Liu, P. Xie, C. Xiong, J. Pei, P. S. Yu, and L. Sun, "A comprehensive survey on pretrained foundation models: A history from BERT to ChatGPT," *arXiv preprint arXiv:2302.09419*, 2023.

[4] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, and J.-R. Wen, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.

[5] J. Gu, Z. Han, S. Chen, A. Beirami, B. He, G. Zhang, R. Liao, Y. Qin, V. Tresp, and P. Torr, "A systematic survey of prompt engineering on vision-language foundation models," *arXiv preprint arXiv:2307.12980*, 2023.

[6] Z. Ye, C. Cai, R. Dong, J. Qi, L. Feng, P.-Y. Chen, and F. Liu, "Neural network reprogrammability: A unified theme on model reprogramming, prompt tuning, and prompt instruction," *arXiv preprint arXiv:2506.04650*, 2025.

[7] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *ICML*, 2019.

[8] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," in *EMNLP*, 2021.

[9] M. Luo, X. Xu, Y. Liu, P. Pasupat, and M. Kazemi, "In-context learning with retrieved demonstrations for language models: A survey," *TMLR*, 2024.

[10] G. F. Elsayed, I. Goodfellow, and J. Sohl-Dickstein, "Adversarial reprogramming of neural networks," in *ICLR*, 2019.

[11] P. Neekhara, S. Hussain, S. Dubnov, and F. Koushanfar, "Adversarial reprogramming of text classification neural networks," in *EMNLP/IJCNLP*, 2019, pp. 5215–5224.

[12] P.-Y. Chen, "Model reprogramming: Resource-efficient cross-domain machine learning," in *AAAI*, 2024.

[13] A. Chen, Y. Yao, P.-Y. Chen, Y. Zhang, and S. Liu, "Understanding and improving visual prompting: A label-mapping perspective," in *CVPR*, 2023.

[14] C. Cai, Z. Ye, L. Feng, J. Qi, and F. Liu, "Sample-specific masks for visual reprogramming-based prompting," in *ICML*, 2024.

[15] ——, "Bayesian-guided label mapping for visual reprogramming," in *NeurIPS*, 2024.

[16] Y.-Y. Tsai, P.-Y. Chen, and T.-Y. Ho, "Transfer learning without knowing: Reprogramming black-box machine learning models with scarce data and limited resources," in *ICML*, 2020.

[17] C.-H. H. Yang, Y.-Y. Tsai, and P.-Y. Chen, "Voice2series: Reprogramming acoustic models for time series classification," in *ICML*, 2021.

[18] I. Melnyk, V. Chenthamarakshan, P.-Y. Chen, P. Das, A. Dhurandhar, I. Padhi, and D. Das, "Reprogramming pretrained language models for antibody sequence infilling," in *International conference on machine learning*. PMLR, 2023, pp. 24 398–24 419.

[19] R. Vinod, P.-Y. Chen, and P. Das, "Reprogramming pretrained language models for protein sequence representation learning," *Digital Discovery*, vol. 4, no. 6, pp. 1591–1601, 2025.

[20] P. Neekhara, S. Hussain, J. Du, S. Dubnov, F. Koushanfar, and J. McAuley, "Cross-modal adversarial reprogramming," in *CVPR*, 2022.

[21] M. Englert and R. Lazic, "Adversarial reprogramming revisited," in *NeurIPS*, 2022.

[22] M.-Y. Chung, J. Fan, H. Ye, Q. Wang, W.-C. Shen, C.-M. Yu, P.-Y. Chen, and S.-Y. Kuo, "Model reprogramming demystified: A neural tangent kernel perspective," *arXiv preprint arXiv:2506.00620*, 2025.

[23] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 2013.

[24] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning*, 2nd ed. MIT Press, 2018.

[25] V. N. Vapnik, "An overview of statistical learning theory," *IEEE transactions on neural networks*, vol. 10, no. 5, pp. 988–999, 1999.

[26] H. Bahng, A. Jahanian, S. Sankaranarayanan, and P. Isola, "Exploring visual prompts for adapting large-scale models," *arXiv preprint arXiv:2203.17274*, 2022.

[27] H.-A. Tsao, L. Hsiung, P.-Y. Chen, S. Liu, and T.-Y. Ho, "Autovp: An automated visual prompting framework and benchmark," in *ICLR*, 2024.

[28] B. Yu, "Assouad, fano, and le cam," in *Festschrift for Lucien Le Cam: research papers in probability and statistics*. Springer, 1997.

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

[30] B. K. Natarajan, "On learning sets and functions," *Machine Learning*, 1989.

[31] A. Daniely, S. Sabato, S. Ben-David, and S. Shalev-Shwartz, "Multiclass learnability and the erm principle," *JMLR*, 2015.

[32] S. Ben-David, N. Cesa-Bianchi, and P. M. Long, "Characterizations of learnability for classes of $\{0, \ldots, n\}$-valued functions," in *Proceedings of the fifth annual workshop on Computational learning theory*, 1992, pp. 333–340.

[33] K. Crammer and Y. Singer, "On the algorithmic implementation of multiclass kernel-based vector machines," *JMLR*, 2001.

[34] T. Zaslavsky, *Facing up to arrangements: Face-count formulas for partitions of space by hyperplanes*. American Mathematical Soc., 1975, vol. 154.

[35] H. E. Warren, "Lower bounds for approximation by nonlinear manifolds," *Transactions of the American Mathematical Society*, 1968.

[36] T. M. Cover, "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition," *IEEE Transactions on Electronic Computers*, vol. EC-14, no. 3, pp. 326–334, 1965.

[37] E. B. Baum and D. Haussler, "What size net gives valid generalization?" *Neural Computation*, vol. 1, no. 1, pp. 151–160, 1989.

[38] M. Karpinski and A. Macintyre, "Polynomial bounds for VC dimension of sigmoidal and general Pfaffian neural networks," *Journal of Computer and System Sciences*, vol. 54, no. 1, pp. 169–176, 1997.

[39] P. L. Bartlett, N. Harvey, C. Liaw, and A. Mehrabian, "Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks," *JMLR*, 2019.

[40] G. Montúfar, R. Pascanu, K. Cho, and Y. Bengio, "On the number of linear regions of deep neural networks," in *NeurIPS*, 2014.

[41] M. Raghu, B. Poole, J. Kleinberg, S. Ganguli, and J. Sohl-Dickstein, "On the expressive power of deep neural networks," in *ICML*, 2017.

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[43] P. M. Long and H. Sedghi, "Generalization bounds for deep convolutional neural networks," in *ICLR*, 2020.

[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017.

[45] J. Trauger and A. Tewari, "Sequence length independent norm-based generalization bounds for transformers," in *AISTATS*, 2024, pp. 1405–1413.

[46] A. B. Tsybakov, *Introduction to Nonparametric Estimation*, ser. Springer Series in Statistics. New York, NY: Springer, 2009.

[47] Y. Mansour, M. Mohri, and A. Rostamizadeh, "Domain adaptation: Learning bounds and algorithms," in *COLT*, 2009.

[48] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine learning*, vol. 79, no. 1, pp. 151–175, 2010.

[49] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[50] S. Geman, E. Bienenstock, and R. Doursat, "Neural networks and the bias/variance dilemma," *Neural computation*, 1992.

[51] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer New York Inc., 2001.

[52] C. Jin, Y. Li, M. Zhao, S. Zhao, Z. Wang, X. He, L. Han, T. Che, and D. N. Metaxas, "Lor-vp: Low-rank visual prompting for efficient vision model adaptation," in *ICLR*, 2025.

[53] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[54] P. L. Bartlett and S. Mendelson, "Rademacher and gaussian complexities: Risk bounds and structural results," *JMLR*, 2002.

[55] A. Maurer, "A vector-contraction inequality for rademacher complexities," in *ALT*, 2016.

[56] T. Zhang, "Statistical analysis of some multi-category large margin classification methods," *JMLR*, 2004.

[57] A. Tewari and P. L. Bartlett, "On the consistency of multiclass classification methods." *JMLR*, 2007.

[58] Y. Lee, Y. Lin, and G. Wahba, "Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data," *Journal of the American Statistical Association*, 2004.

[59] M. J. Kearns, R. E. Schapire, and L. M. Sellie, "Toward efficient agnostic learning," in *Proceedings of the fifth annual workshop on Computational learning theory*, 1992, pp. 341–352.

[60] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019, pp. 8024–8035.

Appendix Contents

A Notations and Assumptions **16**

B Approximation Analysis on Logit-Exposing Source Models **20**

C Proof of the Factorized Excess-Risk Bound and Comparison Chain **25**

D Proofs for the Additive Specialization and Representation-Side Discharges **30**

E Proof of the Lower-Bound Result **36**

F Proofs for Exact Representation-Conditioned Approximation Identities **41**

G Proof of the Hard-Label Source-Conditional Upper Certificate **45**

H Proofs for the Score-Output Certificates and Empirical Corollary **47**

I Uniform Empirical Score-Output Certificate over a Fixed Candidate Class **52**

J Worked Examples **56**

K Controlled Empirical Verification: Protocols and Complete Results **58**

A Notations and Assumptions

This appendix consolidates all notation used across the main text and appendices into a four-layer reference table (Table 2), records our global tie-breaking, reachable-alphabet, and degenerate-regime conventions, and indexes the mathematical assumptions (`A1`–`A17` in Table 3) by the exact results that invoke them.

Global tie-breaking, reachable-alphabet, and edge conventions. Whenever a classifier is induced by maximizing a real-valued score vector $t = (t_1, \dots, t_K) \in \mathbb{R}^K$ ($K \geq 1$), ties among maximizers are broken deterministically in lexicographic order:

$$\operatorname{lexargmax}_{j\in[K]} t_j \;:=\; \min \arg\max_{j\in[K]} t_j.$$

Because $\{t \in \mathbb{R}^K : \operatorname{lexargmax}_{j\in[K]} t_j = k\}$ equals $\bigcap_{j<k}\{t_k > t_j\} \cap \bigcap_{j>k}\{t_k \geq t_j\}$, a finite intersection of open and closed half-spaces, composing lexargmax with any measurable score map preserves measurability (Section 4.3 and Appendix B). For any sample size $m \geq 1$ and source-label vector $z = (z_1, \dots, z_m) \in [K_{\mathrm{S}}]^m$, we write

$$T(z) \;:=\; \{z_1, \dots, z_m\} \;\subseteq\; [K_{\mathrm{S}}], \qquad |z|_{\#} \;:=\; |T(z)|,$$

for the set of distinct source labels appearing in z and its cardinality (Example 1). Given a target tuple $x_{1:m} = (x_1, \dots, x_m) \in \mathcal{X}_{\mathrm{T}}^m$, the reachable source-label pattern set is $\mathrm{Rea}_{\mathcal{G}}(x_{1:m}) := \{(r(x_1), \dots, r(x_m)) : r \in \mathcal{R}\} \subseteq [K_{\mathrm{S}}]^m$, and the realizable target-label pattern set on $z \in [K_{\mathrm{S}}]^m$ is $\mathrm{Rel}_{\mathcal{H}}(z) := \{(h(z_1), \dots, h(z_m)) : h \in \mathcal{H}\} \subseteq [K_{\mathrm{T}}]^m$ (Definition 2). The worst-case reachable alphabet size and restricted head Natarajan dimension on samples of size m are

$$t_{\mathcal{R}}(m) \;:=\; \sup_{x_{1:m}\in\mathcal{X}_{\mathrm{T}}^m} \sup_{z\in\mathrm{Rea}_{\mathcal{G}}(x_{1:m})} |z|_{\#}, \qquad d_{\mathcal{H}}^{\mathcal{R}}(m) \;:=\; \sup_{x_{1:m}\in\mathcal{X}_{\mathrm{T}}^m} \sup_{z\in\mathrm{Rea}_{\mathcal{G}}(x_{1:m})} \mathrm{Ndim}\big(\mathcal{H}|_{T(z)}\big).$$

Throughout the main text and appendices, we adopt four standard degenerate-regime conventions: the empty product satisfies $\prod_{s=1}^{0}(\cdot) := 1$, the zeroth combinatorial power satisfies $\binom{K_{\mathrm{T}}}{2}^0 := 1$ when $K_{\mathrm{T}} = 1$, the additive-MR output-capacity term $B_m := d_m \log\big(e\binom{K_{\mathrm{T}}}{2} m/d_m\big)$ with $d_m := d_{\mathcal{H}} \wedge m$ satisfies $B_m := 0$ when $d_m = 0$ or $K_{\mathrm{T}} = 1$ (Theorem 2), and $0\log(a/0) := 0$ for every $a \geq 0$. Following standard empirical-process conventions [24, 54], whenever an empirical supremum or symmetrization step is taken over an uncountable hypothesis or loss class (including $\mathcal{F}_{\mathrm{MR}}$ in Theorems 1 and 2 and Corollaries 1 and 2 and the score-output classes in Appendix I), the class is assumed pointwise measurable (i.e., admitting a countable subset dense under pointwise convergence) so that empirical suprema and Rademacher averages are measurable and symmetrization applies.

Table 2: Grouped summary of notation across the four structural layers of the analysis.

Symbol	Meaning and definition
(1) Spaces and Distributions	
$\mathcal{X}_{\mathrm{T}}$, $\mathcal{X}_{\mathrm{S}}$	Target and source input spaces ($\mathcal{X}_{\mathrm{T}} \subseteq \mathbb{R}^d$ and $\mathcal{X}_{\mathrm{S}} = \mathbb{R}^d$ in the additive setting of Section 4).
$[K_{\mathrm{T}}]$, $[K_{\mathrm{S}}]$	Finite target and source label alphabets, where $[K] := \{1, \dots, K\}$.
P_{T}, P_{S}	Target joint distribution on $\mathcal{X}_{\mathrm{T}} \times [K_{\mathrm{T}}]$ and source joint distribution on $\mathcal{X}_{\mathrm{S}} \times [K_{\mathrm{S}}]$.
$P_{\mathrm{T},X}$, $P_{\mathrm{S},X}$; $g_{\#}P_{\mathrm{T},X}$	Target- and source-input marginals, and pushforward of $P_{\mathrm{T},X}$ under $g : \mathcal{X}_{\mathrm{T}} \to \mathcal{X}_{\mathrm{S}}$.
D_{T}, D_{S}; n_{T}, n_{S}	i.i.d. target sample $\{(X_i, Y_i)\}_{i=1}^{n_{\mathrm{T}}} \sim P_{\mathrm{T}}^{n_{\mathrm{T}}}$ and unlabeled source sample $\{U_j\}_{j=1}^{n_{\mathrm{S}}} \sim P_{\mathrm{S},X}^{n_{\mathrm{S}}}$.
b_{T}, b_{S}	Measurable Bayes-optimal classifiers $b_{\mathrm{T}} : \mathcal{X}_{\mathrm{T}} \to [K_{\mathrm{T}}]$ and $b_{\mathrm{S}} : \mathcal{X}_{\mathrm{S}} \to [K_{\mathrm{S}}]$.
$\eta_y^g(u)$	Regular conditional posterior $\mathbb{P}_{P_{\mathrm{T}}}(Y = y \mid U_g = u)$ on standard Borel spaces (Appendix B.2).

Symbol	Meaning and definition
(2) Wrapper and Hypothesis Classes	
f_{S}; s_{S}, B_s	Frozen hard-label source classifier $f_{\mathrm{S}} : \mathcal{X}_{\mathrm{S}} \to [K_{\mathrm{S}}]$ and frozen source score map $s_{\mathrm{S}} : \mathcal{X}_{\mathrm{S}} \to [-B_s, B_s]^{K_{\mathrm{S}}}$ ($B_s \in [0, \infty)$).
lexargmax; $\hat{b}_{\mathrm{S}}$, $c(u)$	Lexicographic argmax $\min \arg\max$, score-induced source classifier $\hat{b}_{\mathrm{S}}(u) := \mathrm{lexargmax}_{k \in [K_{\mathrm{S}}]} s_{\mathrm{S},k}(u)$, and one-hot vector $c(u) := e_{\hat{b}_{\mathrm{S}}(u)} \in \mathbb{R}^{K_{\mathrm{S}}}$.
$g \in \mathcal{G}$; $\mathcal{G}_p$, M	Learnable input transformation $\mathcal{X}_{\mathrm{T}} \to \mathcal{X}_{\mathrm{S}}$ and additive family $\mathcal{G}_p = \{g_\theta(x) = x + M\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$ with fixed $M \in \mathbb{R}^{d \times p}$.
$h \in \mathcal{H}$; $W \in \mathcal{W}$, B_h	Discrete output mapping $h : [K_{\mathrm{S}}] \to [K_{\mathrm{T}}]$ and linear score-output map $W \in \mathcal{W} \subseteq \mathbb{R}^{K_{\mathrm{T}} \times K_{\mathrm{S}}}$ (identified with $u \mapsto Wu$) with $\|W\|_F \le B_h$ ($B_h \in [0, \infty)$).
$\mathcal{R}$; $\mathcal{F}_{\mathrm{MR}}$, $\mathcal{F}_{\mathrm{score}}$	Reachable source-label class $\{f_{\mathrm{S}} \circ g : g \in \mathcal{G}\}$, hard-label wrapper class $\{f_{g,h} := h \circ f_{\mathrm{S}} \circ g\}$, and logit-exposing wrapper class $\{f_{g,W}\}$ (Definition 4).
$\widehat{R}_n(f)$; $\widehat{f}_n$, η_n	Empirical 0–1 risk $n_{\mathrm{T}}^{-1} \sum_{i=1}^{n_{\mathrm{T}}} \mathbf{1}\{f(X_i) \neq Y_i\}$ and measurable η_n-approximate ERM satisfying $\widehat{R}_n(\widehat{f}_n) \le \inf_{f \in \mathcal{F}_{\mathrm{MR}}} \widehat{R}_n(f) + \eta_n$ ($\eta_n \ge 0$).
$\mathfrak{F}^{\mathrm{bin}}_{d_{\mathcal{R}}, d_{\mathcal{H}}}(\gamma)$	Explicit binary target distribution family on $m = d_{\mathcal{R}} + d_{\mathcal{H}}$ canonical points with margin $\gamma \in (0, 1/2]$ (Proposition 3 and Appendix E.1).
(3) Reachability and Relabeling Growth Measures	
$\mathrm{Rea}_{\mathcal{G}}(x_{1:m})$, $\mathrm{Rel}_{\mathcal{H}}(z)$	Reachable source-label patterns on $x_{1:m} \in \mathcal{X}_{\mathrm{T}}^m$ and realizable target-label patterns on $z \in [K_{\mathrm{S}}]^m$ (Definition 2).
$T(z)$, $\lvert z \rvert_{\#}$, $t_{\mathcal{R}}(m)$	Distinct source labels $\{z_1, \ldots, z_m\}$, alphabet size $\lvert T(z) \rvert$, and worst-case reachable alphabet size (Example 1).
$\Pi_{\mathcal{R}}(m)$, $\Pi_{\mathcal{H}\mid\mathcal{R}}(m)$, $\Pi^{\mathrm{all}}_{\mathcal{H}}(m)$	Sample-wise reachability growth $\sup_{x_{1:m}} \lvert \mathrm{Rea}_{\mathcal{G}}(x_{1:m}) \rvert$, conditional-relabeling growth $\sup_{x_{1:m}} \sup_{z \in \mathrm{Rea}_{\mathcal{G}}(x_{1:m})} \lvert \mathrm{Rel}_{\mathcal{H}}(z) \rvert$, and structure-oblivious proxy $\sup_{z \in [K_{\mathrm{S}}]^m} \lvert \mathrm{Rel}_{\mathcal{H}}(z) \rvert$ (Definition 3).
$\mathrm{Ndim}(\mathcal{H})$, $d_{\mathcal{H}}$, $d^{\mathcal{R}}_{\mathcal{H}}(m)$	Multiclass Natarajan dimension of $\mathcal{H}$, its global upper bound $d_{\mathcal{H}}$ (Condition B), and its reachable-alphabet restriction $\sup_{x_{1:m}} \sup_{z \in \mathrm{Rea}_{\mathcal{G}}(x_{1:m})} \mathrm{Ndim}(\mathcal{H}\vert_{T(z)})$.
$c_{\mathcal{R}}$, $C_{\mathcal{R}}$; $\Phi_p(N)$	Cell-count exponent and scale constants in $(C_{\mathcal{R}} m)^{c_{\mathcal{R}} p}$ (Condition A), and weak affine sign-pattern bound $\sum_{j=0}^{\min\{p,N\}} 2^j \binom{N}{j}$ (Lemma 3).
(4) Risks, Residuals, and Certificates	
$R_{\mathrm{T}}(f)$, $R_{\mathrm{T}}^{\mathcal{F}_{\mathrm{MR}},*}$, R^*_{T}	Population target 0–1 risk $\mathbb{P}_{P_{\mathrm{T}}}(f(X) \neq Y)$, in-class optimum $\inf_{f \in \mathcal{F}_{\mathrm{MR}}} R_{\mathrm{T}}(f)$, and unrestricted target Bayes risk $\inf_f R_{\mathrm{T}}(f)$.
$\mathcal{E}^{\mathrm{est}}_{\mathrm{T}}(\widehat{f}_n)$, a_{MR}	Target estimation error $R_{\mathrm{T}}(\widehat{f}_n) - R_{\mathrm{T}}^{\mathcal{F}_{\mathrm{MR}},*}$ relative to $\mathcal{F}_{\mathrm{MR}}$ and population approximation error $R_{\mathrm{T}}^{\mathcal{F}_{\mathrm{MR}},*} - R^*_{\mathrm{T}}$ (Definition 1).
Z_g, U_g; $R^*_{\mathcal{H}}(g)$, $R^*_{\mathrm{all}}(g)$	Induced source label $f_{\mathrm{S}}(g(X))$ and logit vector $s_{\mathrm{S}}(g(X))$; $\mathcal{H}$-restricted and unrestricted 0–1 channel optima given Z_g (Proposition 4).
$R^\star_{\mathrm{T}}(U_g)$, $L^\star_\ell(U_g)$	Unrestricted 0–1 and surrogate channel Bayes risks given U_g (Propositions 5 and 6).
$\mathcal{L}_{\mathrm{T},\ell}(g, W)$, $\mathcal{L}^*_{\mathrm{T},\ell}$; $a_{\mathrm{score},0/1}$, $a_{\mathrm{score},\ell}$	Population surrogate target risk $\mathbb{E}_{P_{\mathrm{T}}}[\ell(Y, W s_{\mathrm{S}}(g(X)))]$, unrestricted surrogate Bayes risk, and logit-wrapper 0–1 and surrogate approximation errors (Appendix B.1).
$\mathrm{mis}_{\mathrm{hard}}$, ϵ_{hard}, $\mathrm{disc}_{\mathrm{hard}}$, τ_{hard}	Hard-label target Bayes mismatch $\mathbb{P}_{P_{\mathrm{T},X}}(b_{\mathrm{T}}(X) \neq h(b_{\mathrm{S}}(g(X))))$, source relabeled error $\mathbb{P}_{P_{\mathrm{S},X}}(h(f_{\mathrm{S}}(U)) \neq h(b_{\mathrm{S}}(U)))$, pushforward discrepancy $\lvert \mathbb{E}_{g_\# P_{\mathrm{T},X}}[\phi_h] - \mathbb{E}_{P_{\mathrm{S},X}}[\phi_h] \rvert$ with $\phi_h(u) := \mathbf{1}\{h(f_{\mathrm{S}}(u)) \neq h(b_{\mathrm{S}}(u))\}$, and population certificate $\tau_{\mathrm{hard}} := \inf_{g,h}\{\mathrm{mis}_{\mathrm{hard}} + \epsilon_{\mathrm{hard}} + \mathrm{disc}_{\mathrm{hard}}\}$ (Theorem 4).
$\mathrm{mis}_{\mathrm{score}}$, ζ_W, $\epsilon_{S,\ell}$, disc_ℓ, τ_{score}, $\widehat{\tau_{\mathrm{score}}}$	Score-output plug-in mismatch, pointwise surrogate distortion, source distortion, pushforward discrepancy, population certificate $\tau_{\mathrm{score}} := \mathrm{mis}_{\mathrm{score}} + \epsilon_{S,\ell} + \mathrm{disc}_\ell$, and empirical plug-in certificate $\widehat{\tau_{\mathrm{score}}}$ (Appendices B.3 and B.4).

$\widehat{\mathfrak{R}}_{Z_{1:n}}(\mathcal{F}), \Gamma_Q; R_{U1}, B_\lambda, B_\zeta$	Empirical Rademacher complexity on $Z_{1:n}$ and uniform concentration slack on candidate set Q; score-ball radius $B_h \max\{B_s\sqrt{K_S}, 1\}$ and loss/distortion envelopes (Corollary 6 and Theorem 6).

Minimal requirements for exact identities versus certificates. The population approximation identities require no boundedness, Lipschitz, concentration, or calibration conditions. Specifically, Proposition 4 uses only measurability of f_S and $g \in \mathcal{G}$, finite label alphabets, and nonemptiness of $\mathcal{G}$ and $\mathcal{H} \subseteq [K_T]^{[K_S]}$. Proposition 5 adds standard Borel target spaces to guarantee regular conditional posteriors $\eta_y^g(u) = \mathbb{P}_{P_T}(Y = y \mid U_g = u)$ under lexargmax tie-breaking, and Proposition 6 assumes measurable s_S, measurable $g \in \mathcal{G}$, nonempty $\mathcal{G}$ and $\mathcal{W}$, measurable $\ell : [K_T] \times \mathbb{R}^{K_T} \to [0, \infty)$, and finite risks $\mathcal{L}^*_{T,\ell} < \infty$ and $\inf_{W \in \mathcal{W}} \mathcal{L}_{T,\ell}(g, W) < \infty$ (with the pointwise conditional Bayes representation in Remark 9 invoking standard measurable-selection conditions). By contrast, the score-output certificates (Theorems 5 and 6 and Corollary 6) invoke the norm, Lipschitz, and evaluation-independence assumptions indexed in A8–A11 below and Appendix I.1.

Table 3: Theorem-indexed assumptions (A1–A17) used across the main text and appendices.

Label	Assumption	Used in
A1	Deterministic hard-label wrapper $f_{g,h} = h \circ f_S \circ g$ with measurable $f_S : \mathcal{X}_S \to [K_S]$ and $g \in \mathcal{G}$ (with $\mathcal{F}_{MR}$ pointwise measurable as above when uncountable).	Theorems 7 to 10 and Corollary 7
A2	Bounded 0–1 classification loss $\mathbf{1}\{f(x) \neq y\} \in \{0, 1\}$.	Theorems 7 to 10 and Corollary 7
A3	i.i.d. target sample $D_T = \{(X_i, Y_i)\}_{i=1}^{n_T} \sim P_T^{n_T}$.	Theorems 7 to 9 and Corollary 7
A4	Measurable η_n-approximate ERM $\widehat{f}_n \in \mathcal{F}_{MR}$ with optimization slack $\eta_n \geq 0$.	Theorems 7 and 8 and Corollary 7
A5	Measurable Bayes-optimal selectors $b_T : \mathcal{X}_T \to [K_T]$ and $b_S : \mathcal{X}_S \to [K_S]$.	Theorems 4 and 10
A6	Nonempty discrete output-mapping class $\mathcal{H} \subseteq [K_T]^{[K_S]}$.	Proposition 4 and Theorems 4 and 10
A7	Finite-support binary construction defining $\mathfrak{F}^{\mathrm{bin}}_{d_\mathcal{R}, d_\mathcal{H}}(\gamma)$ (Appendix E.1).	Proposition 3 and Theorems 3 and 9
A8	Bounded source score map $s_S : \mathcal{X}_S \to [-B_s, B_s]^{K_S}$ with $B_s \in [0, \infty)$.	Theorems 5, 11 and 12 and Corollary 11
A9	Norm-bounded linear score-output mappings $\|W\|_F \leq B_h$ ($B_h \in [0, \infty)$) for all $W \in \mathcal{W}$.	Theorems 5, 11 and 12 and Corollary 11
A10	Measurable surrogate $\ell \geq 0$, L_ℓ-Lipschitz on $B(0, R_{U1})$, bounded by $B_\lambda < \infty$, and admitting multiclass regret transform ψ_ℓ^{-1} when calibrated.	Theorems 5, 11 and 12 and Corollary 11
A11	Score-induced source classifier $\hat{b}_S(u) = \mathrm{lexargmax}_{k \in [K_S]} s_{S,k}(u)$ under deterministic lexicographic tie-breaking.	Theorems 5, 11 and 12 and Corollary 11
A12	Additive input-transformation class $\mathcal{G}_p = \{g_\theta(x) = x + M\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$ with fixed $M \in \mathbb{R}^{d \times p}$.	Theorems 2 and 8 and Corollaries 2, 8 and 9
A13	Reachability arrangement partition of Θ into at most $(C_\mathcal{R} m)^{c_\mathcal{R} p}$ constant-source-label cells on every $x_{1:m} \in \mathcal{X}_T^m$ (Condition A).	Theorems 2 and 8 and Corollary 2
A14	Output-mapping Natarajan dimension bound $\mathrm{Ndim}(\mathcal{H}) \leq d_\mathcal{H}$ (Condition B).	Theorems 2 and 8 and Corollary 2
A15	Affine source-score subclass $s_S(u) = Au + b$ with $f_S(u) = \mathrm{lexargmax}_{k \in [K_S]} s_{S,k}(u)$.	Corollaries 3 and 8
A16	Lower-bound sample-size and margin scaling $d_\mathcal{R} + d_\mathcal{H} \leq c_0 n_T$ and $\gamma = c_1\sqrt{(d_\mathcal{R} + d_\mathcal{H})/n_T}$ (with capped margin $\gamma = \min\{1/4, c_1\sqrt{(d_\mathcal{R} + d_\mathcal{H})/n_T}\}$ in Corollary 5).	Theorems 3 and 9 and Corollary 5
A17	Fixed finite-depth feedforward ReLU source-score network with hidden widths $q_1, \ldots, q_L$ and $f_S(u) = \mathrm{lexargmax}_{k \in [K_S]} s_{S,k}(u)$.	Corollaries 4 and 9

Additional evaluation-independence conditions for empirical certificates. Corollary 6 (Corollary 11) requires the finite candidate library Q of size M to be fixed independently of the certification samples $(D_{\mathrm{T}}, D_{\mathrm{S}})$, or selected on an independent holdout or cross-fitting fold prior to evaluation. Theorem 6 (Theorem 12) requires the candidate set $Q \subseteq \mathcal{G} \times \mathcal{W}$, the mutually independent samples $(D_{\mathrm{T}}, D_{\mathrm{S}})$, and the induced scalar function classes $(\mathcal{L}_Q, \mathcal{Z}_{\mathrm{T},Q}, \mathcal{Z}_{\mathrm{S},Q})$ to satisfy the measurability and independence conditions detailed in Appendix I.1.

B Approximation Analysis on Logit-Exposing Source Models

Section 6 characterizes the population approximation error a_{MR} when the frozen source model exposes discrete labels in $[K_{\mathrm{S}}]$. We now extend that approximation analysis to source models exposing real-valued logit (or score) vectors in $\mathbb{R}^{K_{\mathrm{S}}}$, where a learnable linear output mapping replaces discrete relabeling[2].

Appendix B.1 defines the logit-exposing wrapper class together with its 0–1 and surrogate approximation errors. Appendix B.2 establishes exact population decompositions at both interfaces (Propositions 5 and 6, proved in Appendix F), paralleling Proposition 4. Appendix B.3 derives a pointwise population certificate from observable plug-in quantities (Theorem 5, proved in Appendix H), paralleling Theorem 4. Appendix B.4 converts this population certificate into two finite-sample evaluation guarantees via an empirical plug-in certificate: a simultaneous Hoeffding bound over an evaluation-independent finite library (Corollary 6, proved in Appendix H) and a uniform empirical Rademacher bound over a fixed candidate set (Theorem 6, proved in Appendix I).

B.1 Logit-exposing wrapper architecture

Replacing the discrete source classifier $f_{\mathrm{S}} : \mathcal{X}_{\mathrm{S}} \to [K_{\mathrm{S}}]$ of Section 2.1 with a frozen measurable score map $s_{\mathrm{S}} : \mathcal{X}_{\mathrm{S}} \to \mathbb{R}^{K_{\mathrm{S}}}$ exposes a continuous logit vector at each transformed target input. A learnable linear output mapping $W \in \mathcal{W} \subseteq \mathbb{R}^{K_{\mathrm{T}} \times K_{\mathrm{S}}}$ projects $s_{\mathrm{S}}(g(x))$ to a target score vector $W s_{\mathrm{S}}(g(x)) \in \mathbb{R}^{K_{\mathrm{T}}}$, and the induced classifier selects a target label via the deterministic lexicographic rule $\mathrm{lexargmax}_{j \in [K]} t_j := \min \arg\max_{j \in [K]} t_j$ (Section 4.3 and Appendix A).

Definition 4 (Source-out MR predictor)**.** Given measurable input spaces $\mathcal{X}_{\mathrm{T}}$ and $\mathcal{X}_{\mathrm{S}}$ and finite label alphabets $[K_{\mathrm{T}}]$ and $[K_{\mathrm{S}}]$, for any fixed measurable score map $s_{\mathrm{S}} : \mathcal{X}_{\mathrm{S}} \to \mathbb{R}^{K_{\mathrm{S}}}$, any class $\mathcal{G}$ of measurable input transformations $g : \mathcal{X}_{\mathrm{T}} \to \mathcal{X}_{\mathrm{S}}$, and any linear output-mapping class $\mathcal{W} \subseteq \mathbb{R}^{K_{\mathrm{T}} \times K_{\mathrm{S}}}$, the *source-out MR predictor* associated with $(g, W) \in \mathcal{G} \times \mathcal{W}$ is

$$f_{g,W}(x) \;:=\; \mathrm{lexargmax}_{y \in [K_{\mathrm{T}}]} \big(W\, s_{\mathrm{S}}(g(x))\big)_y, \qquad \text{for every } x \in \mathcal{X}_{\mathrm{T}},$$

and the induced logit-exposing wrapper hypothesis class is

$$\mathcal{F}_{\mathrm{score}} \;:=\; \big\{\, f_{g,W} : g \in \mathcal{G},\; W \in \mathcal{W} \,\big\} \;\subseteq\; [K_{\mathrm{T}}]^{\mathcal{X}_{\mathrm{T}}}.$$

For nonempty $\mathcal{G}$ and $\mathcal{W}$ and a target distribution P_{T} over $\mathcal{X}_{\mathrm{T}} \times [K_{\mathrm{T}}]$ with 0–1 risk $R_{\mathrm{T}}(f) = \mathbb{P}_{(X,Y)\sim P_{\mathrm{T}}}(f(X) \neq Y)$ and unrestricted Bayes risk $R^*_{\mathrm{T}} = \inf_f R_{\mathrm{T}}(f)$ over all measurable $f : \mathcal{X}_{\mathrm{T}} \to [K_{\mathrm{T}}]$, the 0–1 approximation error of $\mathcal{F}_{\mathrm{score}}$ is

$$a_{\mathrm{score},0/1} \;:=\; \inf_{(g,W)\in\mathcal{G}\times\mathcal{W}} R_{\mathrm{T}}(f_{g,W}) - R^*_{\mathrm{T}}.$$

Because practical optimization replaces the discontinuous 0–1 loss with a measurable surrogate loss $\ell : [K_{\mathrm{T}}] \times \mathbb{R}^{K_{\mathrm{T}}} \to [0,\infty)$, we also compare the surrogate target risk $\mathcal{L}_{\mathrm{T},\ell}(g,W) := \mathbb{E}_{(X,Y)\sim P_{\mathrm{T}}}\big[\ell\big(Y, W s_{\mathrm{S}}(g(X))\big)\big]$ against the unrestricted surrogate Bayes risk $\mathcal{L}^*_{\mathrm{T},\ell} := \inf_q \mathbb{E}_{(X,Y)\sim P_{\mathrm{T}}}[\ell(Y, q(X))]$ over all measurable $q : \mathcal{X}_{\mathrm{T}} \to \mathbb{R}^{K_{\mathrm{T}}}$, defining the surrogate approximation error

$$a_{\mathrm{score},\ell} \;:=\; \inf_{(g,W)\in\mathcal{G}\times\mathcal{W}} \mathcal{L}_{\mathrm{T},\ell}(g,W) - \mathcal{L}^*_{\mathrm{T},\ell}.$$

When ℓ is multiclass classification-calibrated in the sense of Zhang [56] and Tewari and Bartlett [57] (including multiclass logistic loss and centered Lee–Lin–Wahba hinge losses [58]), a nondecreasing regret transform $\psi_\ell^{-1} : [0,\infty) \to [0,\infty)$ converts any upper bound on the surrogate excess risk $\mathcal{L}_{\mathrm{T},\ell}(g,W) - \mathcal{L}^*_{\mathrm{T},\ell}$ into an upper bound on the 0–1 excess risk $R_{\mathrm{T}}(f_{g,W}) - R^*_{\mathrm{T}}$.

B.2 Approximation identities

To test whether exposing continuous logits alters the coupled two-stage structure of Proposition 4, fix $g \in \mathcal{G}$ and isolate the lowest target 0–1 risk achievable from the induced logit vector $U_g :=$

[2]Unlike the population approximation identities and held-out evaluation certificates established here, the discrete product factorization of Proposition 1 (and hence the agnostic-PAC estimation bound of Theorem 1) relies on the finite alphabet $[K_{\mathrm{S}}]^m$ and does not transfer automatically. Section 8 outlines the metric-complexity route required for same-sample ERM estimation over continuous logits.

$s_{\mathrm{S}}(g(X)) \in \mathbb{R}^{K_{\mathrm{S}}}$ on $(X,Y) \sim P_{\mathrm{T}}$ before restricting the head to $\mathcal{W}$. For measurable $a : \mathbb{R}^{K_{\mathrm{S}}} \to [K_{\mathrm{T}}]$, we abbreviate $R_{\mathrm{T}}(a(U_g)) := R_{\mathrm{T}}(a \circ s_{\mathrm{S}} \circ g) = \mathbb{P}_{(X,Y)\sim P_{\mathrm{T}}}(a(U_g) \neq Y)$. On standard Borel spaces admitting regular conditional posteriors $\eta_y^g(u) := \mathbb{P}_{P_{\mathrm{T}}}(Y = y \mid U_g = u)$ for $y \in [K_{\mathrm{T}}]$, the unrestricted logit-conditioned 0–1 Bayes risk under g is

$$R_{\mathrm{T}}^\star(U_g) \;:=\; \inf_{a:\mathbb{R}^{K_{\mathrm{S}}}\to[K_{\mathrm{T}}] \text{ measurable}} R_{\mathrm{T}}\big(a(U_g)\big) \;=\; \mathbb{E}_{(X,Y)\sim P_{\mathrm{T}}}\Big[1 - \max_{y\in[K_{\mathrm{T}}]} \eta_y^g(U_g)\Big],$$

which is attained by the measurable selector $b_g^\star(u) := \mathrm{lexargmax}_{y\in[K_{\mathrm{T}}]}\, \eta_y^g(u)$. Telescoping each fixed-g excess risk around $R_{\mathrm{T}}^\star(U_g)$ inside the nested infimum over (g, W) (Appendix F.2 in Appendix F) yields the exact 0–1 decomposition below.

Proposition 5 (Score-output 0–1 approximation identity)**.** *Suppose $\mathcal{X}_{\mathrm{T}} \times [K_{\mathrm{T}}]$ is a standard Borel space. For every target distribution P_{T} over $\mathcal{X}_{\mathrm{T}} \times [K_{\mathrm{T}}]$, every fixed measurable score map $s_{\mathrm{S}} : \mathcal{X}_{\mathrm{S}} \to \mathbb{R}^{K_{\mathrm{S}}}$, every nonempty class $\mathcal{G}$ of measurable maps $\mathcal{X}_{\mathrm{T}} \to \mathcal{X}_{\mathrm{S}}$, and every nonempty linear output-mapping class $\mathcal{W} \subseteq \mathbb{R}^{K_{\mathrm{T}}\times K_{\mathrm{S}}}$, let $U_g := s_{\mathrm{S}}(g(X))$ for $(X,Y) \sim P_{\mathrm{T}}$ and $a_W(u) := \mathrm{lexargmax}_{y\in[K_{\mathrm{T}}]}(Wu)_y$ for $u \in \mathbb{R}^{K_{\mathrm{S}}}$. Then*

$$a_{\mathrm{score},0/1} \;=\; \inf_{g\in\mathcal{G}}\Big\{ \underbrace{R_{\mathrm{T}}^\star(U_g) - R_{\mathrm{T}}^*}_{\text{logit-channel information loss}} \;+\; \inf_{W\in\mathcal{W}} \underbrace{\big(R_{\mathrm{T}}(a_W(U_g)) - R_{\mathrm{T}}^\star(U_g)\big)}_{\text{linear-mapping restriction}} \Big\},$$

where both underbraced summands are nonnegative for every $g \in \mathcal{G}$.

Proposition 5 parallels Proposition 4 term for term at the continuous interface: $R_{\mathrm{T}}^\star(U_g) - R_{\mathrm{T}}^*$ measures target Bayes information lost when X is compressed into U_g (vanishing if and only if $R_{\mathrm{T}}^\star(U_g) = R_{\mathrm{T}}^*$, for instance when $s_{\mathrm{S}} \circ g$ is Borel-injective), whereas $\inf_{W\in\mathcal{W}}\big(R_{\mathrm{T}}(a_W(U_g)) - R_{\mathrm{T}}^\star(U_g)\big)$ measures the 0–1 excess risk of restricting the head to linear score projections a_W rather than arbitrary measurable classifiers on $\mathbb{R}^{K_{\mathrm{S}}}$ (vanishing whenever some $W \in \mathcal{W}$ satisfies $a_W(U_g) = b_g^\star(U_g)$ almost surely). As in Proposition 4, both summands depend on the same g inside the outer infimum.

To isolate the corresponding bottlenecks under a measurable surrogate loss ℓ, define the unrestricted logit-conditioned surrogate Bayes risk at fixed $g \in \mathcal{G}$ by

$$L_\ell^\star(U_g) \;:=\; \inf_{q:\mathbb{R}^{K_{\mathrm{S}}}\to\mathbb{R}^{K_{\mathrm{T}}} \text{ measurable}} \mathbb{E}_{(X,Y)\sim P_{\mathrm{T}}}\big[\ell\big(Y, q(U_g)\big)\big].$$

Adding and subtracting the finite intermediate optimum $L_\ell^\star(U_g)$ at each $g \in \mathcal{G}$ (Appendix F.3 and Remark 9 in Appendix F) gives the exact surrogate decomposition.

Proposition 6 (Score-output surrogate approximation identity)**.** *For every target distribution P_{T} over $\mathcal{X}_{\mathrm{T}} \times [K_{\mathrm{T}}]$, every fixed measurable score map $s_{\mathrm{S}} : \mathcal{X}_{\mathrm{S}} \to \mathbb{R}^{K_{\mathrm{S}}}$, every nonempty class $\mathcal{G}$ of measurable maps $\mathcal{X}_{\mathrm{T}} \to \mathcal{X}_{\mathrm{S}}$, every nonempty linear output-mapping class $\mathcal{W} \subseteq \mathbb{R}^{K_{\mathrm{T}}\times K_{\mathrm{S}}}$, and every measurable surrogate loss $\ell : [K_{\mathrm{T}}] \times \mathbb{R}^{K_{\mathrm{T}}} \to [0,\infty)$ satisfying $\mathcal{L}_{\mathrm{T},\ell}^* < \infty$ and $\inf_{W\in\mathcal{W}} \mathcal{L}_{\mathrm{T},\ell}(g,W) < \infty$ for every $g \in \mathcal{G}$, let $U_g := s_{\mathrm{S}}(g(X))$ for $(X,Y) \sim P_{\mathrm{T}}$. Then*

$$a_{\mathrm{score},\ell} \;=\; \inf_{g\in\mathcal{G}}\Big\{ \underbrace{L_\ell^\star(U_g) - \mathcal{L}_{\mathrm{T},\ell}^*}_{\text{surrogate information loss}} \;+\; \inf_{W\in\mathcal{W}} \underbrace{\big(\mathcal{L}_{\mathrm{T},\ell}(g,W) - L_\ell^\star(U_g)\big)}_{\text{surrogate linear-mapping restriction}} \Big\},$$

where both underbraced summands are nonnegative for every $g \in \mathcal{G}$.

Propositions 5 and 6 require no score boundedness or Lipschitz assumptions, and when $\mathcal{W}$ contains an optimal measurable head $u \mapsto Wu$ achieving $L_\ell^\star(U_g)$, surrogate linear-mapping restriction vanishes. Evaluating $R_{\mathrm{T}}^\star(U_g)$, R_{T}^*, $L_\ell^\star(U_g)$, and $\mathcal{L}_{\mathrm{T},\ell}^*$ directly, however, requires unknown conditional posteriors and unrestricted Bayes regressors, motivating an observable upper bound on the surrogate excess risk $\mathcal{L}_{\mathrm{T},\ell}(g,W) - \mathcal{L}_{\mathrm{T},\ell}^*$.

B.3 Source-conditional surrogate certificate

Whereas the hard-label certificate of Theorem 4 interpolates the unknown source Bayes selector b_{S}, a logit-exposing source model directly furnishes its own deterministic score-induced classifier $\hat{b}_{\mathrm{S}}(u) := \mathrm{lexargmax}_{k\in[K_{\mathrm{S}}]}\, s_{\mathrm{S},k}(u)$ and one-hot plug-in vector $c(u) := e_{\hat{b}_{\mathrm{S}}(u)} \in \mathbb{R}^{K_{\mathrm{S}}}$. For every measurable pair $(g,W) \in \mathcal{G} \times \mathcal{W}$, quantizing s_{S} to c on P_{T} defines the target plug-in surrogate

mismatch $\mathrm{mis}_{\mathrm{score}}(g, W)$, while comparing full logits $Ws_{\mathrm{S}}(u)$ against $Wc(u)$ defines the pointwise surrogate distortion $\zeta_W(u)$, its source-marginal expectation $\epsilon_{S,\ell}(W)$, and its pushforward discrepancy $\mathrm{disc}_\ell(g, W)$:

$$\begin{aligned}
\mathrm{mis}_{\mathrm{score}}(g, W) &:= \mathbb{E}_{(X,Y)\sim P_{\mathrm{T}}}\big[\ell\big(Y,\, Wc(g(X))\big)\big], \\
\zeta_W(u) &:= \max_{y\in[K_{\mathrm{T}}]} \big|\ell\big(y, Ws_{\mathrm{S}}(u)\big) - \ell\big(y, Wc(u)\big)\big|, \\
\epsilon_{S,\ell}(W) &:= \mathbb{E}_{U\sim P_{\mathrm{S},X}}\big[\zeta_W(U)\big], \\
\mathrm{disc}_\ell(g, W) &:= \big|\mathbb{E}_{X\sim P_{\mathrm{T},X}}\big[\zeta_W(g(X))\big] \;-\; \mathbb{E}_{U\sim P_{\mathrm{S},X}}\big[\zeta_W(U)\big]\big|.
\end{aligned}$$

Summing these three nonnegative terms defines the population certificate $\tau_{\mathrm{score}}(g, W) := \mathrm{mis}_{\mathrm{score}}(g, W) + \epsilon_{S,\ell}(W) + \mathrm{disc}_\ell(g, W)$. Interpolating $\ell(Y, Wc(g(X)))$, transferring ζ_W from $P_{\mathrm{S},X}$ to $g_\# P_{\mathrm{T},X}$, and using $\mathcal{L}^*_{\mathrm{T},\ell} \geq 0$ (Theorem 11 in Appendix H) yields the pointwise bound below.

Theorem 5 (Source-conditional surrogate certificate)**.** *For every target distribution P_{T} over $\mathcal{X}_{\mathrm{T}} \times [K_{\mathrm{T}}]$, every source-input marginal $P_{\mathrm{S},X}$ on $\mathcal{X}_{\mathrm{S}}$, and constants $B_s, B_h, L_\ell \in [0, \infty)$, suppose $s_{\mathrm{S}} : \mathcal{X}_{\mathrm{S}} \to [-B_s, B_s]^{K_{\mathrm{S}}}$ is measurable, every $W \in \mathcal{W}$ satisfies $\|W\|_F \leq B_h$, and the measurable surrogate loss $\ell : [K_{\mathrm{T}}] \times \mathbb{R}^{K_{\mathrm{T}}} \to [0, \infty)$ is L_ℓ-Lipschitz in its second argument with respect to $\|\cdot\|_2$ on the closed Euclidean ball $B(0, R_{U1}) \subseteq \mathbb{R}^{K_{\mathrm{T}}}$ of radius $R_{U1} := B_h \max\{B_s\sqrt{K_{\mathrm{S}}}, 1\}$ (A8–A11 in Table 3). Then for every measurable pair $(g, W) \in \mathcal{G} \times \mathcal{W}$,*

$$\mathcal{L}_{\mathrm{T},\ell}(g, W) - \mathcal{L}^*_{\mathrm{T},\ell} \;\leq\; \tau_{\mathrm{score}}(g, W) \;=\; \mathrm{mis}_{\mathrm{score}}(g, W) + \epsilon_{S,\ell}(W) + \mathrm{disc}_\ell(g, W).$$

Consequently, whenever ℓ is multiclass classification-calibrated with nondecreasing regret transform $\psi_\ell^{-1} : [0, \infty) \to [0, \infty)$,

$$R_{\mathrm{T}}(f_{g,W}) - R^*_{\mathrm{T}} \;\leq\; \psi_\ell^{-1}\big(\tau_{\mathrm{score}}(g, W)\big) \;=\; \psi_\ell^{-1}\big(\mathrm{mis}_{\mathrm{score}}(g, W) + \epsilon_{S,\ell}(W) + \mathrm{disc}_\ell(g, W)\big).$$

Theorem 5 upper-bounds the surrogate excess risk $\mathcal{L}_{\mathrm{T},\ell}(g, W) - \mathcal{L}^*_{\mathrm{T},\ell}$ and the 0–1 excess risk $R_{\mathrm{T}}(f_{g,W}) - R^*_{\mathrm{T}}$ via the population certificate $\tau_{\mathrm{score}}(g, W)$: $\mathrm{mis}_{\mathrm{score}}(g, W)$ pays for target surrogate loss after quantizing s_{S} to its one-hot plug-in c, $\epsilon_{S,\ell}(W)$ pays for the deviation between full logits and c under W on $P_{\mathrm{S},X}$, and $\mathrm{disc}_\ell(g, W)$ pays for the shift of that distortion between $g_\# P_{\mathrm{T},X}$ and $P_{\mathrm{S},X}$. When $s_{\mathrm{S}}(u) = c(u)$ is already one-hot across $\mathcal{X}_{\mathrm{S}}$, $\zeta_W \equiv 0$ so $\epsilon_{S,\ell}(W) = \mathrm{disc}_\ell(g, W) = 0$, reducing $\tau_{\mathrm{score}}(g, W)$ to $\mathrm{mis}_{\mathrm{score}}(g, W)$; when $g_\# P_{\mathrm{T},X} = P_{\mathrm{S},X}$, $\mathrm{disc}_\ell(g, W) = 0$. Taking the infimum over nonempty $\mathcal{G} \times \mathcal{W}$ gives $a_{\mathrm{score},\ell} \leq \inf_{(g,W)\in\mathcal{G}\times\mathcal{W}} \tau_{\mathrm{score}}(g, W)$. Like Theorem 4, $\tau_{\mathrm{score}}(g, W)$ is a one-directional upper certificate because the proof lower-bounds $\mathcal{L}^*_{\mathrm{T},\ell}$ by 0 and evaluates ζ_W pointwise at W rather than via a symmetric supremum over $\mathcal{W}$.

B.4 Empirical certificates

Because $\mathrm{mis}_{\mathrm{score}}(g, W)$, $\epsilon_{S,\ell}(W)$, and $\mathrm{disc}_\ell(g, W)$ depend only on the known maps $(s_{\mathrm{S}}, c, \ell)$, they can be estimated directly from labeled target examples and unlabeled source inputs without auxiliary posterior estimation. We establish two finite-sample guarantees for the resulting empirical plug-in certificate under evaluation-independent candidate selection.

Finite-library empirical certificate. Let $Q = \{(g_r, W_r)\}_{r=1}^M \subseteq \mathcal{G} \times \mathcal{W}$ ($M \geq 1$) be a finite library of measurable candidate pairs fixed independently of the certification samples (for example by preregistration or selection on a disjoint training split). Given mutually independent i.i.d. certification samples $D_{\mathrm{T}} = \{(X_i, Y_i)\}_{i=1}^{n_{\mathrm{T}}} \sim P_{\mathrm{T}}^{n_{\mathrm{T}}}$ and $D_{\mathrm{S}} = \{U_j\}_{j=1}^{n_{\mathrm{S}}} \sim P_{\mathrm{S},X}^{n_{\mathrm{S}}}$ drawn independently of Q, define for each $(g, W) \in Q$ the empirical plug-in certificate $\widehat{\tau_{\mathrm{score}}}(g, W)$ and its constituent sample means by

$$\begin{aligned}
\widehat{\mathrm{mis}_{\mathrm{score}}}(g, W) &:= \frac{1}{n_{\mathrm{T}}} \sum_{i=1}^{n_{\mathrm{T}}} \ell\big(Y_i, Wc(g(X_i))\big), \\
\widehat{T}(g, W) &:= \frac{1}{n_{\mathrm{T}}} \sum_{i=1}^{n_{\mathrm{T}}} \zeta_W(g(X_i)), \quad \widehat{S}(W) := \frac{1}{n_{\mathrm{S}}} \sum_{j=1}^{n_{\mathrm{S}}} \zeta_W(U_j), \\
\widehat{\tau_{\mathrm{score}}}(g, W) &:= \widehat{\mathrm{mis}_{\mathrm{score}}}(g, W) + \widehat{S}(W) + \big|\widehat{T}(g, W) - \widehat{S}(W)\big|.
\end{aligned}$$

Decomposing $\tau_{\mathrm{score}}(g,W) - \widehat{\tau_{\mathrm{score}}}(g,W)$ deterministically and applying Hoeffding's inequality across the $3M$ scalar sample means (Corollary 11 in Appendix H) yields the simultaneous bound below.

Corollary 6 (Finite-library empirical certificate)**.** *Under the assumptions of Theorem 5, set* $B_\zeta := L_\ell B_h(B_s\sqrt{K_{\mathrm{S}}}+1)$ *and* $B_\lambda := \sup_{y\in[K_{\mathrm{T}}],\, q\in B(0,R_{U1})} \ell(y,q) < \infty$*. For every integer* $M \geq 1$*, every finite candidate library* $Q = \{(g_r,W_r)\}_{r=1}^{M} \subseteq \mathcal{G}\times\mathcal{W}$ *of measurable pairs, all sample sizes* $n_{\mathrm{T}}, n_{\mathrm{S}} \geq 1$*, and every* $\delta \in (0,1)$*, if* $D_{\mathrm{T}} \sim P_{\mathrm{T}}^{n_{\mathrm{T}}}$ *and* $D_{\mathrm{S}} \sim P_{\mathrm{S},X}^{n_{\mathrm{S}}}$ *are mutually independent i.i.d. samples drawn independently of* Q*, then with probability at least* $1-\delta$ *over* $(D_{\mathrm{T}}, D_{\mathrm{S}})$*, simultaneously for all* $(g,W)\in Q$*,*

$$\mathcal{L}_{\mathrm{T},\ell}(g,W) - \mathcal{L}^*_{\mathrm{T},\ell} \;\leq\; \widehat{\tau_{\mathrm{score}}}(g,W) + B_\lambda\sqrt{\frac{\log(6M/\delta)}{2n_{\mathrm{T}}}} + B_\zeta\sqrt{\frac{\log(6M/\delta)}{2n_{\mathrm{T}}}} + 2B_\zeta\sqrt{\frac{\log(6M/\delta)}{2n_{\mathrm{S}}}}.$$

Consequently, whenever ℓ *is multiclass classification-calibrated with nondecreasing regret transform* $\psi_\ell^{-1}:[0,\infty)\to[0,\infty)$*, on the same* $(1-\delta)$*-probability event, simultaneously for all* $(g,W)\in Q$*,*

$$R_{\mathrm{T}}(f_{g,W}) - R^*_{\mathrm{T}} \;\leq\; \psi_\ell^{-1}\left(\widehat{\tau_{\mathrm{score}}}(g,W) + (B_\lambda+B_\zeta)\sqrt{\frac{\log(6M/\delta)}{2n_{\mathrm{T}}}} + 2B_\zeta\sqrt{\frac{\log(6M/\delta)}{2n_{\mathrm{S}}}}\right).$$

In Corollary 6, the factor 2 on the source-sample concentration term arises because $\widehat{S}(W)$ enters both the source-distortion estimate and the empirical discrepancy $|\widehat{T}(g,W) - \widehat{S}(W)|$ (Eqn. (15) in Appendix H). At $M=1$, Corollary 6 certifies a single held-out wrapper learned on an independent training split, whereas selecting Q on $(D_{\mathrm{T}}, D_{\mathrm{S}})$ without sample splitting invalidates the independence hypothesis.

Uniform empirical certificate over a fixed candidate set. When the candidate set $Q \subseteq \mathcal{G}\times\mathcal{W}$ is large or infinite, the $\sqrt{\log(6M/\delta)}$ union-bound term in Corollary 6 becomes vacuous. To replace finite-library counting with empirical capacity control while keeping Q independent of $(D_{\mathrm{T}}, D_{\mathrm{S}})$, write $\widehat{\mathfrak{R}}_{Z_{1:n}}(\mathcal{F}) := n^{-1}\,\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^n \sigma_i f(Z_i) \;\middle|\; Z_{1:n}\right]$ for the empirical Rademacher complexity of a scalar class $\mathcal{F}$ on $Z_{1:n}$ with independent uniform signs $\sigma_1,\dots,\sigma_n\in\{-1,+1\}$, and define the three induced scalar function classes

$$\begin{aligned}
\mathcal{L}_Q &:= \big\{(x,y)\mapsto \ell\big(y, Wc(g(x))\big) : (g,W)\in Q\big\},\\
\mathcal{Z}_{\mathrm{T},Q} &:= \big\{x\mapsto \zeta_W(g(x)) : (g,W)\in Q\big\}, \qquad \mathcal{Z}_{\mathrm{S},Q} := \big\{u\mapsto \zeta_W(u) : (g,W)\in Q\big\}.
\end{aligned}$$

Applying two-sided empirical Rademacher uniform-deviation bounds [54] across $\mathcal{L}_Q$ on D_{T}, $\mathcal{Z}_{\mathrm{T},Q}$ on the target-input tuple $X_{1:n_{\mathrm{T}}} := (X_1,\dots,X_{n_{\mathrm{T}}})$ of D_{T}, and $\mathcal{Z}_{\mathrm{S},Q}$ on $D_{\mathrm{S}} = \{U_j\}_{j=1}^{n_{\mathrm{S}}}$ (Theorem 12 in Appendix I) yields the uniform guarantee below.

Theorem 6 (Uniform empirical certificate over a fixed candidate set)**.** *Under the assumptions of Theorem 5, let B_λ, B_ζ be defined as in Corollary 6, and let $Q \subseteq \mathcal{G} \times \mathcal{W}$ be a nonempty set of measurable candidate pairs whose induced scalar classes $\mathcal{L}_Q$, $\mathcal{Z}_{\mathrm{T},Q}$, and $\mathcal{Z}_{\mathrm{S},Q}$ satisfy the measurability conditions of Appendix I.1. For all sample sizes $n_\mathrm{T}, n_\mathrm{S} \geq 1$ and every $\delta \in (0,1)$, if $D_\mathrm{T} = \{(X_i, Y_i)\}_{i=1}^{n_\mathrm{T}} \sim P_\mathrm{T}^{n_\mathrm{T}}$ (with input tuple $X_{1:n_\mathrm{T}}$) and $D_\mathrm{S} = \{U_j\}_{j=1}^{n_\mathrm{S}} \sim P_{\mathrm{S},X}^{n_\mathrm{S}}$ are mutually independent i.i.d. samples drawn independently of Q, then with probability at least $1-\delta$ over $(D_\mathrm{T}, D_\mathrm{S})$, simultaneously for all $(g, W) \in Q$,*

$$\mathcal{L}_{\mathrm{T},\ell}(g,W) - \mathcal{L}_{\mathrm{T},\ell}^* \;\leq\; \widehat{\tau_{\mathrm{score}}}(g,W) + \Gamma_Q,$$

where the uniform concentration slack Γ_Q is

$$\begin{aligned}\Gamma_Q \;:=\; & 2\widehat{\mathfrak{R}}_{D_\mathrm{T}}(\mathcal{L}_Q) + 3B_\lambda\sqrt{\frac{\log(12/\delta)}{2n_\mathrm{T}}} \\ & + 2\widehat{\mathfrak{R}}_{X_{1:n_\mathrm{T}}}(\mathcal{Z}_{\mathrm{T},Q}) + 3B_\zeta\sqrt{\frac{\log(12/\delta)}{2n_\mathrm{T}}} \\ & + 4\widehat{\mathfrak{R}}_{D_\mathrm{S}}(\mathcal{Z}_{\mathrm{S},Q}) + 6B_\zeta\sqrt{\frac{\log(12/\delta)}{2n_\mathrm{S}}}.\end{aligned}$$

Consequently, whenever ℓ is multiclass classification-calibrated with nondecreasing regret transform $\psi_\ell^{-1} : [0,\infty) \to [0,\infty)$, on the same $(1-\delta)$-probability event, simultaneously for all $(g, W) \in Q$,

$$R_\mathrm{T}(f_{g,W}) - R_\mathrm{T}^* \;\leq\; \psi_\ell^{-1}\big(\widehat{\tau_{\mathrm{score}}}(g,W) + \Gamma_Q\big).$$

Theorem 6 bounds the surrogate excess risk $\mathcal{L}_{\mathrm{T},\ell}(g,W) - \mathcal{L}_{\mathrm{T},\ell}^*$ and calibrated 0–1 excess risk $R_\mathrm{T}(f_{g,W}) - R_\mathrm{T}^*$ via the empirical plug-in certificate $\widehat{\tau_{\mathrm{score}}}(g,W)$ plus empirical Rademacher complexities of $\mathcal{L}_Q$, $\mathcal{Z}_{\mathrm{T},Q}$, and $\mathcal{Z}_{\mathrm{S},Q}$; when $|Q| = 1$, all three empirical Rademacher terms vanish identically, recovering a single-candidate concentration bound up to constants. Because Q is fixed independently of $(D_\mathrm{T}, D_\mathrm{S})$, Theorem 6 certifies held-out candidate sets rather than same-sample ERM over $\mathcal{G} \times \mathcal{W}$; extending to in-sample ERM over continuous logits requires bounding the complexity of the composed vector class $\{s_\mathrm{S} \circ g : g \in \mathcal{G}\}$ via vector contraction [55], as outlined in Section 8.

C Proof of the Factorized Excess-Risk Bound and Comparison Chain

Proof roadmap. This section establishes the finite-sample estimation guarantees of Section 3 together with the restricted head-growth lemma invoked later in Appendix D. We first prove an elementary fixed-point inversion lemma (Lemma 1) that converts logarithmic sample-size inequalities into explicit polynomial sample-complexity thresholds, and then factorize the sample-wise growth of $\mathcal{F}_{\mathrm{MR}}$ into reachability and conditional-relabeling components (Proposition 1). Inserting this product bound into a two-sided growth-function uniform-deviation inequality yields the distribution-free agnostic-PAC estimation bound (Theorem 7, restating Theorem 1), which combines with Lemma 1 to establish polynomial-growth sample complexity (Corollary 7, restating Corollary 1). Finally, we prove dominance of conditional relabeling over the structure-oblivious proxy (Proposition 2) and establish the reachable-alphabet Sauer–Shelah bound on $\Pi_{\mathcal{H}|\mathcal{R}}(m)$ (Lemma 2), whose coarse specialization (Remark 7) supplies the output-capacity bound in Theorem 8.

C.1 Auxiliary fixed-point inversion lemma

Lemma 1 (Logarithmic fixed-point inversion)**.** *Let $a \geq 1$, $b \geq 1$, and $c \geq 0$. For every real number n satisfying*

$$n \;\geq\; 4a\log(2ab) + 2c,$$

we have $n \geq a\log(bn) + c$.

Proof. Define $q(n) := n - a\log(bn) - c$ on $(0,\infty)$. For every $n \geq a$, the derivative satisfies $q'(n) = 1 - a/n \geq 0$, so q is nondecreasing on $[a,\infty)$. Set $L := \log(2ab)$ and $n_0 := 4aL + 2c$. Since $a \geq 1$ and $b \geq 1$, we have $2ab \geq 2$, which implies $L \geq \log 2 > 1/2$ and, using $4\log 2 > 1$ and $c \geq 0$, gives $n_0 \geq 4a\log 2 > a$. By monotonicity of q on $[a,\infty)$, it therefore suffices to show $q(n_0) \geq 0$, or equivalently $a\log(bn_0) + c \leq n_0$.

Expanding $\log(bn_0)$ at $n_0 = 4aL + 2c$ yields

$$\begin{aligned}
\log(bn_0) &= \log(4abL + 2bc) \\
&= \log(4abL) + \log\Big(1 + \frac{c}{2aL}\Big) \\
&= L + \log(2L) + \log\Big(1 + \frac{c}{2aL}\Big) && (\log(2ab) = L) \\
&\leq L + \log(2L) + \frac{c}{2aL}, && (\log(1+u) \leq u)
\end{aligned}$$

where the final step uses $c/(2aL) \geq 0$. Moreover, for every $L > 0$, truncating the exponential series at the quadratic term gives

$$\begin{aligned}
e^L &\geq 1 + L + \frac{L^2}{2} && (\text{since } L > 0) \\
&= 2L + \frac{1}{2}\big((L-1)^2 + 1\big) \\
&> 2L, && ((L-1)^2 + 1 \geq 1)
\end{aligned}$$

so taking logarithms gives $\log(2L) < L$. Combining this bound with $1/(2L) < 1$ (since $L > 1/2$), $a \geq 1$, and $c \geq 0$ yields

$$\begin{aligned}
a\log(bn_0) + c &\leq a\left(L + \log(2L) + \frac{c}{2aL}\right) + c \\
&\leq 2aL + \frac{c}{2L} + c && (\text{since } \log(2L) < L) \\
&\leq 2aL + 2c && (1/(2L) < 1,\ c \geq 0) \\
&\leq 4aL + 2c && (\text{since } aL > 0) \\
&= n_0.
\end{aligned}$$

Hence $q(n_0) \geq 0$, and monotonicity of q on $[a,\infty)$ implies $n \geq a\log(bn) + c$ for every $n \geq n_0$. □

Remark 4 (Role of the inversion constants). The numerical factors 4 and 2 in Lemma 1 are chosen for algebraic simplicity rather than tightness, converting the implicit logarithmic condition of Corollary 7 into an explicit $\widetilde{O}(\cdot)$ sample-complexity threshold.

C.2 Proof of Proposition 1

Proof. Fix an integer $m \geq 1$ and a target input tuple $x_{1:m} = (x_1, \ldots, x_m) \in \mathcal{X}_{\mathrm{T}}^m$. For each reachable source-label predictor $r = f_{\mathrm{S}} \circ g \in \mathcal{R}$, write $z_r := (r(x_1), \ldots, r(x_m)) \in [K_{\mathrm{S}}]^m$, so that $\mathrm{Rea}_{\mathcal{G}}(x_{1:m}) = \{z_r : r \in \mathcal{R}\}$ by Definition 2. Every wrapper predictor $f \in \mathcal{F}_{\mathrm{MR}}$ takes the form $f = h \circ r$ for some $r = f_{\mathrm{S}} \circ g \in \mathcal{R}$ and $h \in \mathcal{H}$, and evaluating f coordinatewise on $x_{1:m}$ places its target-label vector in $\mathrm{Rel}_{\mathcal{H}}(z_r)$:

$$(f(x_1), \ldots, f(x_m)) \;=\; \big(h(r(x_1)), \ldots, h(r(x_m))\big) \;\in\; \mathrm{Rel}_{\mathcal{H}}(z_r).$$

Thus $\{(f(x_1), \ldots, f(x_m)) : f \in \mathcal{F}_{\mathrm{MR}}\} \subseteq \bigcup_{z \in \mathrm{Rea}_{\mathcal{G}}(x_{1:m})} \mathrm{Rel}_{\mathcal{H}}(z)$, and counting realized target-label patterns gives

$$\begin{aligned}
\big|\{(f(x_1), \ldots, f(x_m)) : f \in \mathcal{F}_{\mathrm{MR}}\}\big| &\leq \Big| \bigcup_{z \in \mathrm{Rea}_{\mathcal{G}}(x_{1:m})} \mathrm{Rel}_{\mathcal{H}}(z) \Big| \\
&\leq \sum_{z \in \mathrm{Rea}_{\mathcal{G}}(x_{1:m})} |\mathrm{Rel}_{\mathcal{H}}(z)| \\
&\leq |\mathrm{Rea}_{\mathcal{G}}(x_{1:m})| \sup_{z \in \mathrm{Rea}_{\mathcal{G}}(x_{1:m})} |\mathrm{Rel}_{\mathcal{H}}(z)| \\
&\leq \Pi_{\mathcal{R}}(m)\, \Pi_{\mathcal{H}|\mathcal{R}}(m),
\end{aligned}$$

where the second inequality applies the finite union bound and the fourth invokes Definition 3. Taking the supremum over $x_{1:m} \in \mathcal{X}_{\mathrm{T}}^m$ yields $\Pi_{\mathcal{F}_{\mathrm{MR}}}(m) \leq \Pi_{\mathcal{R}}(m)\, \Pi_{\mathcal{H}|\mathcal{R}}(m)$. □

C.3 Proof of Theorem 1

Before proving Theorem 1, we record the formal definition of agnostic-PAC learnability referenced in Section 3.2. The theorem controls the class-relative estimation error $\mathcal{E}_{\mathrm{T}}^{\mathrm{est}}(\widehat{f}_n) = R_{\mathrm{T}}(\widehat{f}_n) - R_{\mathrm{T}}^{\mathcal{F}_{\mathrm{MR}},*}$, while the population approximation error $a_{\mathrm{MR}} = R_{\mathrm{T}}^{\mathcal{F}_{\mathrm{MR}},*} - R_{\mathrm{T}}^{*}$ enters only when converting estimation guarantees into target Bayes excess-risk bounds via Definition 1.

Definition 5 (Agnostic-PAC learnability [59])**.** A hypothesis class $\mathcal{F}_{\mathrm{MR}} \subseteq [K_{\mathrm{T}}]^{\mathcal{X}_{\mathrm{T}}}$ is *agnostically PAC learnable* by η_n-approximate ERM ($\eta_n \geq 0$) if there exists a sample-complexity function $n_0 : (0,1)^2 \to \mathbb{N}$ such that, for every target distribution P_{T} on $\mathcal{X}_{\mathrm{T}} \times [K_{\mathrm{T}}]$, every $\varepsilon, \delta \in (0,1)$, and every $n_{\mathrm{T}} \geq n_0(\varepsilon, \delta)$, with probability at least $1 - \delta$ over an i.i.d. target sample of size n_{T} drawn from P_{T}, every measurable η_n-approximate ERM $\widehat{f}_n \in \mathcal{F}_{\mathrm{MR}}$ satisfies $\mathcal{E}_{\mathrm{T}}^{\mathrm{est}}(\widehat{f}_n) \leq \eta_n + \varepsilon$.

Theorem 7 (Factorized agnostic-PAC bound; restatement of Theorem 1)**.** *Let $f_{\mathrm{S}} : \mathcal{X}_{\mathrm{S}} \to [K_{\mathrm{S}}]$ be a fixed measurable source classifier, $\mathcal{G}$ a nonempty class of measurable maps $\mathcal{X}_{\mathrm{T}} \to \mathcal{X}_{\mathrm{S}}$, and $\mathcal{H} \subseteq [K_{\mathrm{T}}]^{[K_{\mathrm{S}}]}$ a nonempty output-mapping class, inducing $\mathcal{F}_{\mathrm{MR}} = \{h \circ f_{\mathrm{S}} \circ g : g \in \mathcal{G},\, h \in \mathcal{H}\}$. For every target distribution P_{T} over $\mathcal{X}_{\mathrm{T}} \times [K_{\mathrm{T}}]$, every $n_{\mathrm{T}} \geq 1$, every $\delta \in (0,1)$, and every $\eta_n \geq 0$, with probability at least $1 - \delta$ over an i.i.d. sample of size n_{T} from P_{T}, every measurable η_n-approximate ERM $\widehat{f}_n \in \mathcal{F}_{\mathrm{MR}}$ satisfies*

$$\mathcal{E}_{\mathrm{T}}^{\mathrm{est}}(\widehat{f}_n) \;\leq\; \eta_n + 4\sqrt{6}\,\sqrt{\frac{\log \Pi_{\mathcal{R}}(2n_{\mathrm{T}}) + \log \Pi_{\mathcal{H}|\mathcal{R}}(2n_{\mathrm{T}}) + \log(1/\delta)}{n_{\mathrm{T}}}}.$$

Proof. Because $\mathcal{G}$ and $\mathcal{H}$ are nonempty and P_{T} is a probability distribution on $\mathcal{X}_{\mathrm{T}} \times [K_{\mathrm{T}}]$, both $\mathcal{F}_{\mathrm{MR}}$ and $\mathcal{X}_{\mathrm{T}}$ are nonempty. Define the induced binary 0–1 loss class on $\mathcal{X}_{\mathrm{T}} \times [K_{\mathrm{T}}]$ by

$$\mathcal{L}_{\mathrm{MR}} \;:=\; \big\{(x,y) \mapsto \mathbf{1}\{f(x) \neq y\} : f \in \mathcal{F}_{\mathrm{MR}}\big\} \;\subseteq\; \{0,1\}^{\mathcal{X}_{\mathrm{T}} \times [K_{\mathrm{T}}]}.$$

Fix any integer $m \geq 1$ and any labeled tuple $((x_1,y_1), \ldots, (x_m,y_m)) \in (\mathcal{X}_{\mathrm{T}} \times [K_{\mathrm{T}}])^m$. Once $(y_1, \ldots, y_m)$ is fixed, the map $(u_1, \ldots, u_m) \mapsto (\mathbf{1}\{u_1 \neq y_1\}, \ldots, \mathbf{1}\{u_m \neq y_m\})$ is deterministic, so any two predictors $f, f' \in \mathcal{F}_{\mathrm{MR}}$ that agree on $x_{1:m}$ induce identical loss vectors on $((x_i,y_i))_{i=1}^m$. Taking the supremum over labeled tuples of size m and applying Proposition 1 yields, for every $m \geq 1$,

$$\Pi_{\mathcal{L}_{\mathrm{MR}}}(m) \;\leq\; \Pi_{\mathcal{F}_{\mathrm{MR}}}(m) \;\leq\; \Pi_{\mathcal{R}}(m)\, \Pi_{\mathcal{H}|\mathcal{R}}(m). \tag{5}$$

Set $G_n := \Pi_{\mathcal{R}}(2n_{\mathrm{T}})\,\Pi_{\mathcal{H}|\mathcal{R}}(2n_{\mathrm{T}})$. Since $\mathcal{F}_{\mathrm{MR}} \neq \varnothing$ and $\mathcal{X}_{\mathrm{T}} \neq \varnothing$, Eqn. (5) gives $1 \le \Pi_{\mathcal{L}_{\mathrm{MR}}}(2n_{\mathrm{T}}) \le G_n$.

Applying the two-sided growth-function uniform-deviation inequality of Mohri et al. [24, Corollary 3.9 and Eq. (3.23)] to the $\{0,1\}$-valued class $\mathcal{L}_{\mathrm{MR}}$ on the i.i.d. sample $\{(X_i, Y_i)\}_{i=1}^{n_{\mathrm{T}}} \sim P_{\mathrm{T}}^{n_{\mathrm{T}}}$ (under the standard pointwise-measurability convention of Appendix A) and bounding $\Pi_{\mathcal{L}_{\mathrm{MR}}}(2n_{\mathrm{T}}) \le G_n$ via Eqn. (5) gives an event of probability at least $1-\delta$ on which

$$\sup_{f \in \mathcal{F}_{\mathrm{MR}}} \left| R_{\mathrm{T}}(f) - \widehat{R}_n(f) \right| \;\le\; \sqrt{\frac{8\{\log G_n + \log 4 + \log(1/\delta)\}}{n_{\mathrm{T}}}}, \tag{6}$$

where $\widehat{R}_n(f) = n_{\mathrm{T}}^{-1} \sum_{i=1}^{n_{\mathrm{T}}} \mathbf{1}\{f(X_i) \neq Y_i\}$. Fix an arbitrary $\nu > 0$ and choose a deterministic reference predictor $f_\nu \in \mathcal{F}_{\mathrm{MR}}$ satisfying $R_{\mathrm{T}}(f_\nu) \le R_{\mathrm{T}}^{\mathcal{F}_{\mathrm{MR}},*} + \nu$ by definition of $R_{\mathrm{T}}^{\mathcal{F}_{\mathrm{MR}},*} = \inf_{f \in \mathcal{F}_{\mathrm{MR}}} R_{\mathrm{T}}(f)$. Every measurable η_n-approximate ERM $\widehat{f}_n \in \mathcal{F}_{\mathrm{MR}}$ satisfies $\widehat{R}_n(\widehat{f}_n) \le \inf_{f \in \mathcal{F}_{\mathrm{MR}}} \widehat{R}_n(f) + \eta_n \le \widehat{R}_n(f_\nu) + \eta_n$. Telescoping $\mathcal{E}_{\mathrm{T}}^{\mathrm{est}}(\widehat{f}_n)$ into four differences gives

$$\begin{aligned}
\mathcal{E}_{\mathrm{T}}^{\mathrm{est}}(\widehat{f}_n) &= R_{\mathrm{T}}(\widehat{f}_n) - R_{\mathrm{T}}^{\mathcal{F}_{\mathrm{MR}},*} && \text{(by Definition 1)} \\
&= \big(R_{\mathrm{T}}(\widehat{f}_n) - \widehat{R}_n(\widehat{f}_n)\big) + \big(\widehat{R}_n(\widehat{f}_n) - \widehat{R}_n(f_\nu)\big) \\
&\quad + \big(\widehat{R}_n(f_\nu) - R_{\mathrm{T}}(f_\nu)\big) + \big(R_{\mathrm{T}}(f_\nu) - R_{\mathrm{T}}^{\mathcal{F}_{\mathrm{MR}},*}\big) \\
&\le 2 \sup_{f \in \mathcal{F}_{\mathrm{MR}}} \left| R_{\mathrm{T}}(f) - \widehat{R}_n(f) \right| \\
&\quad + \big(\widehat{R}_n(\widehat{f}_n) - \widehat{R}_n(f_\nu)\big) + \big(R_{\mathrm{T}}(f_\nu) - R_{\mathrm{T}}^{\mathcal{F}_{\mathrm{MR}},*}\big) && \text{(two-sided supremum)} \\
&\le 2 \sup_{f \in \mathcal{F}_{\mathrm{MR}}} \left| R_{\mathrm{T}}(f) - \widehat{R}_n(f) \right| + \eta_n + \nu. && (\eta_n\text{-ERM and } f_\nu)
\end{aligned}$$

Letting $\nu \downarrow 0$ and substituting Eqn. (6) with $2\sqrt{8} = 4\sqrt{2}$ yields, on the same $(1-\delta)$-probability event,

$$\mathcal{E}_{\mathrm{T}}^{\mathrm{est}}(\widehat{f}_n) \;\le\; \eta_n + 4\sqrt{2}\sqrt{\frac{\log G_n + \log 4 + \log(1/\delta)}{n_{\mathrm{T}}}}. \tag{7}$$

To absorb the additive $\log 4$ term, we split on the integer value of $G_n \ge 1$.

Case 1 ($G_n = 1$)**.** Monotonicity of sample-wise growth functions and Eqn. (5) yield $1 \le \Pi_{\mathcal{F}_{\mathrm{MR}}}(1) \le \Pi_{\mathcal{F}_{\mathrm{MR}}}(2n_{\mathrm{T}}) \le G_n = 1$, so $\Pi_{\mathcal{F}_{\mathrm{MR}}}(1) = 1$. Every predictor in $\mathcal{F}_{\mathrm{MR}}$ therefore takes a common value at each $x \in \mathcal{X}_{\mathrm{T}}$, forcing $R_{\mathrm{T}}(f) = R_{\mathrm{T}}^{\mathcal{F}_{\mathrm{MR}},*}$ for all $f \in \mathcal{F}_{\mathrm{MR}}$ and hence $\mathcal{E}_{\mathrm{T}}^{\mathrm{est}}(\widehat{f}_n) = 0$. Since $\eta_n \ge 0$, $\log G_n = \log \Pi_{\mathcal{R}}(2n_{\mathrm{T}}) + \log \Pi_{\mathcal{H}|\mathcal{R}}(2n_{\mathrm{T}}) = 0$, and $\log(1/\delta) > 0$, the claimed bound holds deterministically.

Case 2 ($G_n \ge 2$)**.** Because $\log 4 = 2\log 2 \le 2\log G_n$ and $\log(1/\delta) > 0$, we obtain

$$\begin{aligned}
\log G_n + \log 4 + \log(1/\delta) &\le 3\log G_n + \log(1/\delta) && (\text{since } G_n \ge 2) \\
&\le 3\{\log G_n + \log(1/\delta)\}. && (\text{since } \log(1/\delta) > 0)
\end{aligned}$$

Substituting this inequality into Eqn. (7), using $4\sqrt{2} \cdot \sqrt{3} = 4\sqrt{6}$, and expanding $\log G_n = \log \Pi_{\mathcal{R}}(2n_{\mathrm{T}}) + \log \Pi_{\mathcal{H}|\mathcal{R}}(2n_{\mathrm{T}})$ completes the proof. □

Remark 5 (Target Bayes excess-risk consequence). Combining Theorem 7 with the exact identity $R_{\mathrm{T}}(\widehat{f}_n) - R_{\mathrm{T}}^{*} = \mathcal{E}_{\mathrm{T}}^{\mathrm{est}}(\widehat{f}_n) + a_{\mathrm{MR}}$ (Definition 1) gives, on the same $(1-\delta)$-probability event,

$$R_{\mathrm{T}}(\widehat{f}_n) - R_{\mathrm{T}}^{*} \;\le\; a_{\mathrm{MR}} + \eta_n + 4\sqrt{6}\sqrt{\frac{\log \Pi_{\mathcal{R}}(2n_{\mathrm{T}}) + \log \Pi_{\mathcal{H}|\mathcal{R}}(2n_{\mathrm{T}}) + \log(1/\delta)}{n_{\mathrm{T}}}}.$$

C.4 Proof of Corollary 1

Corollary 7 (Polynomial-growth sample complexity; restatement of Corollary 1)**.** *Under the wrapper setup of Theorem 7, suppose there exist constants $d_{\mathcal{R}}, d_{\mathcal{H}} \ge 0$ and $C_{\mathcal{R}}, C_{\mathcal{H}} \ge 1$ such that*

$\log \Pi_{\mathcal{R}}(m) \le d_{\mathcal{R}} \log(C_{\mathcal{R}} m)$ *and* $\log \Pi_{\mathcal{H}|\mathcal{R}}(m) \le d_{\mathcal{H}} \log(C_{\mathcal{H}} m)$ *for every* $m \ge 1$. *Then for every target distribution* P_{T} *over* $\mathcal{X}_{\mathrm{T}} \times [K_{\mathrm{T}}]$, *every* $\varepsilon, \delta \in (0,1)$, *and every* $\eta_n \ge 0$, *there exists* $n_0 = \widetilde{O}\big((d_{\mathcal{R}} + d_{\mathcal{H}} + \log(1/\delta))/\varepsilon^2\big)$ *such that whenever* $n_{\mathrm{T}} \ge n_0$, *every measurable* η_n*-approximate ERM* $\widehat{f}_n \in \mathcal{F}_{\mathrm{MR}}$ *satisfies* $\mathcal{E}_{\mathrm{T}}^{\mathrm{est}}(\widehat{f}_n) \le \eta_n + \varepsilon$ *with probability at least* $1 - \delta$ *over* $P_{\mathrm{T}}^{n_{\mathrm{T}}}$.

Proof. Fix a target distribution P_{T} over $\mathcal{X}_{\mathrm{T}} \times [K_{\mathrm{T}}]$, $\varepsilon, \delta \in (0,1)$, and $\eta_n \ge 0$. By Theorem 7, it suffices to exhibit an integer $n_0 \ge 1$ such that the statistical term in Theorem 7 is at most ε for all $n_{\mathrm{T}} \ge n_0$. Evaluating the assumed polynomial growth bounds at $m = 2n_{\mathrm{T}}$ gives

$$\log \Pi_{\mathcal{R}}(2n_{\mathrm{T}}) + \log \Pi_{\mathcal{H}|\mathcal{R}}(2n_{\mathrm{T}}) + \log(1/\delta) \;\le\; d_{\mathcal{R}} \log(2C_{\mathcal{R}} n_{\mathrm{T}}) + d_{\mathcal{H}} \log(2C_{\mathcal{H}} n_{\mathrm{T}}) + \log(1/\delta).$$

Since $(4\sqrt{6})^2 = 96$, the statistical term in Theorem 7 is at most ε whenever

$$n_{\mathrm{T}} \;\ge\; \frac{96}{\varepsilon^2}\Big\{ d_{\mathcal{R}} \log(2C_{\mathcal{R}} n_{\mathrm{T}}) + d_{\mathcal{H}} \log(2C_{\mathcal{H}} n_{\mathrm{T}}) + \log(1/\delta) \Big\}. \tag{8}$$

Set $d := d_{\mathcal{R}} + d_{\mathcal{H}} \ge 0$ and $C_* := \max\{C_{\mathcal{R}}, C_{\mathcal{H}}\} \ge 1$. If $d = 0$, then $d_{\mathcal{R}} = d_{\mathcal{H}} = 0$, and Eqn. (8) reduces directly to $n_{\mathrm{T}} \ge 96 \log(1/\delta)/\varepsilon^2$, which holds for all $n_{\mathrm{T}} \ge n_0 := \lceil 96 \log(1/\delta)/\varepsilon^2 \rceil = O\big(\log(1/\delta)/\varepsilon^2\big)$.

Assume henceforth that $d > 0$, and define

$$a \;:=\; \max\Big\{1,\, \frac{96d}{\varepsilon^2}\Big\} \;\ge\; 1, \qquad b \;:=\; 2C_* \;\ge\; 2, \qquad c \;:=\; \frac{96 \log(1/\delta)}{\varepsilon^2} \;>\; 0.$$

For every $n_{\mathrm{T}} \ge 1$, since $bn_{\mathrm{T}} \ge 2$ implies $\log(bn_{\mathrm{T}}) > 0$, the right-hand side of Eqn. (8) satisfies

$$\begin{aligned}
&\frac{96}{\varepsilon^2}\Big\{ d_{\mathcal{R}} \log(2C_{\mathcal{R}} n_{\mathrm{T}}) + d_{\mathcal{H}} \log(2C_{\mathcal{H}} n_{\mathrm{T}}) + \log(1/\delta) \Big\} \\
&\quad\le \frac{96}{\varepsilon^2}\big\{ d \log(2C_* n_{\mathrm{T}}) + \log(1/\delta) \big\} \quad (\text{since } C_{\mathcal{R}}, C_{\mathcal{H}} \le C_*) \\
&\quad= \frac{96d}{\varepsilon^2} \log(bn_{\mathrm{T}}) + c \\
&\quad\le a \log(bn_{\mathrm{T}}) + c. \quad (\text{since } a \ge 96d/\varepsilon^2 \text{ and } \log(bn_{\mathrm{T}}) > 0)
\end{aligned}$$

Consequently, every integer $n_{\mathrm{T}} \ge 1$ satisfying $n_{\mathrm{T}} \ge a \log(bn_{\mathrm{T}}) + c$ also satisfies Eqn. (8). Invoking Lemma 1 shows that $n_0 := \lceil 4a \log(2ab) + 2c \rceil$ suffices, with $d_{\mathcal{R}}$ and $d_{\mathcal{H}}$ entering linearly while $C_{\mathcal{R}}, C_{\mathcal{H}}$ appear only inside $\log(2ab)$, giving $n_0 = \widetilde{O}\big((d_{\mathcal{R}} + d_{\mathcal{H}} + \log(1/\delta))/\varepsilon^2\big)$. Moreover, the explicit formula $n_0 = \lceil 4a \log(2ab) + 2c \rceil$ (and $\lceil 96 \log(1/\delta)/\varepsilon^2 \rceil$ when $d = 0$) depends only on $(d_{\mathcal{R}}, d_{\mathcal{H}}, C_{\mathcal{R}}, C_{\mathcal{H}}, \varepsilon, \delta)$ and in particular not on P_{T} or η_n, establishing agnostic-PAC learnability under Definition 5. □

Remark 6 (Bayes-excess form under polynomial growth). Combining Corollary 7 with Definition 1 yields $R_{\mathrm{T}}(\widehat{f}_n) - R_{\mathrm{T}}^* \le a_{\mathrm{MR}} + \eta_n + \varepsilon$ with probability at least $1 - \delta$ whenever $n_{\mathrm{T}} \ge n_0$, which is the form invoked when pairing the estimation rate with a population approximation certificate.

C.5 Proof of Proposition 2 and restricted head-growth lemma

Proof of Proposition 2. Fix an integer $m \ge 1$. For every target tuple $x_{1:m} \in \mathcal{X}_{\mathrm{T}}^m$, Definition 2 gives $\mathrm{Rea}_{\mathcal{G}}(x_{1:m}) \subseteq [K_{\mathrm{S}}]^m$, so maximizing $|\mathrm{Rel}_{\mathcal{H}}(z)|$ over reachable source-label patterns cannot exceed maximizing over $[K_{\mathrm{S}}]^m$:

$$\sup_{z \in \mathrm{Rea}_{\mathcal{G}}(x_{1:m})} |\mathrm{Rel}_{\mathcal{H}}(z)| \;\le\; \sup_{z \in [K_{\mathrm{S}}]^m} |\mathrm{Rel}_{\mathcal{H}}(z)| \;=\; \Pi_{\mathcal{H}}^{\mathrm{all}}(m).$$

Taking the supremum over $x_{1:m} \in \mathcal{X}_{\mathrm{T}}^m$ gives $\Pi_{\mathcal{H}|\mathcal{R}}(m) \le \Pi_{\mathcal{H}}^{\mathrm{all}}(m)$. Combining Proposition 1 with $\Pi_{\mathcal{R}}(m) \ge 0$ yields

$$\Pi_{\mathcal{F}_{\mathrm{MR}}}(m) \;\le\; \Pi_{\mathcal{R}}(m)\, \Pi_{\mathcal{H}|\mathcal{R}}(m) \;\le\; \Pi_{\mathcal{R}}(m)\, \Pi_{\mathcal{H}}^{\mathrm{all}}(m).$$ □

We conclude this section with the reachable-alphabet Sauer–Shelah bound on $\Pi_{\mathcal{H}|\mathcal{R}}(m)$, which restricts output-side counting to the set of source labels realized on a target sample and supplies the head-capacity bound invoked in Appendix D. Recall from Appendix A that for $z = (z_1, \dots, z_m) \in [K_{\mathrm{S}}]^m$, $T(z) = \{z_1, \dots, z_m\} \subseteq [K_{\mathrm{S}}]$ denotes the reached source-label set of cardinality $|z|_{\#} = |T(z)|$, with sample-wise maxima $t_{\mathcal{R}}(m) = \sup_{x_{1:m} \in \mathcal{X}_{\mathrm{T}}^m} \sup_{z \in \mathrm{Rea}_{\mathcal{G}}(x_{1:m})} |T(z)|$ and $d_{\mathcal{H}}^{\mathcal{R}}(m) = \sup_{x_{1:m} \in \mathcal{X}_{\mathrm{T}}^m} \sup_{z \in \mathrm{Rea}_{\mathcal{G}}(x_{1:m})} \mathrm{Ndim}(\mathcal{H}|_{T(z)})$.

Lemma 2 (Restricted head growth on the reachable source alphabet)**.** *Under the wrapper setup of Theorem 7, for every integer $m \geq 1$, the conditional-relabeling growth satisfies*

$$\Pi_{\mathcal{H}|\mathcal{R}}(m) \;\leq\; \sum_{i=0}^{d_{\mathcal{H}}^{\mathcal{R}}(m) \wedge t_{\mathcal{R}}(m)} \binom{t_{\mathcal{R}}(m)}{i} \binom{K_{\mathrm{T}}}{2}^{i}, \tag{9}$$

with the convention $\binom{K_{\mathrm{T}}}{2}^{0} := 1$ when $K_{\mathrm{T}} = 1$.

Proof. Fix an integer $m \geq 1$. If $K_{\mathrm{T}} = 1$, then every $h \in \mathcal{H} \subseteq \{1\}^{[K_{\mathrm{S}}]}$ is identically 1, so $|\mathrm{Rel}_{\mathcal{H}}(z)| \leq 1$ on every $z \in \mathrm{Rea}_{\mathcal{G}}(x_{1:m})$ and hence $\Pi_{\mathcal{H}|\mathcal{R}}(m) \leq 1$. Moreover, no coordinate admits two distinct target labels when $K_{\mathrm{T}} = 1$, giving $d_{\mathcal{H}}^{\mathcal{R}}(m) = 0$ and $\binom{t_{\mathcal{R}}(m)}{0}\binom{1}{2}^{0} = 1$ under $\binom{K_{\mathrm{T}}}{2}^{0} := 1$, which establishes Eqn. (9).

Assume henceforth that $K_{\mathrm{T}} \geq 2$. Fix any target tuple $x_{1:m} \in \mathcal{X}_{\mathrm{T}}^m$ and any reachable source-label vector $z = (z_1, \dots, z_m) \in \mathrm{Rea}_{\mathcal{G}}(x_{1:m})$, and set $A_z := T(z) = \{z_1, \dots, z_m\} \subseteq [K_{\mathrm{S}}]$ and $d_z := \mathrm{Ndim}(\mathcal{H}|_{A_z})$. Two maps $h, h' \in \mathcal{H}$ satisfy $(h(z_1), \dots, h(z_m)) = (h'(z_1), \dots, h'(z_m))$ if and only if $h(k) = h'(k)$ for every $k \in A_z$, because A_z consists of the exact coordinates appearing in z. Thus the evaluation map $h|_{A_z} \mapsto (h(z_1), \dots, h(z_m))$ is a bijection from $\mathcal{H}|_{A_z} := \{h|_{A_z} : h \in \mathcal{H}\}$ onto $\mathrm{Rel}_{\mathcal{H}}(z)$, giving $|\mathrm{Rel}_{\mathcal{H}}(z)| = |\mathcal{H}|_{A_z}|$. Applying the multivalued Sauer–Shelah lemma [32, Theorem 9] to $\mathcal{H}|_{A_z} \subseteq [K_{\mathrm{T}}]^{A_z}$ on the finite domain A_z, together with $|A_z| \leq t_{\mathcal{R}}(m)$ and $d_z \leq d_{\mathcal{H}}^{\mathcal{R}}(m)$, yields

$$\begin{aligned}
|\mathrm{Rel}_{\mathcal{H}}(z)| &= |\mathcal{H}|_{A_z}| && \text{(restriction bijection)} \\
&\leq \sum_{i=0}^{d_z \wedge |A_z|} \binom{|A_z|}{i} \binom{K_{\mathrm{T}}}{2}^{i} && \text{(multivalued Sauer–Shelah)} \\
&\leq \sum_{i=0}^{d_z \wedge |A_z|} \binom{t_{\mathcal{R}}(m)}{i} \binom{K_{\mathrm{T}}}{2}^{i} && \text{(since } |A_z| \leq t_{\mathcal{R}}(m)\text{)} \\
&\leq \sum_{i=0}^{d_{\mathcal{H}}^{\mathcal{R}}(m) \wedge t_{\mathcal{R}}(m)} \binom{t_{\mathcal{R}}(m)}{i} \binom{K_{\mathrm{T}}}{2}^{i}, && \text{(since } d_z \leq d_{\mathcal{H}}^{\mathcal{R}}(m)\text{)}
\end{aligned}$$

where the last inequality extends a sum of nonnegative terms from $d_z \wedge |A_z|$ to $d_{\mathcal{H}}^{\mathcal{R}}(m) \wedge t_{\mathcal{R}}(m)$. Taking the supremum over $z \in \mathrm{Rea}_{\mathcal{G}}(x_{1:m})$ and then over $x_{1:m} \in \mathcal{X}_{\mathrm{T}}^m$ proves Eqn. (9). □

Remark 7 (Recovery of the coarse Natarajan head bound). Because every $z \in [K_{\mathrm{S}}]^m$ contains at most m distinct symbols and domain restriction cannot increase Natarajan dimension, we always have $t_{\mathcal{R}}(m) \leq m$ and $d_{\mathcal{H}}^{\mathcal{R}}(m) \leq \mathrm{Ndim}(\mathcal{H})$. Whenever $\mathrm{Ndim}(\mathcal{H}) \leq d_{\mathcal{H}}$, integrality of $\mathrm{Ndim}(\mathcal{H}) \in \mathbb{N}_0$ gives $d_{\mathcal{H}}^{\mathcal{R}}(m) \leq \lfloor d_{\mathcal{H}} \rfloor$, and Lemma 2 therefore implies

$$\Pi_{\mathcal{H}|\mathcal{R}}(m) \;\leq\; \sum_{i=0}^{\lfloor d_{\mathcal{H}} \rfloor \wedge m} \binom{m}{i} \binom{K_{\mathrm{T}}}{2}^{i},$$

which is the source-free output-capacity bound used in Theorem 8.

D Proofs for the Additive Specialization and Representation-Side Discharges

Proof roadmap. This section proves the explicit PAC rate and sample complexity for additive input transformations (Theorem 2 and Corollary 2) and discharges the reachability-arrangement condition (Condition A) for affine and fixed finite-depth ReLU source-score maps (Corollaries 3 and 4). In Appendix D.1, Theorem 8 combines the cell-partition bound of Condition A on $\Pi_{\mathcal{R}}(m)$ with the restricted Sauer–Shelah bound of Lemma 2 (Remark 7) under Condition B on $\Pi_{\mathcal{H}|\mathcal{R}}(m)$ and substitutes both into Theorem 7, after which Corollary 7 yields Corollary 2. In Appendix D.2, we establish the affine hyperplane weak-sign-pattern recursion on $\mathbb{R}^p$ (Lemma 3), apply it with the lexicographic weak-sign characterization Eqn. (10) to the $m\binom{K_S}{2}$ pairwise score differences of an affine score map (Corollary 8, proving Corollary 3), and propagate the refinement inductively across L hidden ReLU layers before the output-layer score comparisons (Corollary 9, proving Corollary 4).

D.1 Proof of Theorem 2 and Corollary 2

Theorem 8 (Explicit PAC bound for additive MR; restatement of Theorem 2)**.** *For $\mathcal{X}_T \subseteq \mathbb{R}^d$, $\mathcal{X}_S = \mathbb{R}^d$, and the additive input-transformation class $\mathcal{G}_p = \{g_\theta(x) = x + M\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$ with fixed $M \in \mathbb{R}^{d\times p}$, assume Conditions A and B and set $d_m := d_{\mathcal{H}} \wedge m$ (with $B_m := 0$ when $d_m = 0$ or $K_T = 1$). For every integer $m \geq 1$,*

$$\log \Pi_{\mathcal{R}}(m) \;\leq\; \underbrace{c_{\mathcal{R}}\, p \log(C_{\mathcal{R}}\, m)}_{=:A_m}, \qquad \log \Pi_{\mathcal{H}|\mathcal{R}}(m) \;\leq\; \underbrace{d_m \log\left(\frac{e\binom{K_T}{2}m}{d_m}\right)}_{=:B_m}.$$

Consequently, for every target distribution P_T, every $n_T \geq 1$, every $\delta \in (0,1)$, and every $\eta_n \geq 0$, with probability at least $1-\delta$ over $P_T^{n_T}$, every measurable η_n-approximate ERM $\widehat{f}_n \in \mathcal{F}_{MR}$ satisfies

$$\mathcal{E}_T^{\mathrm{est}}(\widehat{f}_n) \;\leq\; \eta_n + 4\sqrt{6}\sqrt{\frac{A_{2n_T} + B_{2n_T} + \log(1/\delta)}{n_T}}.$$

Proof. Fix an integer $m \geq 1$.

Bounding $\log \Pi_{\mathcal{R}}(m)$ (A_m). Fix an arbitrary target tuple $x_{1:m} \in \mathcal{X}_T^m$. For each parameter $\theta \in \Theta$, write the induced source-label vector as

$$z_\theta(x_{1:m}) \;:=\; \big(f_S(g_\theta(x_1)), \ldots, f_S(g_\theta(x_m))\big) \;\in\; [K_S]^m,$$

so that $\mathrm{Rea}_{\mathcal{G}}(x_{1:m}) = \{z_\theta(x_{1:m}) : \theta \in \Theta\}$. By Condition A, Θ admits a partition into at most $(C_{\mathcal{R}}\, m)^{c_{\mathcal{R}}\, p}$ cells on each of which $\theta \mapsto z_\theta(x_{1:m})$ is constant. Because at most one source-label vector is realized per cell, $|\mathrm{Rea}_{\mathcal{G}}(x_{1:m})| \leq (C_{\mathcal{R}}\, m)^{c_{\mathcal{R}}\, p}$. Taking the supremum over $x_{1:m} \in \mathcal{X}_T^m$ gives $\Pi_{\mathcal{R}}(m) \leq (C_{\mathcal{R}}\, m)^{c_{\mathcal{R}}\, p}$, and taking logarithms yields $\log \Pi_{\mathcal{R}}(m) \leq c_{\mathcal{R}}\, p \log(C_{\mathcal{R}}\, m) = A_m$.

Bounding $\log \Pi_{\mathcal{H}|\mathcal{R}}(m)$ (B_m). Under Condition B, $\mathrm{Ndim}(\mathcal{H}) \leq d_{\mathcal{H}} < \infty$, and since $\mathrm{Ndim}(\mathcal{H}) \in \mathbb{N}_0$, we have $d_{\mathcal{H}} \geq 0$ and $d_{\mathcal{H}}^{\mathcal{R}}(m) \leq \mathrm{Ndim}(\mathcal{H}) \leq \lfloor d_{\mathcal{H}} \rfloor$, while $t_{\mathcal{R}}(m) \leq m$. If $K_T = 1$ or $d_m = 0$, then Lemma 2 (Remark 7) gives $\Pi_{\mathcal{H}|\mathcal{R}}(m) = 1$, so $\log \Pi_{\mathcal{H}|\mathcal{R}}(m) = 0 = B_m$. If $K_T \geq 2$ and $0 < d_m < 1$, then $m \geq 1$ forces $0 < d_{\mathcal{H}} < 1$ and hence $\mathrm{Ndim}(\mathcal{H}) \leq \lfloor d_{\mathcal{H}} \rfloor = 0$, so Lemma 2 gives $\Pi_{\mathcal{H}|\mathcal{R}}(m) = 1$ and $\log \Pi_{\mathcal{H}|\mathcal{R}}(m) = 0 < d_m \log\big(e\binom{K_T}{2}m/d_m\big) = B_m$ (since $e\binom{K_T}{2}m/d_m > e > 1$). Assume henceforth that $K_T \geq 2$ and $1 \leq d_m \leq m$, and set $k_m := \lfloor d_m \rfloor \in \{1, \ldots, m\}$. Then $d_{\mathcal{H}}^{\mathcal{R}}(m) \wedge m \leq \lfloor d_{\mathcal{H}} \rfloor \wedge m = k_m \leq d_m$, while $\binom{K_T}{2} \geq 1$ gives $\binom{K_T}{2}^i \leq \binom{K_T}{2}^{d_m}$ and $m \geq d_m$ gives $(m/d_m)^{d_m}(d_m/m)^i = (m/d_m)^{d_m - i} \geq 1$ for every $0 \leq i \leq k_m$. Starting from the restricted

Sauer–Shelah bound (Lemma 2 and Remark 7), we obtain

$$\begin{aligned}
\Pi_{\mathcal{H}|\mathcal{R}}(m) &\le \sum_{i=0}^{k_m} \binom{m}{i} \binom{K_{\mathrm{T}}}{2}^{i} \\
&\le \binom{K_{\mathrm{T}}}{2}^{d_m} \sum_{i=0}^{k_m} \binom{m}{i} && (\tbinom{K_{\mathrm{T}}}{2} \ge 1,\ i \le d_m) \\
&\le \binom{K_{\mathrm{T}}}{2}^{d_m} \left(\frac{m}{d_m}\right)^{d_m} \sum_{i=0}^{k_m} \binom{m}{i} \left(\frac{d_m}{m}\right)^{i} && ((m/d_m)^{d_m - i} \ge 1) \\
&\le \binom{K_{\mathrm{T}}}{2}^{d_m} \left(\frac{m}{d_m}\right)^{d_m} \left(1 + \frac{d_m}{m}\right)^{m} && (\text{extending sum to } m) \\
&\le \left(\frac{e\binom{K_{\mathrm{T}}}{2} m}{d_m}\right)^{d_m}, && (1 + u \le e^{u})
\end{aligned}$$

and taking logarithms yields $\log \Pi_{\mathcal{H}|\mathcal{R}}(m) \le d_m \log\big(e\binom{K_{\mathrm{T}}}{2} m / d_m\big) = B_m$.

Evaluating both growth bounds at $m = 2n_{\mathrm{T}}$ gives $\log \Pi_{\mathcal{R}}(2n_{\mathrm{T}}) \le A_{2n_{\mathrm{T}}}$ and $\log \Pi_{\mathcal{H}|\mathcal{R}}(2n_{\mathrm{T}}) \le B_{2n_{\mathrm{T}}}$. Since $u \mapsto \sqrt{u}$ is nondecreasing on $[0, \infty)$, substituting both inequalities into Theorem 7 yields the stated finite-sample bound on $\mathcal{E}^{\mathrm{est}}_{\mathrm{T}}(\widehat{f}_n)$. □

Proof of Corollary 2. Under Conditions A and B, Theorem 8 gives $\log \Pi_{\mathcal{R}}(m) \le A_m = (c_{\mathcal{R}}\, p) \log(C_{\mathcal{R}}\, m)$ for every integer $m \ge 1$, with $c_{\mathcal{R}}\, p \ge 0$ and $C_{\mathcal{R}} \ge 1$. Define $C_{\mathcal{H}} := \max\{1, e\binom{K_{\mathrm{T}}}{2}\} \ge 1$. If $K_{\mathrm{T}} = 1$ or $0 \le d_{\mathcal{H}} < 1$, then $\mathrm{Ndim}(\mathcal{H}) \in \mathbb{N}_0$ forces $\mathrm{Ndim}(\mathcal{H}) = 0$ whenever $0 \le d_{\mathcal{H}} < 1$, so Lemma 2 gives $\Pi_{\mathcal{H}|\mathcal{R}}(m) = 1$ and $\log \Pi_{\mathcal{H}|\mathcal{R}}(m) = 0 \le d_{\mathcal{H}} \log(C_{\mathcal{H}}\, m)$ for every $m \ge 1$ (since $d_{\mathcal{H}} \ge 0$ and $C_{\mathcal{H}}\, m \ge 1$). Otherwise, $d_{\mathcal{H}} \ge 1$ and $K_{\mathrm{T}} \ge 2$, so for every $m \ge 1$ we have $1 \le d_m \le d_{\mathcal{H}}$, $C_{\mathcal{H}} = e\binom{K_{\mathrm{T}}}{2}$, and $\log\big(e\binom{K_{\mathrm{T}}}{2} m\big) \ge 1 > 0$; Theorem 8 then yields

$$\begin{aligned}
\log \Pi_{\mathcal{H}|\mathcal{R}}(m) \le B_m \;&=\; d_m \log\left(\frac{e\binom{K_{\mathrm{T}}}{2} m}{d_m}\right) \\
&\le d_m \log\left(e\binom{K_{\mathrm{T}}}{2} m\right) && (d_m \ge 1) \\
&\le d_{\mathcal{H}} \log\left(e\binom{K_{\mathrm{T}}}{2} m\right) \;=\; d_{\mathcal{H}} \log(C_{\mathcal{H}}\, m). && (d_m \le d_{\mathcal{H}})
\end{aligned}$$

In both cases, the polynomial-growth hypotheses of Corollary 7 hold with exponents $(c_{\mathcal{R}}\, p, d_{\mathcal{H}})$ and scale constants $(C_{\mathcal{R}}, C_{\mathcal{H}})$. Invoking Corollary 7 gives $n_0 = \widetilde{O}\big((c_{\mathcal{R}}\, p + d_{\mathcal{H}} + \log(1/\delta))/\varepsilon^2\big)$ such that whenever $n_{\mathrm{T}} \ge n_0$, with probability at least $1 - \delta$ over $P_{\mathrm{T}}^{n_{\mathrm{T}}}$, every measurable η_n-approximate ERM $\widehat{f}_n \in \mathcal{F}_{\mathrm{MR}}$ satisfies $\mathcal{E}^{\mathrm{est}}_{\mathrm{T}}(\widehat{f}_n) \le \eta_n + \varepsilon$, and hence $R_{\mathrm{T}}(\widehat{f}_n) - R^*_{\mathrm{T}} = \mathcal{E}^{\mathrm{est}}_{\mathrm{T}}(\widehat{f}_n) + a_{\mathrm{MR}} \le a_{\mathrm{MR}} + \eta_n + \varepsilon$ (Definition 1). □

D.2 Weak affine sign-pattern bound and proofs of Corollaries 3 and 4

Lemma 3 (Weak affine sign-pattern count)**.** *For $p, N \in \mathbb{N}_0$, define*

$$\Phi_p(N) \;:=\; \sum_{j=0}^{\min\{p,N\}} 2^j \binom{N}{j}, \qquad \Phi_p(0) \;=\; 1.$$

Let $\ell_1, \dots, \ell_N : \mathbb{R}^p \to \mathbb{R}$ be affine functions, and define the three-valued weak sign $\mathrm{sgn} : \mathbb{R} \to \{-1, 0, +1\}$ *by*

$$\mathrm{sgn}(t) \;:=\; \begin{cases} +1, & t > 0, \\ 0, & t = 0, \\ -1, & t < 0. \end{cases}$$

Then for every subset $\Theta \subseteq \mathbb{R}^p$, the number of distinct weak sign vectors $\big(\operatorname{sgn}(\ell_1(\theta)), \dots, \operatorname{sgn}(\ell_N(\theta))\big)$ realized as θ ranges over Θ is at most $\Phi_p(N)$. Moreover, $\Phi_p(N) \le (1+2N)^p$ for every $p \ge 1$, while $\Phi_0(N) = 1$, so $\log \Phi_p(N) \le p \log(1+2N)$ for all $p, N \in \mathbb{N}_0$.

Proof. Because restricting θ from $\mathbb{R}^p$ to $\Theta \subseteq \mathbb{R}^p$ can only decrease the set of realized sign vectors, it suffices to take $\Theta = \mathbb{R}^p$. For each $\sigma \in \{-1, 0, +1\}^N$, define its weak sign cell

$$C_\sigma \;:=\; \big\{\theta \in \mathbb{R}^p : \operatorname{sgn}(\ell_j(\theta)) = \sigma_j \text{ for all } j \in [N]\big\}.$$

The cells $\{C_\sigma : \sigma \in \{-1,0,+1\}^N\}$ form a pairwise disjoint partition of $\mathbb{R}^p$, and each nonempty C_σ is convex as an intersection of open affine halfspaces ($\sigma_j = \pm 1$) and affine hyperplanes ($\sigma_j = 0$). For any N-tuple $(\ell_1, \dots, \ell_N)$ of affine functions on $\mathbb{R}^p$, let $C(\ell_1, \dots, \ell_N) := \big|\{\sigma \in \{-1,0,+1\}^N : C_\sigma \neq \varnothing\}\big|$ denote the number of nonempty weak sign cells it generates on $\mathbb{R}^p$, and define the worst-case cell count

$$F_p(N) \;:=\; \sup_{(\ell'_1, \dots, \ell'_N)} C(\ell'_1, \dots, \ell'_N),$$

where the supremum (attained as a maximum since $C(\ell'_1, \dots, \ell'_N) \le 3^N$) ranges over all N-tuples of affine functions on $\mathbb{R}^p$. In the boundary cases, $F_p(0) = 1 = \Phi_p(0)$ (the single empty tuple) and $F_0(N) = 1 = \Phi_0(N)$ (since $\mathbb{R}^0$ is a singleton).

Now assume $p, N \ge 1$, fix an arbitrary N-tuple $(\ell_1, \dots, \ell_N)$ of affine functions on $\mathbb{R}^p$, and consider adding ℓ_N to $(\ell_1, \dots, \ell_{N-1})$. If ℓ_N is constant on $\mathbb{R}^p$, its weak sign takes a single value everywhere, so each nonempty cell of $(\ell_1, \dots, \ell_{N-1})$ produces a single nonempty cell of $(\ell_1, \dots, \ell_N)$ and

$$C(\ell_1, \dots, \ell_N) \;=\; C(\ell_1, \dots, \ell_{N-1}) \;\le\; F_p(N-1) \;\le\; F_p(N-1) + 2F_{p-1}(N-1).$$

Otherwise, ℓ_N is nonconstant, and its zero set $H := \{\theta \in \mathbb{R}^p : \ell_N(\theta) = 0\}$ is an affine hyperplane isomorphic to $\mathbb{R}^{p-1}$. Each nonempty cell C_τ ($\tau \in \{-1,0,+1\}^{N-1}$) of $(\ell_1, \dots, \ell_{N-1})$ falls into one of three mutually exclusive cases relative to H:

1. *Non-intersecting ($C_\tau \cap H = \varnothing$).* Because C_τ is convex (hence connected) and ℓ_N is continuous and nonzero on C_τ, the Intermediate Value Theorem forces $\operatorname{sgn}(\ell_N(\theta))$ to take a single value ($+1$ or -1) throughout C_τ; thus C_τ contributes 1 nonempty cell after ℓ_N is added.

2. *Contained ($C_\tau \subseteq H$).* Here $\operatorname{sgn}(\ell_N(\theta)) = 0$ on C_τ, so C_τ again contributes 1 nonempty cell.

3. *Crossing ($C_\tau \cap H \neq \varnothing$ and $C_\tau \not\subseteq H$).* Splitting C_τ according to $\operatorname{sgn}(\ell_N(\theta)) \in \{-1,0,+1\}$ partitions C_τ into at most 3 nonempty subcells (at most one each for $\ell_N > 0$, $\ell_N = 0$, and $\ell_N < 0$), adding at most 2 nonempty cells beyond the 1 counted in $C(\ell_1, \dots, \ell_{N-1})$.

Let $\mathcal{T}_{\text{cross}}$ denote the set of indices τ in the crossing case. For every $\tau \in \mathcal{T}_{\text{cross}}$, the trace $C_\tau \cap H$ is nonempty and equals the weak sign cell $\{\theta \in H : \operatorname{sgn}(\ell_j|_H(\theta)) = \tau_j \text{ for all } j \in [N-1]\}$ of the $(N-1)$ restricted affine functions $\ell_1|_H, \dots, \ell_{N-1}|_H$ on H. Moreover, the map $\tau \mapsto C_\tau \cap H$ on $\mathcal{T}_{\text{cross}}$ is injective: if $\theta_0 \in (C_\tau \cap H) \cap (C_{\tau'} \cap H)$, then $\tau_j = \operatorname{sgn}(\ell_j(\theta_0)) = \tau'_j$ for every $j \in [N-1]$, so $\tau = \tau'$. Identifying H affinely with $\mathbb{R}^{p-1}$ bounds $|\mathcal{T}_{\text{cross}}| \le F_{p-1}(N-1)$, and summing over all nonempty cells of $(\ell_1, \dots, \ell_{N-1})$ gives

$$C(\ell_1, \dots, \ell_N) \;\le\; C(\ell_1, \dots, \ell_{N-1}) + 2|\mathcal{T}_{\text{cross}}| \;\le\; F_p(N-1) + 2F_{p-1}(N-1).$$

In both cases, $C(\ell_1, \dots, \ell_N) \le F_p(N-1) + 2F_{p-1}(N-1)$; taking the supremum over all N-tuples $(\ell_1, \dots, \ell_N)$ of affine functions on $\mathbb{R}^p$ yields the recursion

$$F_p(N) \;\le\; F_p(N-1) + 2F_{p-1}(N-1).$$

For all $p, N \geq 1$, Pascal's identity $\binom{N-1}{j} + \binom{N-1}{j-1} = \binom{N}{j}$ shows that Φ_p satisfies the matching recursion with equality:

$$\begin{aligned}\Phi_p(N-1) + 2\Phi_{p-1}(N-1) &= \sum_{j=0}^{\min\{p,N-1\}} 2^j \binom{N-1}{j} + \sum_{j=1}^{\min\{p,N\}} 2^j \binom{N-1}{j-1} \\ &= \sum_{j=0}^{\min\{p,N\}} 2^j \binom{N}{j} \;=\; \Phi_p(N),\end{aligned}$$

where when $N-1 < p$, the $j = N$ summand in the first sum vanishes via $\binom{N-1}{N} = 0$. Since $F_p(N) = \Phi_p(N) = 1$ on the boundary $p = 0$ or $N = 0$, induction on $p + N$ gives $C(\ell_1, \ldots, \ell_N) \leq F_p(N) \leq \Phi_p(N)$ for all $p, N \in \mathbb{N}_0$.

Finally, if $p = 0$, then $\Phi_0(N) = 1$ and $\log \Phi_0(N) = 0$. For $p \geq 1$, if $N = 0$ then $\Phi_p(0) = 1 = (1 + 2 \cdot 0)^p$; if $N \geq 1$, then for every $0 \leq j \leq \min\{p, N\}$,

$$\begin{aligned}\binom{N}{j} &= \frac{1}{j!} \prod_{r=0}^{j-1} (N - r) \\ &\leq \frac{N^j}{j!} && (0 < N - r \leq N) \\ &\leq N^j && (j! \geq 1) \\ &\leq \binom{p}{j} N^j, && (\tbinom{p}{j} \geq 1,\ j \leq p)\end{aligned}$$

with the empty product at $j = 0$ equal to 1. Summing over j gives

$$\begin{aligned}\Phi_p(N) &= \sum_{j=0}^{\min\{p,N\}} 2^j \binom{N}{j} \\ &\leq \sum_{j=0}^{\min\{p,N\}} \binom{p}{j} (2N)^j \\ &\leq \sum_{j=0}^{p} \binom{p}{j} (2N)^j && \text{(extending sum to } p\text{)} \\ &= (1 + 2N)^p, && \text{(binomial theorem)}\end{aligned}$$

and taking logarithms yields $\log \Phi_p(N) \leq p \log(1 + 2N)$. □

Corollary 8 (Affine source-score maps; restatement of Corollary 3)**.** *For $\mathcal{G}_p$, if $f_{\mathrm{S}}(u) = \operatorname{lexargmax}_{k \in [K_{\mathrm{S}}]} s_{\mathrm{S},k}(u)$ with $s_{\mathrm{S}}(u) = Au + b$ ($A \in \mathbb{R}^{K_{\mathrm{S}} \times d}$, $b \in \mathbb{R}^{K_{\mathrm{S}}}$), then for every $m \geq 1$,*

$$\log \Pi_{\mathcal{R}}(m) \;\leq\; \underbrace{\log \Phi_p\big(m \tbinom{K_{\mathrm{S}}}{2}\big)}_{=: \, C_{\mathrm{Aff}}(m)} \;\leq\; p \log\big(1 + 2m \tbinom{K_{\mathrm{S}}}{2}\big).$$

Proof. Under deterministic lexicographic tie-breaking $\operatorname{lexargmax}_{j \in [K_{\mathrm{S}}]} t_j = \min \arg\max_{j \in [K_{\mathrm{S}}]} t_j$, any score vector $t = (t_1, \ldots, t_{K_{\mathrm{S}}}) \in \mathbb{R}^{K_{\mathrm{S}}}$ ($K_{\mathrm{S}} \geq 1$) selects class $k \in [K_{\mathrm{S}}]$ if and only if t_k strictly exceeds every lower-indexed coordinate and weakly exceeds every higher-indexed coordinate:

$$\operatorname{lexargmax}_{j \in [K_{\mathrm{S}}]} t_j = k \quad \Longleftrightarrow \quad \begin{cases} \operatorname{sgn}(t_r - t_k) = -1 & \text{for all } 1 \leq r < k, \\ \operatorname{sgn}(t_k - t_\ell) \in \{0, +1\} & \text{for all } k < \ell \leq K_{\mathrm{S}}, \end{cases} \tag{10}$$

where both index ranges are vacuous when $K_{\mathrm{S}} = 1$.

Fix $m \geq 1$ and an arbitrary target tuple $x_{1:m} \in \mathcal{X}_{\mathrm{T}}^m$. If $K_{\mathrm{S}} = 1$, then $\binom{K_{\mathrm{S}}}{2} = 0$ and $f_{\mathrm{S}} \equiv 1$, so Θ itself forms a single constant-label cell, $|\mathrm{Rea}_{\mathcal{G}}(x_{1:m})| \leq 1 = \Phi_p(0)$, $\Pi_{\mathcal{R}}(m) \leq 1 = \Phi_p(0)$,

and $\log \Pi_{\mathcal{R}}(m) \le 0 = \mathcal{C}_{\mathrm{Aff}}(m) = p\log(1+0)$ for all $p \in \mathbb{N}_0$. Assume henceforth $K_{\mathrm{S}} \ge 2$, and denote the k-th row of A by $A_k^\top \in \mathbb{R}^{1\times d}$. For each sample index $i \in [m]$ and ordered label pair $1 \le k < \ell \le K_{\mathrm{S}}$, the pairwise score difference

$$\delta_{i,k\ell}(\theta) \;:=\; s_{\mathrm{S},k}\big(g_\theta(x_i)\big) - s_{\mathrm{S},\ell}\big(g_\theta(x_i)\big) \;=\; (A_k - A_\ell)^\top M\theta + (A_k - A_\ell)^\top x_i + (b_k - b_\ell)$$

is an affine function of $\theta \in \mathbb{R}^p$. Applying Lemma 3 to the $N := m\binom{K_{\mathrm{S}}}{2}$ affine functions $\{\delta_{i,k\ell} : i \in [m],\, 1 \le k < \ell \le K_{\mathrm{S}}\}$ partitions Θ into at most $\Phi_p\big(m\binom{K_{\mathrm{S}}}{2}\big)$ nonempty weak sign cells on each of which every $\mathrm{sgn}(\delta_{i,k\ell}(\theta))$ is constant. Applying Eqn. (10) to $t = s_{\mathrm{S}}(g_\theta(x_i))$ shows that the source-label vector $\big(f_{\mathrm{S}}(g_\theta(x_1)), \dots, f_{\mathrm{S}}(g_\theta(x_m))\big)$ is constant on each cell, discharging Condition A with $c_{\mathcal{R}} = 1$ and $C_{\mathcal{R}} = 1 + 2\binom{K_{\mathrm{S}}}{2}$ (since $1 + 2m\binom{K_{\mathrm{S}}}{2} \le \big(1 + 2\binom{K_{\mathrm{S}}}{2}\big)m$ for $m \ge 1$). Hence $|\mathrm{Rea}_{\mathcal{G}}(x_{1:m})| \le \Phi_p\big(m\binom{K_{\mathrm{S}}}{2}\big)$, and taking the supremum over $x_{1:m} \in \mathcal{X}_{\mathrm{T}}^m$ gives $\Pi_{\mathcal{R}}(m) \le \Phi_p\big(m\binom{K_{\mathrm{S}}}{2}\big)$; taking logarithms and applying $\log \Phi_p(N) \le p\log(1+2N)$ (Lemma 3) for all $p \in \mathbb{N}_0$ yields $\log \Pi_{\mathcal{R}}(m) \le \mathcal{C}_{\mathrm{Aff}}(m) \le p\log\big(1 + 2m\binom{K_{\mathrm{S}}}{2}\big)$. □

For the fixed-depth ReLU discharge (Corollary 4), consider a fixed feedforward ReLU source-score network $s_{\mathrm{S}} : \mathbb{R}^d \to \mathbb{R}^{K_{\mathrm{S}}}$ with $L \ge 1$ hidden layers of widths $q_1, \dots, q_L \ge 1$, defined recursively for $u \in \mathbb{R}^d$ by

$$z_0(u) \;:=\; u, \qquad a_\ell(u) \;:=\; A_\ell z_{\ell-1}(u) + b_\ell, \qquad z_\ell(u) \;:=\; \sigma\big(a_\ell(u)\big) \quad (\ell = 1, \dots, L),$$

where $\sigma(t) := \max\{t, 0\}$ acts coordinatewise, layer ℓ has fixed weights $A_\ell \in \mathbb{R}^{q_\ell \times q_{\ell-1}}$ ($q_0 := d$) and biases $b_\ell \in \mathbb{R}^{q_\ell}$, and the output score map is $s_{\mathrm{S}}(u) := A_{L+1} z_L(u) + b_{L+1}$ with $A_{L+1} \in \mathbb{R}^{K_{\mathrm{S}} \times q_L}$ and $b_{L+1} \in \mathbb{R}^{K_{\mathrm{S}}}$ ($K_{\mathrm{S}} \ge 1$).

Corollary 9 (Fixed finite-depth ReLU source-score networks; restatement of Corollary 4)**.** *For $\mathcal{G}_p$, if $f_{\mathrm{S}}(u) = \mathrm{lexargmax}_{k\in[K_{\mathrm{S}}]}\, s_{\mathrm{S},k}(u)$ where s_{S} is a fixed feedforward ReLU network with $L \ge 1$ hidden layers of widths $q_1, \dots, q_L$, then for every $m \ge 1$,*

$$\log \Pi_{\mathcal{R}}(m) \;\le\; \underbrace{\log \Phi_p\Big(m\binom{K_{\mathrm{S}}}{2}\Big) + \sum_{\ell=1}^{L} \log \Phi_p(m\, q_\ell)}_{=:\, \mathcal{C}_{\mathrm{ReLU}}(m)} \;=\; O\big(p(L+1)\log m\big).$$

Proof. Fix $m \ge 1$ and an arbitrary target tuple $x_{1:m} \in \mathcal{X}_{\mathrm{T}}^m$, and write $u_i(\theta) := g_\theta(x_i) = x_i + M\theta$ for $i \in [m]$. We construct finite partitions $\mathcal{P}_0, \mathcal{P}_1, \dots, \mathcal{P}_L$ of Θ by induction on $\ell \in \{0, \dots, L\}$ maintaining two invariants: (a) $|\mathcal{P}_\ell| \le \prod_{s=1}^{\ell} \Phi_p(m\, q_s)$ (with the empty product equal to 1), and (b) on every cell $C \in \mathcal{P}_\ell$ and for every $i \in [m]$, the hidden representation $\theta \mapsto z_\ell(u_i(\theta)) \in \mathbb{R}^{q_\ell}$ is the restriction to C of an affine map on $\mathbb{R}^p$.

Base case ($\ell = 0$). Set $\mathcal{P}_0 := \{\Theta\}$ (assuming $\Theta \ne \varnothing$; otherwise $\Pi_{\mathcal{R}}(m) = 0$ trivially). Then $|\mathcal{P}_0| = 1$, matching the empty product in (a), and $z_0(u_i(\theta)) = x_i + M\theta$ is globally affine on $\mathbb{R}^p$ for every $i \in [m]$, establishing (b).

Inductive step ($\ell - 1 \to \ell$). For $1 \le \ell \le L$, assume $\mathcal{P}_{\ell-1}$ satisfies (a) and (b), and fix a cell $C \in \mathcal{P}_{\ell-1}$. By invariant (b), for each $i \in [m]$ there exist $U_{\ell-1}^{(i)} \in \mathbb{R}^{q_{\ell-1}\times p}$ and $v_{\ell-1}^{(i)} \in \mathbb{R}^{q_{\ell-1}}$ such that $z_{\ell-1}(u_i(\theta)) = U_{\ell-1}^{(i)}\theta + v_{\ell-1}^{(i)}$ for all $\theta \in C$. Consequently, each preactivation coordinate $(i, j) \in [m] \times [q_\ell]$ agrees on C with the globally defined affine function

$$\widetilde{a}_{\ell,j}^{(i)}(\theta) \;:=\; (A_\ell)_{j,\cdot}\big(U_{\ell-1}^{(i)}\theta + v_{\ell-1}^{(i)}\big) + (b_\ell)_j \qquad (\theta \in \mathbb{R}^p).$$

Applying Lemma 3 to the family $\{\widetilde{a}_{\ell,j}^{(i)} : i \in [m],\, j \in [q_\ell]\}$ of $m\, q_\ell$ affine functions over $C \subseteq \mathbb{R}^p$ partitions C into at most $\Phi_p(m\, q_\ell)$ nonempty subcells on each of which every weak sign $\mathrm{sgn}\big(\widetilde{a}_{\ell,j}^{(i)}(\theta)\big) \in \{-1, 0, +1\}$ is constant. On any such subcell C', for every $(i, j) \in [m] \times [q_\ell]$, the post-activation coordinate $z_{\ell,j}(u_i(\theta)) = \max\big\{\widetilde{a}_{\ell,j}^{(i)}(\theta), 0\big\}$ equals $\widetilde{a}_{\ell,j}^{(i)}(\theta)$ throughout C' when the constant sign is $+1$ and equals 0 throughout C' when the constant sign is -1 or 0; in all three cases it agrees on C' with an affine map on $\mathbb{R}^p$. Collecting the refined subcells C' across all $C \in \mathcal{P}_{\ell-1}$ yields a partition $\mathcal{P}_\ell$ of Θ satisfying $|\mathcal{P}_\ell| \le |\mathcal{P}_{\ell-1}|\, \Phi_p(m\, q_\ell) \le \prod_{s=1}^{\ell} \Phi_p(m\, q_s)$, completing the induction.

At the output layer, fix any terminal hidden-layer cell $C \in \mathcal{P}_L$. By invariant (b) at $\ell = L$, the top hidden representation $\theta \mapsto z_L(u_i(\theta))$ agrees on C with an affine map on $\mathbb{R}^p$ for every $i \in [m]$, and therefore so do every output score coordinate $s_{\mathrm{S},k}(u_i(\theta)) = (A_{L+1})_{k,\cdot} z_L(u_i(\theta)) + (b_{L+1})_k$ and every pairwise score difference

$$\Delta_{i,kr}(\theta) \;:=\; s_{\mathrm{S},k}\big(u_i(\theta)\big) - s_{\mathrm{S},r}\big(u_i(\theta)\big), \qquad i \in [m],\ 1 \le k < r \le K_{\mathrm{S}}.$$

Applying Lemma 3 to these $m\binom{K_{\mathrm{S}}}{2}$ affine extensions over C (with $\Phi_p(0) = 1$ when $K_{\mathrm{S}} = 1$) refines C into at most $\Phi_p\big(m\binom{K_{\mathrm{S}}}{2}\big)$ nonempty subcells on each of which all output pairwise weak signs are constant. By the lexicographic weak-sign characterization Eqn. (10) applied to $t = s_{\mathrm{S}}(u_i(\theta))$, fixed pairwise weak signs uniquely fix $f_{\mathrm{S}}(u_i(\theta))$ for every $i \in [m]$. Taking the union of refined subcells across all $C \in \mathcal{P}_L$ partitions Θ into at most

$$\Phi_p\left(m\binom{K_{\mathrm{S}}}{2}\right) |\mathcal{P}_L| \;\le\; \Phi_p\left(m\binom{K_{\mathrm{S}}}{2}\right) \prod_{\ell=1}^{L} \Phi_p(m\, q_\ell)$$

constant-source-label cells, so $|\mathrm{Rea}_{\mathcal{G}}(x_{1:m})|$ satisfies the same bound and taking the supremum over $x_{1:m} \in \mathcal{X}_{\mathrm{T}}^m$ gives $\Pi_{\mathcal{R}}(m) \le \Phi_p\big(m\binom{K_{\mathrm{S}}}{2}\big) \prod_{\ell=1}^{L} \Phi_p(m\, q_\ell)$. Setting $C_{\mathcal{R}} := 1 + 2\max\big\{\binom{K_{\mathrm{S}}}{2}, q_1, \ldots, q_L\big\}$, taking logarithms, and applying $\log \Phi_p(N) \le p \log(1 + 2N)$ (Lemma 3) yields

$$\begin{aligned}
\log \Pi_{\mathcal{R}}(m) &\le \log \Phi_p\left(m\binom{K_{\mathrm{S}}}{2}\right) + \sum_{\ell=1}^{L} \log \Phi_p(m\, q_\ell) \;=\; \mathcal{C}_{\mathrm{ReLU}}(m) \\
&\le p \log\left(1 + 2m\binom{K_{\mathrm{S}}}{2}\right) + \sum_{\ell=1}^{L} p \log(1 + 2m\, q_\ell) && \text{(Lemma 3)} \\
&\le p(L+1)\log(C_{\mathcal{R}}\, m) \;=\; O\big(p(L+1)\log m\big), && (1 + 2mq \le C_{\mathcal{R}}\, m)
\end{aligned}$$

which also shows that the cell count satisfies $\Phi_p\big(m\binom{K_{\mathrm{S}}}{2}\big) \prod_{\ell=1}^{L} \Phi_p(m\, q_\ell) \le (C_{\mathcal{R}}\, m)^{(L+1)p}$, discharging Condition A with $c_{\mathcal{R}} = L + 1$ and $C_{\mathcal{R}} = 1 + 2\max\big\{\binom{K_{\mathrm{S}}}{2}, q_1, \ldots, q_L\big\}$. □

E Proof of the Lower-Bound Result

Proof roadmap. This section constructs the explicit binary reprogramming family of Proposition 3 and proves the matching minimax estimation lower bound of Theorem 3 together with its sample-complexity consequence (Corollary 5). In Appendix E.1, we give a single self-contained finite-support construction of $(\mathcal{X}_{\mathrm{T}}, \mathcal{X}_{\mathrm{S}}, f_{\mathrm{S}}, \mathcal{G}, \mathcal{H}, \mathfrak{F}^{\mathrm{bin}}_{d_{\mathcal{R}},d_{\mathcal{H}}}(\gamma))$ and verify parameter injectivity, exact reachability and conditional-relabeling growth ($\Pi_{\mathcal{R}}(m) = 2^{d_{\mathcal{R}}}$ and $\Pi_{\mathcal{H}|\mathcal{R}}(m) = 2^{d_{\mathcal{H}}}$), full binary hypercube realizability on $x_{1:m}$, and zero approximation error ($a_{\mathrm{MR}} = 0$). In Appendix E.2, we first isolate the neighboring product-law KL divergence bound and hypercube Assouad reduction in Lemma 4, then prove Theorem 9 (restating Theorem 3) and Corollary 5 directly on the construction of Appendix E.1, and finally record the matching specialization of Theorem 1 in Remark 8.

E.1 Full construction of the binary MR family (Proposition 3)

We construct the canonical finite-support objects claimed in Proposition 3 and verify each combinatorial and population property explicitly.

Parameters and label-set identification. Fix integers $d_{\mathcal{R}}, d_{\mathcal{H}} \geq 1$, and set $m := d_{\mathcal{R}} + d_{\mathcal{H}}$, $K_{\mathrm{T}} := 2$, and $K_{\mathrm{S}} := d_{\mathcal{H}} + 2$. For indexing convenience, we identify the binary target label space $[K_{\mathrm{T}}]$ with $\{0,1\}$ and the source label space $[K_{\mathrm{S}}]$ with $\{0,1,\ldots,d_{\mathcal{H}}+1\}$ via the shift bijection $y \leftrightarrow y-1$.

Target and source input spaces. Let $\mathcal{X}_{\mathrm{T}} := \{x_1,\ldots,x_m\}$ be a set of m distinct points equipped with the discrete σ-algebra, and denote the canonical support tuple by $x_{1:m} := (x_1,\ldots,x_m) \in \mathcal{X}_{\mathrm{T}}^m$. Let

$$\mathcal{X}_{\mathrm{S}} \;:=\; \{u_j^0, u_j^1 : 1 \leq j \leq d_{\mathcal{R}}\} \;\cup\; \{v_k : 1 \leq k \leq d_{\mathcal{H}}\}$$

be a set of $2d_{\mathcal{R}} + d_{\mathcal{H}}$ pairwise distinct points, also equipped with the discrete σ-algebra. The paired points (u_j^0, u_j^1) carry the $d_{\mathcal{R}}$ input-side reachability bits, while the singletons v_k carry the $d_{\mathcal{H}}$ output-side relabeling bits.

Frozen source classifier. Define $f_{\mathrm{S}} : \mathcal{X}_{\mathrm{S}} \to [K_{\mathrm{S}}] \cong \{0,1,\ldots,d_{\mathcal{H}}+1\}$ by

$$f_{\mathrm{S}}(u_j^0) \;:=\; 0, \qquad f_{\mathrm{S}}(u_j^1) \;:=\; 1, \qquad f_{\mathrm{S}}(v_k) \;:=\; k+1, \qquad 1 \leq j \leq d_{\mathcal{R}},\ 1 \leq k \leq d_{\mathcal{H}}.$$

Labels 0 and 1 are shared across all $1 \leq j \leq d_{\mathcal{R}}$, whereas each v_k receives its dedicated private source label $k+1 \in \{2,\ldots,d_{\mathcal{H}}+1\}$. Measurability is immediate under the discrete σ-algebra.

Input-transformation and output-mapping classes. For each bit vector $\alpha = (\alpha_1,\ldots,\alpha_{d_{\mathcal{R}}}) \in \{0,1\}^{d_{\mathcal{R}}}$, define $g_\alpha : \mathcal{X}_{\mathrm{T}} \to \mathcal{X}_{\mathrm{S}}$ by

$$g_\alpha(x_j) \;:=\; u_j^{\alpha_j} \quad (1 \leq j \leq d_{\mathcal{R}}), \qquad g_\alpha(x_{d_{\mathcal{R}}+k}) \;:=\; v_k \quad (1 \leq k \leq d_{\mathcal{H}}),$$

and set $\mathcal{G} := \{g_\alpha : \alpha \in \{0,1\}^{d_{\mathcal{R}}}\}$. For each bit vector $\beta = (\beta_1,\ldots,\beta_{d_{\mathcal{H}}}) \in \{0,1\}^{d_{\mathcal{H}}}$, define $h_\beta : \{0,1,\ldots,d_{\mathcal{H}}+1\} \to \{0,1\}$ by

$$h_\beta(0) \;:=\; 0, \qquad h_\beta(1) \;:=\; 1, \qquad h_\beta(k+1) \;:=\; \beta_k \quad (1 \leq k \leq d_{\mathcal{H}}),$$

and set $\mathcal{H} := \{h_\beta : \beta \in \{0,1\}^{d_{\mathcal{H}}}\}$. Both parameterizations are injective: if $\alpha \neq \alpha'$, then $\alpha_j \neq \alpha'_j$ for some $j \leq d_{\mathcal{R}}$ so $g_\alpha(x_j) = u_j^{\alpha_j} \neq u_j^{\alpha'_j} = g_{\alpha'}(x_j)$; if $\beta \neq \beta'$, then $\beta_k \neq \beta'_k$ for some $k \leq d_{\mathcal{H}}$ so $h_\beta(k+1) = \beta_k \neq \beta'_k = h_{\beta'}(k+1)$. Hence $|\mathcal{G}| = 2^{d_{\mathcal{R}}}$ and $|\mathcal{H}| = 2^{d_{\mathcal{H}}}$.

Induced wrapper class and hypercube bijection on $x_{1:m}$. Write $f_{\alpha,\beta} := h_\beta \circ f_{\mathrm{S}} \circ g_\alpha$ and $\mathcal{F}_{\mathrm{MR}} = \{f_{\alpha,\beta} : \alpha \in \{0,1\}^{d_{\mathcal{R}}},\ \beta \in \{0,1\}^{d_{\mathcal{H}}}\}$. On the first $d_{\mathcal{R}}$ support points ($1 \leq j \leq d_{\mathcal{R}}$),

$$f_{\alpha,\beta}(x_j) \;=\; h_\beta\big(f_{\mathrm{S}}(u_j^{\alpha_j})\big) \;=\; h_\beta(\alpha_j) \;=\; \alpha_j,$$

because $\alpha_j \in \{0,1\}$ and h_β fixes 0 and 1. On the remaining $d_{\mathcal{H}}$ support points ($1 \leq k \leq d_{\mathcal{H}}$),

$$f_{\alpha,\beta}(x_{d_{\mathcal{R}}+k}) \;=\; h_\beta\big(f_{\mathrm{S}}(v_k)\big) \;=\; h_\beta(k+1) \;=\; \beta_k.$$

Thus, for every $\omega = (\alpha,\beta) \in \{0,1\}^{d_{\mathcal{R}}} \times \{0,1\}^{d_{\mathcal{H}}} = \{0,1\}^m$, the wrapper $f_\omega := f_{\alpha,\beta} \in \mathcal{F}_{\mathrm{MR}}$ satisfies

$$\big(f_\omega(x_1),\ldots,f_\omega(x_m)\big) \;=\; (\alpha_1,\ldots,\alpha_{d_{\mathcal{R}}},\beta_1,\ldots,\beta_{d_{\mathcal{H}}}) \;=\; \omega,$$

establishing a bijection $(\alpha,\beta) \mapsto (f_{\alpha,\beta}(x_1),\ldots,f_{\alpha,\beta}(x_m))$ from $\{0,1\}^{d_{\mathcal{R}}} \times \{0,1\}^{d_{\mathcal{H}}}$ onto $\{0,1\}^m$.

Exact reachability and conditional-relabeling growth. Because $|\mathrm{Rea}_{\mathcal{G}}(x'_{1:m'})| \le |\mathcal{R}| \le |\mathcal{G}| = 2^{d_{\mathcal{R}}}$ and $|\mathrm{Rel}_{\mathcal{H}}(z')| \le |\mathcal{H}| = 2^{d_{\mathcal{H}}}$ for every $m' \ge 1$, every $x'_{1:m'} \in \mathcal{X}_{\mathrm{T}}^{m'}$, and every $z' \in [K_{\mathrm{S}}]^{m'}$, we have

$$\Pi_{\mathcal{R}}(m') \;\le\; 2^{d_{\mathcal{R}}} \qquad \text{and} \qquad \Pi_{\mathcal{H}|\mathcal{R}}(m') \;\le\; 2^{d_{\mathcal{H}}} \qquad \text{for every } m' \ge 1.$$

On the canonical support $x_{1:m}$ ($m' = m$), each $r_\alpha := f_{\mathrm{S}} \circ g_\alpha$ induces the source-label vector

$$z_\alpha \;:=\; \big(r_\alpha(x_1), \dots, r_\alpha(x_m)\big) \;=\; (\alpha_1, \dots, \alpha_{d_{\mathcal{R}}}, 2, 3, \dots, d_{\mathcal{H}} + 1).$$

Distinct $\alpha \in \{0,1\}^{d_{\mathcal{R}}}$ yield distinct z_α, so $|\mathrm{Rea}_{\mathcal{G}}(x_{1:m})| = 2^{d_{\mathcal{R}}}$ and hence $\Pi_{\mathcal{R}}(m) = 2^{d_{\mathcal{R}}}$ (and $\Pi_{\mathcal{R}}(m') = 2^{d_{\mathcal{R}}}$ for all $m' \ge d_{\mathcal{R}}$). Moreover, for every $z_\alpha \in \mathrm{Rea}_{\mathcal{G}}(x_{1:m})$, relabeling z_α by $h_\beta \in \mathcal{H}$ yields $(h_\beta(z_{\alpha,1}), \dots, h_\beta(z_{\alpha,m})) = (\alpha_1, \dots, \alpha_{d_{\mathcal{R}}}, \beta_1, \dots, \beta_{d_{\mathcal{H}}})$, which takes $2^{d_{\mathcal{H}}}$ distinct values as β ranges over $\{0,1\}^{d_{\mathcal{H}}}$; hence $|\mathrm{Rel}_{\mathcal{H}}(z_\alpha)| = 2^{d_{\mathcal{H}}}$, proving $\Pi_{\mathcal{H}|\mathcal{R}}(m) = 2^{d_{\mathcal{H}}}$ (and $\Pi_{\mathcal{H}|\mathcal{R}}(m') = 2^{d_{\mathcal{H}}}$ for all $m' \ge d_{\mathcal{H}}$).

Target distribution family and zero approximation error. Fix $\gamma \in (0, 1/2]$. For each $\omega \in \{0,1\}^m$, define the target probability measure P_{T}^ω on $\mathcal{X}_{\mathrm{T}} \times \{0,1\}$ by

$$P_{\mathrm{T}}^\omega(X = x_i) = \frac{1}{m},\; P_{\mathrm{T}}^\omega(Y = \omega_i \mid X = x_i) = \frac{1+\gamma}{2},\; P_{\mathrm{T}}^\omega(Y = 1 - \omega_i \mid X = x_i) = \frac{1-\gamma}{2},$$

for all $i \in [m]$, and collect these 2^m distributions into

$$\mathfrak{F}^{\mathrm{bin}}_{d_{\mathcal{R}}, d_{\mathcal{H}}}(\gamma) \;:=\; \big\{P_{\mathrm{T}}^\omega : \omega \in \{0,1\}^m\big\}.$$

Since $\gamma > 0$, $(1+\gamma)/2 > (1-\gamma)/2$, so the unique Bayes-optimal classifier on $\mathcal{X}_{\mathrm{T}}$ under P_{T}^ω is $b_{\mathrm{T}}^\omega(x_i) = \omega_i$ ($i \in [m]$), achieving Bayes risk $R_{\mathrm{T}}^*(P_{\mathrm{T}}^\omega) = (1-\gamma)/2$. Because $f_\omega \in \mathcal{F}_{\mathrm{MR}}$ satisfies $f_\omega(x_i) = \omega_i = b_{\mathrm{T}}^\omega(x_i)$ for all $i \in [m]$, we have $R_{\mathrm{T}}(f_\omega) = R_{\mathrm{T}}^*(P_{\mathrm{T}}^\omega)$ and hence $R_{\mathrm{T}}^{\mathcal{F}_{\mathrm{MR}},*}(P_{\mathrm{T}}^\omega) = R_{\mathrm{T}}^*(P_{\mathrm{T}}^\omega)$, i.e.,

$$a_{\mathrm{MR}}(P_{\mathrm{T}}^\omega) \;=\; R_{\mathrm{T}}^{\mathcal{F}_{\mathrm{MR}},*}(P_{\mathrm{T}}^\omega) - R_{\mathrm{T}}^*(P_{\mathrm{T}}^\omega) \;=\; 0 \qquad \text{for every } P_{\mathrm{T}}^\omega \in \mathfrak{F}^{\mathrm{bin}}_{d_{\mathcal{R}}, d_{\mathcal{H}}}(\gamma).$$

(If an ambient Euclidean domain is desired, embedding $\mathcal{X}_{\mathrm{T}}$ and $\mathcal{X}_{\mathrm{S}}$ into Euclidean spaces and extending $f_{\mathrm{S}} \equiv 0$ outside $\mathcal{X}_{\mathrm{S}}$ preserves measurability without altering any distribution in $\mathfrak{F}^{\mathrm{bin}}_{d_{\mathcal{R}}, d_{\mathcal{H}}}(\gamma)$, since P_{T}^ω is supported on $\mathcal{X}_{\mathrm{T}}$.) This completes the verification of Proposition 3.

E.2 Proof of Theorem 3 and Corollary 5

We first establish the pairwise product-measure KL bound and hypercube Assouad reduction used in the proof of Theorem 3. Write $\rho_{\mathrm{H}}(\omega, \omega') := \sum_{i=1}^m \mathbf{1}\{\omega_i \ne \omega'_i\}$ for Hamming distance on $\{0,1\}^m$, $\mathrm{TV}(P, Q) := \sup_A |P(A) - Q(A)|$ for total variation distance over measurable events A, and $\mathrm{KL}(P \,\|\, Q) := \int \log(\mathrm{d}P/\mathrm{d}Q)\, \mathrm{d}P$ for Kullback–Leibler divergence.

Lemma 4 (Neighboring-law KL bound and hypercube Assouad reduction)**.** *Let $\mathfrak{F}^{\mathrm{bin}}_{d_{\mathcal{R}}, d_{\mathcal{H}}}(\gamma) = \{P_{\mathrm{T}}^\omega : \omega \in \{0,1\}^m\}$ be the family of Appendix E.1 with $m = d_{\mathcal{R}} + d_{\mathcal{H}} \ge 2$ and $\gamma \in (0, 1/2]$, and let $n_{\mathrm{T}} \ge 1$. For any $\omega, \omega' \in \{0,1\}^m$ satisfying $\rho_{\mathrm{H}}(\omega, \omega') = 1$, the n_{T}-fold product laws satisfy*

$$\mathrm{KL}\Big((P_{\mathrm{T}}^\omega)^{\otimes n_{\mathrm{T}}} \,\Big\|\, (P_{\mathrm{T}}^{\omega'})^{\otimes n_{\mathrm{T}}}\Big) \;=\; \frac{n_{\mathrm{T}}\gamma}{m} \log\!\left(\frac{1+\gamma}{1-\gamma}\right) \;\le\; \frac{4 n_{\mathrm{T}} \gamma^2}{m}.$$

Consequently, if $4n_{\mathrm{T}}\gamma^2/m \le \bar{\alpha}$ for some $\bar{\alpha} > 0$, then every measurable estimator $\widehat{\omega} : (\mathcal{X}_{\mathrm{T}} \times \{0,1\})^{n_{\mathrm{T}}} \to \{0,1\}^m$ satisfies

$$\sup_{\omega \in \{0,1\}^m} \mathbb{E}_{(P_{\mathrm{T}}^\omega)^{\otimes n_{\mathrm{T}}}}\big[\rho_{\mathrm{H}}(\widehat{\omega}, \omega)\big] \;\ge\; \frac{m}{2} \max\left\{\frac{1}{2} e^{-\bar{\alpha}},\, 1 - \sqrt{\frac{\bar{\alpha}}{2}}\right\} \;\ge\; \frac{m}{4} e^{-\bar{\alpha}}.$$

Proof. Suppose $\omega, \omega' \in \{0,1\}^m$ differ only at coordinate $i \in [m]$. Under P_{T}^ω and $P_{\mathrm{T}}^{\omega'}$, the marginal law of X is uniform on $\mathcal{X}_{\mathrm{T}}$, and for every $j \ne i$ the conditional laws $P_{\mathrm{T}}^\omega(Y = \cdot \mid X = x_j)$ and $P_{\mathrm{T}}^{\omega'}(Y = \cdot \mid X = x_j)$ coincide; at $X = x_i$ (which carries marginal mass $1/m$), the Bernoulli parameters $(1+\gamma)/2$ and $(1-\gamma)/2$ are swapped. Using additivity of KL divergence across n_{T}

independent draws and $1-t^2 \geq 1-\gamma^2 \geq 3/4 > 1/2$ for all $t \in [0,\gamma] \subseteq [0,1/2]$, direct summation over $\mathcal{X}_{\mathrm{T}} \times \{0,1\}$ gives

$$\begin{aligned}
\mathrm{KL}\Big((P_{\mathrm{T}}^{\omega})^{\otimes n_{\mathrm{T}}} \,\Big\|\, (P_{\mathrm{T}}^{\omega'})^{\otimes n_{\mathrm{T}}}\Big) &= n_{\mathrm{T}}\,\mathrm{KL}(P_{\mathrm{T}}^{\omega} \,\|\, P_{\mathrm{T}}^{\omega'}) \\
&= \frac{n_{\mathrm{T}}}{m}\left\{\frac{1+\gamma}{2}\log\!\left(\frac{1+\gamma}{1-\gamma}\right) + \frac{1-\gamma}{2}\log\!\left(\frac{1-\gamma}{1+\gamma}\right)\right\} \\
&= \frac{n_{\mathrm{T}}\gamma}{m}\log\!\left(\frac{1+\gamma}{1-\gamma}\right) \;=\; \frac{2n_{\mathrm{T}}\gamma}{m}\int_0^{\gamma}\frac{\mathrm{d}t}{1-t^2} \\
&\leq \frac{2n_{\mathrm{T}}\gamma}{m}\int_0^{\gamma} 2\,\mathrm{d}t \;=\; \frac{4n_{\mathrm{T}}\gamma^2}{m},
\end{aligned}$$

where the first equality uses KL tensorization across independent draws and the inequality uses $(1-t^2)^{-1} \leq 2$ on $[0,\gamma]$.

Now suppose $4n_{\mathrm{T}}\gamma^2/m \leq \bar{\alpha}$ with $\bar{\alpha} > 0$. Write $P_\omega := (P_{\mathrm{T}}^{\omega})^{\otimes n_{\mathrm{T}}}$ with probability $\mathbb{P}_\omega$ and expectation $\mathbb{E}_\omega$, and for each $\omega \in \{0,1\}^m$ and $i \in [m]$ let $\omega^{\oplus i} \in \{0,1\}^m$ denote the neighbor obtained by flipping coordinate i of ω, so $\rho_{\mathrm{H}}(\omega,\omega^{\oplus i}) = 1$ and $\mathrm{KL}(P_\omega \,\|\, P_{\omega^{\oplus i}}) \leq \bar{\alpha}$. For any measurable estimator $\widehat{\omega} : (\mathcal{X}_{\mathrm{T}} \times \{0,1\})^{n_{\mathrm{T}}} \to \{0,1\}^m$, averaging over the uniform prior on $\{0,1\}^m$ gives

$$\begin{aligned}
\sup_{\omega\in\{0,1\}^m} \mathbb{E}_\omega\big[\rho_{\mathrm{H}}(\widehat{\omega},\omega)\big] &\geq \frac{1}{2^m}\sum_{\omega\in\{0,1\}^m} \mathbb{E}_\omega\left[\sum_{i=1}^{m}\mathbf{1}\{\widehat{\omega}_i \neq \omega_i\}\right] && \text{(uniform prior)} \\
&= \sum_{i=1}^{m}\frac{1}{2^m}\sum_{\omega\in\{0,1\}^m} \mathbb{P}_\omega(\widehat{\omega}_i \neq \omega_i) && \text{(linearity)} \\
&= \frac{1}{2}\sum_{i=1}^{m}\frac{1}{2^{m-1}}\sum_{\omega:\,\omega_i=0}\Big(\mathbb{P}_\omega(\widehat{\omega}_i = 1) + \mathbb{P}_{\omega^{\oplus i}}(\widehat{\omega}_i = 0)\Big) && \text{(pairing } \omega,\omega^{\oplus i}) \\
&\geq \frac{1}{2}\sum_{i=1}^{m}\frac{1}{2^{m-1}}\sum_{\omega:\,\omega_i=0}\big(1 - \mathrm{TV}(P_\omega, P_{\omega^{\oplus i}})\big) && \text{(Le Cam testing)} \\
&\geq \frac{1}{2}\sum_{i=1}^{m}\frac{1}{2^{m-1}}\sum_{\omega:\,\omega_i=0}\max\left\{\frac{1}{2}e^{-\bar{\alpha}},\, 1-\sqrt{\frac{\bar{\alpha}}{2}}\right\} && (\mathrm{KL} \leq \bar{\alpha}) \\
&= \frac{m}{2}\max\left\{\frac{1}{2}e^{-\bar{\alpha}},\, 1-\sqrt{\frac{\bar{\alpha}}{2}}\right\} \;\geq\; \frac{m}{4}e^{-\bar{\alpha}},
\end{aligned}$$

where the fourth relation applies the two-point testing identity $\mathbb{P}_\omega(\widehat{\omega}_i = 1) + \mathbb{P}_{\omega^{\oplus i}}(\widehat{\omega}_i = 0) = 1 - \big(P_{\omega^{\oplus i}}(\widehat{\omega}_i = 1) - P_\omega(\widehat{\omega}_i = 1)\big) \geq 1 - \mathrm{TV}(P_\omega, P_{\omega^{\oplus i}})$, and the fifth applies the Bretagnolle–Huber inequality $1 - \mathrm{TV}(P,Q) \geq \frac{1}{2}e^{-\mathrm{KL}(P\,\|\,Q)}$ together with Pinsker's inequality $\mathrm{TV}(P,Q) \leq \sqrt{\mathrm{KL}(P\,\|\,Q)/2}$ [46, Lemmas 2.5 and 2.6, Theorem 2.12] under $\mathrm{KL}(P_\omega \,\|\, P_{\omega^{\oplus i}}) \leq \bar{\alpha}$ and $|\{\omega \in \{0,1\}^m : \omega_i = 0\}| = 2^{m-1}$. □

Theorem 9 (Matching minimax lower bound; restatement of Theorem 3)**.** *With $c_0 = 1$, $c_1 = 1/4$, and $c_2 = e^{-1/4}/16$, for all integers $d_{\mathcal{R}}, d_{\mathcal{H}} \geq 1$ and $n_{\mathrm{T}} \geq 1$ satisfying $d_{\mathcal{R}} + d_{\mathcal{H}} \leq c_0 n_{\mathrm{T}}$, on the binary reprogramming family $\mathfrak{F}^{\mathrm{bin}}_{d_{\mathcal{R}},d_{\mathcal{H}}}(\gamma)$ (Proposition 3 and Appendix E.1) with margin $\gamma = c_1\sqrt{(d_{\mathcal{R}}+d_{\mathcal{H}})/n_{\mathrm{T}}}$, every learning rule $\widehat{f}_n$ mapping n_{T} i.i.d. target samples to a measurable predictor $\mathcal{X}_{\mathrm{T}} \to [K_{\mathrm{T}}]$ satisfies*

$$\sup_{P_{\mathrm{T}}\in\mathfrak{F}^{\mathrm{bin}}_{d_{\mathcal{R}},d_{\mathcal{H}}}(\gamma)} \mathbb{E}_{P_{\mathrm{T}}^{n_{\mathrm{T}}}}\big[\,\mathcal{E}^{\mathrm{est}}_{\mathrm{T}}(\widehat{f}_n)\,\big] \;\geq\; c_2\sqrt{\frac{d_{\mathcal{R}}+d_{\mathcal{H}}}{n_{\mathrm{T}}}}.$$

Proof. Let $m := d_{\mathcal{R}} + d_{\mathcal{H}} \geq 2$ and work on the family $\mathfrak{F}^{\mathrm{bin}}_{d_{\mathcal{R}},d_{\mathcal{H}}}(\gamma) = \{P_{\mathrm{T}}^{\omega} : \omega \in \{0,1\}^m\}$ of Appendix E.1, writing $R_{\mathrm{T}}^{\omega}(f) := P_{\mathrm{T}}^{\omega}(f(X) \neq Y)$ for the target 0–1 risk under P_{T}^{ω}. By Appendix E.1,

$a_{\mathrm{MR}}(P_{\mathrm{T}}^{\omega}) = 0$ and $R_{\mathrm{T}}^{\mathcal{F}_{\mathrm{MR}},*}(P_{\mathrm{T}}^{\omega}) = R_{\mathrm{T}}^{*}(P_{\mathrm{T}}^{\omega}) = (1-\gamma)/2$ for every $\omega \in \{0,1\}^m$. Conditioned on the realized n_{T}-sample, any learning rule fixes a measurable predictor $\widehat{f}_n : \mathcal{X}_{\mathrm{T}} \to \{0,1\}$ (whether or not $\widehat{f}_n$ takes values in $\mathcal{F}_{\mathrm{MR}}$) and induces the bit vector $\widehat{\omega} = (\widehat{\omega}_1, \ldots, \widehat{\omega}_m) \in \{0,1\}^m$ defined by $\widehat{\omega}_i := \widehat{f}_n(x_i)$ for $i \in [m]$. At each support point x_i, $P_{\mathrm{T}}^{\omega}\big(\widehat{f}_n(x_i) \neq Y \mid X = x_i\big)$ equals $(1-\gamma)/2$ when $\widehat{\omega}_i = \omega_i$ and $(1+\gamma)/2 = (1-\gamma)/2 + \gamma$ when $\widehat{\omega}_i \neq \omega_i$. Disintegrating $R_{\mathrm{T}}^{\omega}(\widehat{f}_n)$ over the uniform marginal $P_{\mathrm{T}}^{\omega}(X = x_i) = 1/m$ gives

$$
\begin{aligned}
\mathcal{E}_{\mathrm{T}}^{\mathrm{est}}(\widehat{f}_n) &= R_{\mathrm{T}}^{\omega}(\widehat{f}_n) - R_{\mathrm{T}}^{\mathcal{F}_{\mathrm{MR}},*}(P_{\mathrm{T}}^{\omega}) && \text{(Definition 1)} \\
&= R_{\mathrm{T}}^{\omega}(\widehat{f}_n) - R_{\mathrm{T}}^{*}(P_{\mathrm{T}}^{\omega}) && (a_{\mathrm{MR}}(P_{\mathrm{T}}^{\omega}) = 0) \\
&= \frac{1}{m}\sum_{i=1}^{m} P_{\mathrm{T}}^{\omega}\big(\widehat{f}_n(x_i) \neq Y \;\big|\; X = x_i\big) - \frac{1-\gamma}{2} && \text{(uniform marginal)} \\
&= \frac{1}{m}\sum_{i=1}^{m} \left(\frac{1-\gamma}{2} + \gamma\,\mathbf{1}\{\widehat{\omega}_i \neq \omega_i\}\right) - \frac{1-\gamma}{2} \\
&= \frac{1}{m}\sum_{i=1}^{m} \gamma\,\mathbf{1}\{\widehat{\omega}_i \neq \omega_i\} \;=\; \frac{\gamma}{m}\,\rho_{\mathrm{H}}(\widehat{\omega}, \omega). && (11)
\end{aligned}
$$

Under $m \leq c_0 n_{\mathrm{T}} = n_{\mathrm{T}}$ and $\gamma = c_1\sqrt{m/n_{\mathrm{T}}} = \frac{1}{4}\sqrt{m/n_{\mathrm{T}}}$, we have $\gamma \in (0, 1/4] \subset (0, 1/2]$ and

$$
\frac{4n_{\mathrm{T}}\gamma^2}{m} \;=\; \frac{4n_{\mathrm{T}}}{m} \cdot \frac{m}{16 n_{\mathrm{T}}} \;=\; \frac{1}{4}.
$$

Applying Lemma 4 with $\bar{\alpha} := 1/4$ to the measurable estimator $\widehat{\omega}$ induced by $\widehat{f}_n$, taking expectation over $(P_{\mathrm{T}}^{\omega})^{\otimes n_{\mathrm{T}}}$ in Eqn. (11), and taking the supremum over $\omega \in \{0,1\}^m$ yields

$$
\begin{aligned}
\sup_{P_{\mathrm{T}} \in \mathfrak{F}^{\mathrm{bin}}_{d_{\mathcal{R}}, d_{\mathcal{H}}}(\gamma)} \mathbb{E}_{P_{\mathrm{T}}^{n_{\mathrm{T}}}}\big[\,\mathcal{E}_{\mathrm{T}}^{\mathrm{est}}(\widehat{f}_n)\,\big] &= \frac{\gamma}{m} \sup_{\omega \in \{0,1\}^m} \mathbb{E}_{(P_{\mathrm{T}}^{\omega})^{\otimes n_{\mathrm{T}}}}\big[\rho_{\mathrm{H}}(\widehat{\omega}, \omega)\big] && \text{(Eqn. (11))} \\
&\geq \frac{\gamma}{m} \cdot \frac{m}{4} e^{-1/4} && \text{(Lemma 4)} \\
&= \frac{e^{-1/4}}{16}\sqrt{\frac{d_{\mathcal{R}} + d_{\mathcal{H}}}{n_{\mathrm{T}}}}, && (\gamma = \tfrac{1}{4}\sqrt{m/n_{\mathrm{T}}})
\end{aligned}
$$

establishing the claim with $c_2 = e^{-1/4}/16$. □

Proof of Corollary 5. Fix integers $d_{\mathcal{R}}, d_{\mathcal{H}} \geq 1$, set $m := d_{\mathcal{R}} + d_{\mathcal{H}} \geq 2$, fix $\varepsilon \in (0, c_2/2)$, and let $\widehat{f}_n$ be any learning rule on $n_{\mathrm{T}} \geq 1$ i.i.d. target samples achieving $\sup_{P_{\mathrm{T}} \in \mathfrak{F}^{\mathrm{bin}}_{d_{\mathcal{R}}, d_{\mathcal{H}}}(\gamma)} \mathbb{E}_{P_{\mathrm{T}}^{n_{\mathrm{T}}}}\big[\mathcal{E}_{\mathrm{T}}^{\mathrm{est}}(\widehat{f}_n)\big] \leq \varepsilon$ at margin $\gamma = \min\big\{1/4,\, c_1\sqrt{m/n_{\mathrm{T}}}\big\}$ with $c_1 = 1/4$.

Case 1 ($n_{\mathrm{T}} \geq d_{\mathcal{R}} + d_{\mathcal{H}}$). Here $n_{\mathrm{T}} \geq m$, so $\sqrt{m/n_{\mathrm{T}}} \leq 1$ and the margin evaluates to

$$
\gamma \;=\; \min\left\{\frac{1}{4},\, \frac{1}{4}\sqrt{\frac{m}{n_{\mathrm{T}}}}\right\} \;=\; \frac{1}{4}\sqrt{\frac{m}{n_{\mathrm{T}}}} \;=\; c_1\sqrt{\frac{m}{n_{\mathrm{T}}}}.
$$

Since $m \leq n_{\mathrm{T}} = c_0 n_{\mathrm{T}}$, Theorem 9 applies directly and gives

$$
\varepsilon \;\geq\; \sup_{P_{\mathrm{T}} \in \mathfrak{F}^{\mathrm{bin}}_{d_{\mathcal{R}}, d_{\mathcal{H}}}(\gamma)} \mathbb{E}_{P_{\mathrm{T}}^{n_{\mathrm{T}}}}\big[\,\mathcal{E}_{\mathrm{T}}^{\mathrm{est}}(\widehat{f}_n)\,\big] \;\geq\; c_2\sqrt{\frac{m}{n_{\mathrm{T}}}},
$$

which rearranges to $n_{\mathrm{T}} \geq c_2^2 m/\varepsilon^2 = c_2^2(d_{\mathcal{R}} + d_{\mathcal{H}})/\varepsilon^2$.

Case 2 ($1 \leq n_{\mathrm{T}} < d_{\mathcal{R}} + d_{\mathcal{H}}$). Here $1 \leq n_{\mathrm{T}} < m$, so $\sqrt{m/n_{\mathrm{T}}} > 1$ and the margin evaluates to

$$
\gamma \;=\; \min\left\{\frac{1}{4},\, \frac{1}{4}\sqrt{\frac{m}{n_{\mathrm{T}}}}\right\} \;=\; \frac{1}{4} \;\in\; (0, 1/2],
$$

for which $4n_{\mathrm{T}}\gamma^2/m = n_{\mathrm{T}}/(4m) < 1/4 =: \bar{\alpha}$. Combining Eqn. (11) with Lemma 4 at $\gamma = 1/4$ and $\bar{\alpha} = 1/4$ gives

$$\sup_{P_{\mathrm{T}} \in \mathfrak{F}^{\mathrm{bin}}_{d_{\mathcal{R}}, d_{\mathcal{H}}}(1/4)} \mathbb{E}_{P_{\mathrm{T}}^{n_{\mathrm{T}}}}\left[\, \mathcal{E}^{\mathrm{est}}_{\mathrm{T}}(\widehat{f}_n) \,\right] = \frac{1/4}{m} \sup_{\omega \in \{0,1\}^m} \mathbb{E}_{(P_{\mathrm{T}}^{\omega})^{\otimes n_{\mathrm{T}}}}\left[\rho_{\mathrm{H}}(\widehat{\omega}, \omega)\right] \qquad \text{(Eqn. (11))}$$

$$\geq \frac{1}{4m} \cdot \frac{m}{4} e^{-1/4} \;=\; \frac{e^{-1/4}}{16} \;=\; c_2, \qquad \text{(Lemma 4)}$$

contradicting $\sup_{P_{\mathrm{T}} \in \mathfrak{F}^{\mathrm{bin}}_{d_{\mathcal{R}}, d_{\mathcal{H}}}(\gamma)} \mathbb{E}_{P_{\mathrm{T}}^{n_{\mathrm{T}}}}\left[\mathcal{E}^{\mathrm{est}}_{\mathrm{T}}(\widehat{f}_n)\right] \leq \varepsilon < c_2/2 < c_2$. Thus Case 2 cannot occur, and $n_{\mathrm{T}} \geq c_2^2(d_{\mathcal{R}} + d_{\mathcal{H}})/\varepsilon^2 = \Omega\big((d_{\mathcal{R}} + d_{\mathcal{H}})/\varepsilon^2\big)$ is necessary. □

Remark 8 (Specialization of Theorem 1 on $\mathfrak{F}^{\mathrm{bin}}_{d_{\mathcal{R}}, d_{\mathcal{H}}}(\gamma)$ and scope of Theorem 3). On the binary family of Appendix E.1, we proved $a_{\mathrm{MR}} = 0$, $\log \Pi_{\mathcal{R}}(2n_{\mathrm{T}}) \leq d_{\mathcal{R}} \log 2$, and $\log \Pi_{\mathcal{H}|\mathcal{R}}(2n_{\mathrm{T}}) \leq d_{\mathcal{H}} \log 2$. Substituting these bounds into Theorem 7 (Remark 5) shows that for every $P_{\mathrm{T}} \in \mathfrak{F}^{\mathrm{bin}}_{d_{\mathcal{R}}, d_{\mathcal{H}}}(\gamma)$, every $n_{\mathrm{T}} \geq 1$, every $\delta \in (0, 1)$, and every $\eta_n \geq 0$, with probability at least $1 - \delta$ over $P_{\mathrm{T}}^{n_{\mathrm{T}}}$, every measurable η_n-approximate ERM $\widehat{f}_n \in \mathcal{F}_{\mathrm{MR}}$ satisfies

$$R_{\mathrm{T}}(\widehat{f}_n) - R^*_{\mathrm{T}} \;=\; \mathcal{E}^{\mathrm{est}}_{\mathrm{T}}(\widehat{f}_n) \;\leq\; \eta_n + 4\sqrt{6}\sqrt{\frac{(d_{\mathcal{R}} + d_{\mathcal{H}})\log 2 + \log(1/\delta)}{n_{\mathrm{T}}}},$$

matching the $\sqrt{(d_{\mathcal{R}} + d_{\mathcal{H}})/n_{\mathrm{T}}}$ scaling of Theorem 9 up to absolute constants and the high-probability confidence term. As noted in Section 5, Theorem 9 establishes existential rate tightness on $\mathfrak{F}^{\mathrm{bin}}_{d_{\mathcal{R}}, d_{\mathcal{H}}}(\gamma)$, showing that neither $d_{\mathcal{R}}$ nor $d_{\mathcal{H}}$ can be omitted in general, rather than a lower bound holding for every fixed source classifier f_{S}.

F Proofs for Exact Representation-Conditioned Approximation Identities

Proof roadmap. This section proves the three exact population approximation identities of the paper: the hard-label 0–1 decomposition (Proposition 4 in Appendix F.1), the score-output 0–1 decomposition (Proposition 5 in Appendix F.2), and the score-output surrogate decomposition (Proposition 6 in Appendix F.3). Each proof rewrites the wrapper infimum as a nested optimization over the input transformation $g \in \mathcal{G}$ and the output mapping ($h \in \mathcal{H}$ or $W \in \mathcal{W}$), characterizes the unrestricted channel optimum conditioned on the induced representation ($Z_g = f_{\mathrm{S}}(g(X))$ or $U_g = s_{\mathrm{S}}(g(X))$), and telescopes the fixed-g excess risk around that intermediate reference while verifying pointwise nonnegativity of both summands. None of the three identities requires bounded scores, Lipschitz losses, classification calibration, or finite-sample concentration.

F.1 Proof of Proposition 4

Proof. Fix $g \in \mathcal{G}$ and write $Z_g := f_{\mathrm{S}}(g(X)) \in [K_{\mathrm{S}}]$ for the induced source-label random variable under $(X, Y) \sim P_{\mathrm{T}}$. Because $f_{\mathrm{S}} : \mathcal{X}_{\mathrm{S}} \to [K_{\mathrm{S}}]$ and $g : \mathcal{X}_{\mathrm{T}} \to \mathcal{X}_{\mathrm{S}}$ are measurable, Z_g is measurable. For each $(k, y) \in [K_{\mathrm{S}}] \times [K_{\mathrm{T}}]$, set $p_g(k, y) := \mathbb{P}_{(X,Y)\sim P_{\mathrm{T}}}(Z_g = k,\, Y = y)$. For any relabeling $h \in [K_{\mathrm{T}}]^{[K_{\mathrm{S}}]}$, writing $R_{\mathrm{T}}(h(Z_g)) := R_{\mathrm{T}}(h \circ f_{\mathrm{S}} \circ g)$ and partitioning $\{h(Z_g) = Y\}$ across the disjoint level sets $\{Z_g = k\}_{k=1}^{K_{\mathrm{S}}}$ gives

$$
\begin{aligned}
R_{\mathrm{T}}\big(h(Z_g)\big) &= \mathbb{P}_{(X,Y)\sim P_{\mathrm{T}}}\big(h(Z_g) \neq Y\big) \\
&= 1 - \mathbb{P}_{(X,Y)\sim P_{\mathrm{T}}}\big(h(Z_g) = Y\big) \\
&= 1 - \sum_{k=1}^{K_{\mathrm{S}}} \mathbb{P}_{(X,Y)\sim P_{\mathrm{T}}}\big(Z_g = k,\, Y = h(k)\big) && \text{(disjoint level sets)} \\
&= 1 - \sum_{k=1}^{K_{\mathrm{S}}} p_g\big(k, h(k)\big).
\end{aligned}
$$

Since $\mathcal{G}$ and $\mathcal{H} \subseteq [K_{\mathrm{T}}]^{[K_{\mathrm{S}}]}$ are nonempty and $\mathcal{F}_{\mathrm{MR}} = \{h \circ f_{\mathrm{S}} \circ g : g \in \mathcal{G},\, h \in \mathcal{H}\}$, setting $R^*_{\mathcal{H}}(g) := \inf_{h\in\mathcal{H}} R_{\mathrm{T}}\big(h(Z_g)\big)$ factors the wrapper approximation error as

$$
\begin{aligned}
a_{\mathrm{MR}} &= R_{\mathrm{T}}^{\mathcal{F}_{\mathrm{MR}},*} - R^*_{\mathrm{T}} && \text{(by Definition 1)} \\
&= \inf_{f\in\mathcal{F}_{\mathrm{MR}}} R_{\mathrm{T}}(f) - R^*_{\mathrm{T}} \\
&= \inf_{g\in\mathcal{G}} \inf_{h\in\mathcal{H}} R_{\mathrm{T}}\big(h(Z_g)\big) - R^*_{\mathrm{T}} && \text{(nested infimum)} \\
&= \inf_{g\in\mathcal{G}} \big\{R^*_{\mathcal{H}}(g) - R^*_{\mathrm{T}}\big\}. && (\mathcal{G} \neq \varnothing)
\end{aligned}
$$

Now consider the unrestricted relabeling class $\mathcal{H}_{\mathrm{all}} := [K_{\mathrm{T}}]^{[K_{\mathrm{S}}]}$, whose coordinates $h(1), \ldots, h(K_{\mathrm{S}}) \in [K_{\mathrm{T}}]$ vary independently, and denote the active support of Z_g by $\mathcal{Z}_g := \{k \in [K_{\mathrm{S}}] : P_{\mathrm{T}}(Z_g = k) > 0\}$. For every zero-probability atom $k \in [K_{\mathrm{S}}] \setminus \mathcal{Z}_g$, monotonicity of probability gives $0 \le p_g(k, y) \le P_{\mathrm{T}}(Z_g = k) = 0$ for all $y \in [K_{\mathrm{T}}]$, so $\max_{y\in[K_{\mathrm{T}}]} p_g(k, y) = 0$; meanwhile, $P_{\mathrm{T}}(Z_g \notin \mathcal{Z}_g) = \sum_{k\notin\mathcal{Z}_g} P_{\mathrm{T}}(Z_g = k) = 0$, so any version of the conditional probability $P_{\mathrm{T}}(Y = y \mid Z_g)$ agrees with $p_g(k, y)/P_{\mathrm{T}}(Z_g = k)$ on $\{Z_g = k\}$ for every $k \in \mathcal{Z}_g$ and integrates to

0 over $\{Z_g \notin \mathcal{Z}_g\}$. Evaluating $R^*_{\text{all}}(g) := \inf_{h\in\mathcal{H}_{\text{all}}} R_{\text{T}}\big(h(Z_g)\big)$ therefore yields

$$\begin{aligned}
R^*_{\text{all}}(g) &= \inf_{h\in\mathcal{H}_{\text{all}}} \left\{ 1 - \sum_{k=1}^{K_{\text{S}}} p_g\big(k, h(k)\big) \right\} \\
&= 1 - \sup_{h\in\mathcal{H}_{\text{all}}} \sum_{k=1}^{K_{\text{S}}} p_g\big(k, h(k)\big) \\
&= 1 - \sum_{k=1}^{K_{\text{S}}} \max_{y\in[K_{\text{T}}]} p_g(k, y) \\
&= 1 - \sum_{k\in\mathcal{Z}_g} \max_{y\in[K_{\text{T}}]} p_g(k, y) \\
&= 1 - \sum_{k\in\mathcal{Z}_g} P_{\text{T}}(Z_g = k) \max_{y\in[K_{\text{T}}]} P_{\text{T}}(Y = y \mid Z_g = k) \\
&= \mathbb{E}_{P_{\text{T}}}\left[1 - \max_{y\in[K_{\text{T}}]} P_{\text{T}}(Y = y \mid Z_g) \right],
\end{aligned}$$

where the third equality decouples the supremum across $k \in [K_{\text{S}}]$, the fourth drops null atoms $k \notin \mathcal{Z}_g$, the fifth factors $p_g(k, y) = P_{\text{T}}(Z_g = k)\, P_{\text{T}}(Y = y \mid Z_g = k)$ on $\mathcal{Z}_g$, and the sixth rewrites the sum over $\mathcal{Z}_g$ as an expectation under $P_{\text{T}}(Z_g \in \mathcal{Z}_g) = 1$.

Because $R^*_{\mathcal{H}}(g), R^*_{\text{all}}(g), R^*_{\text{T}} \in [0, 1]$ are finite, adding and subtracting $R^*_{\text{all}}(g)$ inside the outer infimum over $g \in \mathcal{G}$ gives

$$a_{\text{MR}} \;=\; \inf_{g\in\mathcal{G}} \Big\{ \big(R^*_{\text{all}}(g) - R^*_{\text{T}}\big) + \big(R^*_{\mathcal{H}}(g) - R^*_{\text{all}}(g)\big) \Big\}.$$

Finally, for every $g \in \mathcal{G}$, each composition $h \circ f_{\text{S}} \circ g$ ($h \in \mathcal{H}_{\text{all}}$) is a measurable classifier $\mathcal{X}_{\text{T}} \to [K_{\text{T}}]$ and $\mathcal{H} \subseteq \mathcal{H}_{\text{all}}$, so monotonicity of the infimum gives

$$\begin{aligned}
R^*_{\text{all}}(g) - R^*_{\text{T}} &= \inf_{h\in\mathcal{H}_{\text{all}}} R_{\text{T}}(h \circ f_{\text{S}} \circ g) - \inf_{f:\mathcal{X}_{\text{T}}\to[K_{\text{T}}] \text{ measurable}} R_{\text{T}}(f) \;\geq\; 0, \\
R^*_{\mathcal{H}}(g) - R^*_{\text{all}}(g) &= \inf_{h\in\mathcal{H}} R_{\text{T}}\big(h(Z_g)\big) - \inf_{h\in\mathcal{H}_{\text{all}}} R_{\text{T}}\big(h(Z_g)\big) \;\geq\; 0,
\end{aligned}$$

establishing Eqn. (3). □

F.2 Proof of Proposition 5

Proof. Recall from Definition 4 that $\mathcal{F}_{\text{score}} = \{f_{g,W} = a_W \circ s_{\text{S}} \circ g : g \in \mathcal{G}, W \in \mathcal{W}\}$, where $a_W(u) = \text{lexargmax}_{y\in[K_{\text{T}}]}(Wu)_y$ for $u \in \mathbb{R}^{K_{\text{S}}}$. For each $W \in \mathcal{W}$ and $k \in [K_{\text{T}}]$, the level set

$$\begin{aligned}
\big\{u \in \mathbb{R}^{K_{\text{S}}} : a_W(u) = k\big\} \;=\; &\bigcap_{r<k} \big\{u \in \mathbb{R}^{K_{\text{S}}} : (Wu)_k > (Wu)_r\big\} \\
&\cap \bigcap_{r>k} \big\{u \in \mathbb{R}^{K_{\text{S}}} : (Wu)_k \geq (Wu)_r\big\}
\end{aligned}$$

is a finite intersection of open and closed half-spaces, so a_W is Borel-measurable; since s_{S} and g are measurable, $f_{g,W} = a_W(U_g)$ with $U_g := s_{\text{S}}(g(X))$ is measurable.

Fix $g \in \mathcal{G}$. Because $\mathcal{X}_{\text{T}} \times [K_{\text{T}}]$ is standard Borel and U_g takes values in $\mathbb{R}^{K_{\text{S}}}$, regular conditional probabilities $\eta^g_y(u) := \mathbb{P}_{P_{\text{T}}}(Y = y \mid U_g = u)$ exist as Borel-measurable functions of $u \in \mathbb{R}^{K_{\text{S}}}$ for each $y \in [K_{\text{T}}]$. Define the lexicographically tie-broken channel selector $b^\star_g(u) := \text{lexargmax}_{y\in[K_{\text{T}}]}\, \eta^g_y(u)$ for $u \in \mathbb{R}^{K_{\text{S}}}$. For each $k \in [K_{\text{T}}]$, $\{u \in \mathbb{R}^{K_{\text{S}}} : b^\star_g(u) = k\} = \bigcap_{r<k}\{u : \eta^g_k(u) > \eta^g_r(u)\} \cap \bigcap_{r>k}\{u : \eta^g_k(u) \geq \eta^g_r(u)\}$ is a finite intersection of Borel sets, so $b^\star_g : \mathbb{R}^{K_{\text{S}}} \to [K_{\text{T}}]$ is Borel-measurable and satisfies $\eta^g_{b^\star_g(u)}(u) = \max_{y\in[K_{\text{T}}]} \eta^g_y(u)$ for every $u \in \mathbb{R}^{K_{\text{S}}}$. For any Borel-measurable map

$a : \mathbb{R}^{K_S} \to [K_T]$, conditioning on U_g gives

$$\begin{aligned}
R_T\big(a(U_g)\big) &= \mathbb{P}_{P_T}\big(Y \neq a(U_g)\big) \\
&= \mathbb{E}_{P_T}\big[\mathbb{P}_{P_T}\big(Y \neq a(U_g) \mid U_g\big)\big] && \text{(tower property)} \\
&= \mathbb{E}_{P_T}\big[1 - \eta^g_{a(U_g)}(U_g)\big] && (\sigma(U_g)\text{-measurability}) \\
&\geq \mathbb{E}_{P_T}\Big[1 - \max_{y \in [K_T]} \eta^g_y(U_g)\Big] && (\eta^g_{a(u)}(u) \leq \max_y \eta^g_y(u)) \\
&= \mathbb{E}_{P_T}\big[1 - \eta^g_{b^\star_g(U_g)}(U_g)\big] && \text{(definition of } b^\star_g) \\
&= R_T\big(b^\star_g(U_g)\big).
\end{aligned}$$

Since $b^\star_g$ itself is Borel-measurable from $\mathbb{R}^{K_S}$ to $[K_T]$, the infimum over measurable $a : \mathbb{R}^{K_S} \to [K_T]$ is attained at $b^\star_g$, giving $R^\star_T(U_g) = R_T\big(b^\star_g(U_g)\big) = \mathbb{E}_{P_T}\big[1 - \max_{y \in [K_T]} \eta^g_y(U_g)\big]$.

Because all 0–1 risks lie in $[0, 1]$ and $\mathcal{G}, \mathcal{W}$ are nonempty, factoring the joint infimum over $(g, W) \in \mathcal{G} \times \mathcal{W}$ and adding and subtracting $R^\star_T(U_g)$ (which is independent of W) yields

$$\begin{aligned}
a_{\text{score},0/1} &= \inf_{(g,W) \in \mathcal{G} \times \mathcal{W}} R_T(f_{g,W}) - R^*_T \\
&= \inf_{g \in \mathcal{G}} \Big\{ \inf_{W \in \mathcal{W}} R_T\big(a_W(U_g)\big) - R^*_T \Big\} && (\mathcal{G}, \mathcal{W} \neq \varnothing) \\
&= \inf_{g \in \mathcal{G}} \Big\{ \big(R^\star_T(U_g) - R^*_T\big) + \Big(\inf_{W \in \mathcal{W}} R_T\big(a_W(U_g)\big) - R^\star_T(U_g) \Big) \Big\} \\
&= \inf_{g \in \mathcal{G}} \Big\{ \big(R^\star_T(U_g) - R^*_T\big) + \inf_{W \in \mathcal{W}} \big(R_T\big(a_W(U_g)\big) - R^\star_T(U_g) \big) \Big\}. && (\mathcal{W} \neq \varnothing)
\end{aligned}$$

Moreover, for every $g \in \mathcal{G}$, $b^\star_g \circ s_S \circ g : \mathcal{X}_T \to [K_T]$ is measurable and $a_W : \mathbb{R}^{K_S} \to [K_T]$ is Borel-measurable for each $W \in \mathcal{W}$, whence

$$R^\star_T(U_g) - R^*_T = R_T(b^\star_g \circ s_S \circ g) - \inf_{f : \mathcal{X}_T \to [K_T] \text{ measurable}} R_T(f) \;\geq\; 0,$$

$$\inf_{W \in \mathcal{W}} \big(R_T\big(a_W(U_g)\big) - R^\star_T(U_g) \big) = \inf_{W \in \mathcal{W}} \mathbb{E}_{P_T}\big[\eta^g_{b^\star_g(U_g)}(U_g) - \eta^g_{a_W(U_g)}(U_g)\big] \;\geq\; 0,$$

where the second inequality uses $\eta^g_{b^\star_g(u)}(u) - \eta^g_{a_W(u)}(u) \geq 0$ pointwise on $\mathbb{R}^{K_S}$. □

F.3 Proof of Proposition 6

Proof. For each $g \in \mathcal{G}$, write $U_g := s_S(g(X))$ and set $A_\ell(g) := \inf_{W \in \mathcal{W}} \mathcal{L}_{T,\ell}(g, W)$. By hypothesis, $\mathcal{G}$ and $\mathcal{W}$ are nonempty, $\ell \geq 0$, $\mathcal{L}^*_{T,\ell} < \infty$, and $A_\ell(g) < \infty$ for every $g \in \mathcal{G}$. Fix $g \in \mathcal{G}$; comparing infima across nested classes of measurable predictors gives

$$\begin{aligned}
0 &\leq \inf_{q_0 : \mathcal{X}_T \to \mathbb{R}^{K_T} \text{ measurable}} \mathbb{E}_{P_T}\big[\ell\big(Y, q_0(X)\big)\big] \;=\; \mathcal{L}^*_{T,\ell} \\
&\leq \inf_{q : \mathbb{R}^{K_S} \to \mathbb{R}^{K_T} \text{ measurable}} \mathbb{E}_{P_T}\big[\ell\big(Y, q(s_S(g(X)))\big)\big] \;=\; L^\star_\ell(U_g) \\
&\leq \inf_{W \in \mathcal{W}} \mathbb{E}_{P_T}\big[\ell\big(Y, W s_S(g(X))\big)\big] \;=\; A_\ell(g) \;<\; \infty,
\end{aligned}$$

where the first inequality uses $\ell \geq 0$, the second holds because $q \circ s_S \circ g : \mathcal{X}_T \to \mathbb{R}^{K_T}$ is measurable whenever $q : \mathbb{R}^{K_S} \to \mathbb{R}^{K_T}$ is Borel-measurable, and the third holds because every linear map $u \mapsto Wu$ ($W \in \mathcal{W}$) is continuous and hence Borel-measurable.

Because $\mathcal{L}^*_{T,\ell}, L^\star_\ell(U_g), A_\ell(g) \in [0, \infty)$ are all finite, adding and subtracting $L^\star_\ell(U_g)$ introduces no $\infty - \infty$ indeterminacy:

$$\begin{aligned}
a_{\text{score},\ell} &= \inf_{(g,W) \in \mathcal{G} \times \mathcal{W}} \mathcal{L}_{T,\ell}(g, W) - \mathcal{L}^*_{T,\ell} \\
&= \inf_{g \in \mathcal{G}} \big\{ A_\ell(g) - \mathcal{L}^*_{T,\ell} \big\} && (\mathcal{G}, \mathcal{W} \neq \varnothing) \\
&= \inf_{g \in \mathcal{G}} \Big\{ \big(L^\star_\ell(U_g) - \mathcal{L}^*_{T,\ell}\big) + \big(A_\ell(g) - L^\star_\ell(U_g)\big) \Big\} && (L^\star_\ell(U_g) < \infty) \\
&= \inf_{g \in \mathcal{G}} \Big\{ \big(L^\star_\ell(U_g) - \mathcal{L}^*_{T,\ell}\big) + \inf_{W \in \mathcal{W}} \big(\mathcal{L}_{T,\ell}(g, W) - L^\star_\ell(U_g) \big) \Big\}. && (\mathcal{W} \neq \varnothing)
\end{aligned}$$

Finally, the ordering $0 \le \mathcal{L}^{*}_{\mathrm{T},\ell} \le L^{\star}_{\ell}(U_g) \le A_\ell(g) < \infty$ directly certifies $L^{\star}_{\ell}(U_g) - \mathcal{L}^{*}_{\mathrm{T},\ell} \ge 0$ and $\inf_{W \in \mathcal{W}} \big(\mathcal{L}_{\mathrm{T},\ell}(g, W) - L^{\star}_{\ell}(U_g)\big) = A_\ell(g) - L^{\star}_{\ell}(U_g) \ge 0$ for every $g \in \mathcal{G}$. □

Remark 9 (Pointwise conditional surrogate Bayes form). If, in addition to the assumptions of Proposition 6, regular conditional posteriors $\eta^g_y(u) = \mathbb{P}_{P_{\mathrm{T}}}(Y = y \mid U_g = u)$ exist and ℓ satisfies measurable-selection or normal-integrand conditions permitting the interchange of expectation and pointwise infimum [56, 57], then $L^{\star}_{\ell}(U_g)$ admits the pointwise representation

$$L^{\star}_{\ell}(U_g) \;=\; \mathbb{E}_{P_{\mathrm{T}}}\left[\inf_{v \in \mathbb{R}^{K_{\mathrm{T}}}} \sum_{y=1}^{K_{\mathrm{T}}} \eta^g_y(U_g)\, \ell(y, v)\right].$$

This formula interprets $L^{\star}_{\ell}(U_g)$ geometrically as a conditional surrogate Bayes risk, while the proof of Proposition 6 holds without invoking measurable selection.

G Proof of the Hard-Label Source-Conditional Upper Certificate

Proof roadmap. This section proves the population source-conditional upper certificate for the hard-label approximation error (Theorem 10, restating Theorem 4), records its finite-sample Bayes excess-risk consequence when combined with Theorem 7 (Corollary 10), and pinpoints the two inequalities responsible for the one-directional boundary discussed in Remark 3 (Remark 10). For each candidate wrapper $(g, h) \in \mathcal{G} \times \mathcal{H}$, the proof of Theorem 10 upper-bounds target Bayes excess risk by disagreement with the target Bayes selector b_{T}, interpolates the relabeled source Bayes selector $h \circ b_{\mathrm{S}} \circ g$ via the indicator triangle inequality, and transfers the resulting source-classifier disagreement indicator ϕ_h from $P_{\mathrm{S},X}$ to the pushforward target marginal $g_{\#}P_{\mathrm{T},X}$.

Recall from Section 6.2 that for measurable Bayes-optimal classifiers $b_{\mathrm{T}} : \mathcal{X}_{\mathrm{T}} \to [K_{\mathrm{T}}]$ and $b_{\mathrm{S}} : \mathcal{X}_{\mathrm{S}} \to [K_{\mathrm{S}}]$ and any measurable pair $(g, h) \in \mathcal{G} \times \mathcal{H}$,

$$\begin{aligned}
\mathrm{mis}_{\mathrm{hard}}(g, h) &:= \mathbb{P}_{X \sim P_{\mathrm{T},X}}\big(b_{\mathrm{T}}(X) \neq h(b_{\mathrm{S}}(g(X)))\big), \\
\epsilon_{\mathrm{hard}}(h) &:= \mathbb{P}_{U \sim P_{\mathrm{S},X}}\big(h(f_{\mathrm{S}}(U)) \neq h(b_{\mathrm{S}}(U))\big), \\
\phi_h(u) &:= \mathbf{1}\big\{h(f_{\mathrm{S}}(u)) \neq h(b_{\mathrm{S}}(u))\big\}, \\
\mathrm{disc}_{\mathrm{hard}}(g, h) &:= \Big|\mathbb{E}_{X \sim P_{\mathrm{T},X}}\big[\phi_h(g(X))\big] - \mathbb{E}_{U \sim P_{\mathrm{S},X}}\big[\phi_h(U)\big]\Big| \;=\; \Big|\mathbb{E}_{g_{\#}P_{\mathrm{T},X}}[\phi_h] - \mathbb{E}_{P_{\mathrm{S},X}}[\phi_h]\Big|,
\end{aligned}$$

and write the population certificate value as

$$\tau_{\mathrm{hard}} \;:=\; \inf_{(g,h) \in \mathcal{G} \times \mathcal{H}} \big\{\mathrm{mis}_{\mathrm{hard}}(g, h) + \epsilon_{\mathrm{hard}}(h) + \mathrm{disc}_{\mathrm{hard}}(g, h)\big\}.$$

Theorem 10 (Population upper certificate for a_{MR}; restatement of Theorem 4)**.** *Suppose P_{T} and P_{S} admit measurable Bayes-optimal classifiers $b_{\mathrm{T}} : \mathcal{X}_{\mathrm{T}} \to [K_{\mathrm{T}}]$ and $b_{\mathrm{S}} : \mathcal{X}_{\mathrm{S}} \to [K_{\mathrm{S}}]$. Then for every fixed measurable $f_{\mathrm{S}} : \mathcal{X}_{\mathrm{S}} \to [K_{\mathrm{S}}]$, every nonempty class $\mathcal{G}$ of measurable maps $\mathcal{X}_{\mathrm{T}} \to \mathcal{X}_{\mathrm{S}}$, and every nonempty $\mathcal{H} \subseteq [K_{\mathrm{T}}]^{[K_{\mathrm{S}}]}$,*

$$a_{\mathrm{MR}} \;\leq\; \tau_{\mathrm{hard}} \;=\; \inf_{(g,h) \in \mathcal{G} \times \mathcal{H}} \big\{\mathrm{mis}_{\mathrm{hard}}(g, h) + \epsilon_{\mathrm{hard}}(h) + \mathrm{disc}_{\mathrm{hard}}(g, h)\big\}.$$

Proof. Fix any pair $(g, h) \in \mathcal{G} \times \mathcal{H}$, and define the induced hard-label wrapper predictor and its relabeled source-Bayes comparator on $\mathcal{X}_{\mathrm{T}}$ by

$$f_{g,h}(x) \;:=\; h\big(f_{\mathrm{S}}(g(x))\big), \qquad m_{g,h}(x) \;:=\; h\big(b_{\mathrm{S}}(g(x))\big), \qquad x \in \mathcal{X}_{\mathrm{T}}.$$

For $(X, Y) \sim P_{\mathrm{T}}$ and each $y \in [K_{\mathrm{T}}]$, let $\eta_y(X) := \mathbb{E}_{P_{\mathrm{T}}}[\mathbf{1}\{Y = y\} \mid X] \in [0, 1]$ denote a version of the conditional class posterior, so that for any measurable classifier $f : \mathcal{X}_{\mathrm{T}} \to [K_{\mathrm{T}}]$, $\mathbb{E}_{P_{\mathrm{T}}}[\mathbf{1}\{f(X) \neq Y\} \mid X] = 1 - \eta_{f(X)}(X)$ almost surely. Because $0 \leq \eta_y(X) \leq 1$ almost surely for all $y \in [K_{\mathrm{T}}]$, the difference $\eta_{b_{\mathrm{T}}(X)}(X) - \eta_{f(X)}(X)$ vanishes on $\{f(X) = b_{\mathrm{T}}(X)\}$ and is at most 1 almost surely on $\{f(X) \neq b_{\mathrm{T}}(X)\}$, giving $\eta_{b_{\mathrm{T}}(X)}(X) - \eta_{f(X)}(X) \leq \mathbf{1}\{f(X) \neq b_{\mathrm{T}}(X)\}$ almost surely. Using $R_{\mathrm{T}}(b_{\mathrm{T}}) = R^*_{\mathrm{T}}$ and conditioning on X therefore yields

$$\begin{aligned}
R_{\mathrm{T}}(f) - R^*_{\mathrm{T}} &= R_{\mathrm{T}}(f) - R_{\mathrm{T}}(b_{\mathrm{T}}) \\
&= \mathbb{E}_{X \sim P_{\mathrm{T},X}}\big[1 - \eta_{f(X)}(X)\big] - \mathbb{E}_{X \sim P_{\mathrm{T},X}}\big[1 - \eta_{b_{\mathrm{T}}(X)}(X)\big] \\
&= \mathbb{E}_{X \sim P_{\mathrm{T},X}}\big[\eta_{b_{\mathrm{T}}(X)}(X) - \eta_{f(X)}(X)\big] \\
&\leq \mathbb{E}_{X \sim P_{\mathrm{T},X}}\big[\mathbf{1}\{f(X) \neq b_{\mathrm{T}}(X)\}\big] \\
&= \mathbb{P}_{X \sim P_{\mathrm{T},X}}\big(f(X) \neq b_{\mathrm{T}}(X)\big),
\end{aligned} \tag{12}$$

where the second equality applies the tower property of conditional expectation.

Next, for every $x \in \mathcal{X}_{\mathrm{T}}$, if $b_{\mathrm{T}}(x) = m_{g,h}(x)$ and $m_{g,h}(x) = f_{g,h}(x)$, then $f_{g,h}(x) = b_{\mathrm{T}}(x)$; taking the contrapositive gives the pointwise indicator bound $\mathbf{1}\{f_{g,h}(x) \neq b_{\mathrm{T}}(x)\} \leq \mathbf{1}\{b_{\mathrm{T}}(x) \neq m_{g,h}(x)\} + \mathbf{1}\{f_{g,h}(x) \neq m_{g,h}(x)\}$. Noting that $\mathbf{1}\{f_{g,h}(x) \neq m_{g,h}(x)\} = \mathbf{1}\{h(f_{\mathrm{S}}(g(x))) \neq h(b_{\mathrm{S}}(g(x)))\} = \phi_h(g(x))$ and applying Eqn. (12) to $f = f_{g,h}$ yields

$$\begin{aligned}
R_{\mathrm{T}}(f_{g,h}) - R^*_{\mathrm{T}} &\leq \mathbb{P}_{X \sim P_{\mathrm{T},X}}\big(f_{g,h}(X) \neq b_{\mathrm{T}}(X)\big) \\
&\leq \mathbb{E}_{X \sim P_{\mathrm{T},X}}\big[\mathbf{1}\{b_{\mathrm{T}}(X) \neq m_{g,h}(X)\}\big] + \mathbb{E}_{X \sim P_{\mathrm{T},X}}\big[\mathbf{1}\{f_{g,h}(X) \neq m_{g,h}(X)\}\big] \\
&= \mathrm{mis}_{\mathrm{hard}}(g, h) + \mathbb{E}_{X \sim P_{\mathrm{T},X}}\big[\phi_h(g(X))\big],
\end{aligned} \tag{13}$$

where the first inequality invokes Eqn. (12), the second applies the pointwise indicator triangle bound, and the equality substitutes the definitions of $\mathrm{mis}_{\mathrm{hard}}$ and ϕ_h.

To transfer the second term from $g_{\#}P_{\mathrm{T},X}$ to $P_{\mathrm{S},X}$, apply the pushforward change-of-variables identity together with $t \le |t|$:

$$\begin{aligned}\mathbb{E}_{X\sim P_{\mathrm{T},X}}\big[\phi_h(g(X))\big] &= \mathbb{E}_{g_{\#}P_{\mathrm{T},X}}[\phi_h] \\ &= \mathbb{E}_{P_{\mathrm{S},X}}[\phi_h] + \Big(\mathbb{E}_{g_{\#}P_{\mathrm{T},X}}[\phi_h] - \mathbb{E}_{P_{\mathrm{S},X}}[\phi_h]\Big) \\ &\le \mathbb{E}_{P_{\mathrm{S},X}}[\phi_h] + \Big|\mathbb{E}_{g_{\#}P_{\mathrm{T},X}}[\phi_h] - \mathbb{E}_{P_{\mathrm{S},X}}[\phi_h]\Big| \\ &= \epsilon_{\mathrm{hard}}(h) + \mathrm{disc}_{\mathrm{hard}}(g,h), \end{aligned} \tag{14}$$

where the final equality substitutes $\mathbb{E}_{P_{\mathrm{S},X}}[\phi_h] = \epsilon_{\mathrm{hard}}(h)$ and $|\mathbb{E}_{g_{\#}P_{\mathrm{T},X}}[\phi_h] - \mathbb{E}_{P_{\mathrm{S},X}}[\phi_h]| = \mathrm{disc}_{\mathrm{hard}}(g,h)$. Substituting Eqn. (14) into Eqn. (13) gives $R_{\mathrm{T}}(f_{g,h}) - R^*_{\mathrm{T}} \le \mathrm{mis}_{\mathrm{hard}}(g,h) + \epsilon_{\mathrm{hard}}(h) + \mathrm{disc}_{\mathrm{hard}}(g,h)$ for every $(g,h) \in \mathcal{G}\times\mathcal{H}$. Taking the infimum over the nonempty product $\mathcal{G}\times\mathcal{H}$ yields

$$\begin{aligned} a_{\mathrm{MR}} &= R_{\mathrm{T}}^{\mathcal{F}_{\mathrm{MR}},*} - R^*_{\mathrm{T}} && \text{(by Definition 1)} \\ &= \inf_{(g,h)\in\mathcal{G}\times\mathcal{H}} \big\{R_{\mathrm{T}}(f_{g,h}) - R^*_{\mathrm{T}}\big\} && (\mathcal{G},\mathcal{H}\neq\varnothing) \\ &\le \inf_{(g,h)\in\mathcal{G}\times\mathcal{H}} \big\{\mathrm{mis}_{\mathrm{hard}}(g,h) + \epsilon_{\mathrm{hard}}(h) + \mathrm{disc}_{\mathrm{hard}}(g,h)\big\} \;=\; \tau_{\mathrm{hard}}, \end{aligned}$$

which establishes $a_{\mathrm{MR}} \le \tau_{\mathrm{hard}}$. □

Corollary 10 (Combined finite-sample Bayes excess-risk certificate)**.** *Under the hypotheses of Theorems 7 and 10, for every $n_{\mathrm{T}} \ge 1$, every $\delta \in (0,1)$, and every $\eta_n \ge 0$, with probability at least $1-\delta$ over an i.i.d. target sample of size n_{T} drawn from P_{T}, every measurable η_n-approximate ERM $\widehat{f}_n \in \mathcal{F}_{\mathrm{MR}}$ satisfies*

$$R_{\mathrm{T}}(\widehat{f}_n) - R^*_{\mathrm{T}} \;\le\; \tau_{\mathrm{hard}} + \eta_n + 4\sqrt{6}\sqrt{\frac{\log\Pi_{\mathcal{R}}(2n_{\mathrm{T}}) + \log\Pi_{\mathcal{H}|\mathcal{R}}(2n_{\mathrm{T}}) + \log(1/\delta)}{n_{\mathrm{T}}}}.$$

Proof. On the $(1-\delta)$-probability event of Theorem 7 (Remark 5), every measurable η_n-approximate ERM $\widehat{f}_n \in \mathcal{F}_{\mathrm{MR}}$ satisfies

$$\begin{aligned} R_{\mathrm{T}}(\widehat{f}_n) - R^*_{\mathrm{T}} &= \mathcal{E}^{\mathrm{est}}_{\mathrm{T}}(\widehat{f}_n) + a_{\mathrm{MR}} \\ &\le a_{\mathrm{MR}} + \eta_n + 4\sqrt{6}\sqrt{\frac{\log\Pi_{\mathcal{R}}(2n_{\mathrm{T}}) + \log\Pi_{\mathcal{H}|\mathcal{R}}(2n_{\mathrm{T}}) + \log(1/\delta)}{n_{\mathrm{T}}}} \\ &\le \tau_{\mathrm{hard}} + \eta_n + 4\sqrt{6}\sqrt{\frac{\log\Pi_{\mathcal{R}}(2n_{\mathrm{T}}) + \log\Pi_{\mathcal{H}|\mathcal{R}}(2n_{\mathrm{T}}) + \log(1/\delta)}{n_{\mathrm{T}}}}, \end{aligned}$$

where the equality applies Definition 1, the first inequality invokes Theorem 7, and the second substitutes $a_{\mathrm{MR}} \le \tau_{\mathrm{hard}}$ from Theorem 10. □

Remark 10 (Proof-level origin of the two converse obstructions in Remark 3). The proof of Theorem 10 localizes the two inequalities that prevent a converse lower bound in Remark 3. First, Eqn. (12) bounds the pointwise posterior regret $\eta_{b_{\mathrm{T}}(X)}(X) - \eta_{f_{g,h}(X)}(X)$ from above by the binary disagreement indicator $\mathbf{1}\{f_{g,h}(X) \neq b_{\mathrm{T}}(X)\}$. Whenever the target posterior $\eta_\cdot(X)$ is nearly uniform across top competing classes, $f_{g,h}$ can disagree with b_{T} with probability bounded away from zero while incurring arbitrarily small 0–1 excess risk, so reversing Eqn. (12) requires an additional low-noise margin condition on P_{T}. Second, Eqn. (14) transfers ϕ_h between $g_{\#}P_{\mathrm{T},X}$ and $P_{\mathrm{S},X}$ through the pointwise pushforward discrepancy $\mathrm{disc}_{\mathrm{hard}}(g,h) = \big|\mathbb{E}_{g_{\#}P_{\mathrm{T},X}}[\phi_h] - \mathbb{E}_{P_{\mathrm{S},X}}[\phi_h]\big|$ evaluated at the single tested head h. Evaluating ϕ_h at a fixed h makes $\mathrm{disc}_{\mathrm{hard}}(g,h)$ estimable from finite unlabeled samples once b_{S} is plugged in, whereas symmetric domain-adaptation divergences [48] take a supremum over $\mathcal{H}\Delta\mathcal{H}$ and sacrifice direct pointwise estimability.

H Proofs for the Score-Output Certificates and Empirical Corollary

Proof roadmap. This section proves the pointwise population surrogate certificate (Theorem 11, restating Theorem 5) and its simultaneous finite-library empirical counterpart (Corollary 11, restating Corollary 6). We first establish an auxiliary lemma (Lemma 5) proving measurability of the lexicographic plug-in quantities ($\hat{b}_{\mathrm{S}}$, c, and ζ_W) together with the Euclidean ball containment $Ws_{\mathrm{S}}(u), Wc(u) \in B(0, R_{U1})$ and finite range envelopes (B_λ and B_ζ) for all $B_s, B_h, L_\ell \in [0, \infty)$. Theorem 11 invokes Lemma 5 to bound the target surrogate risk via one-hot plug-in interpolation, transfers ζ_W from the pushforward target marginal $\mu_g = g_{\#}P_{\mathrm{T},X}$ to the source marginal $P_{\mathrm{S},X}$, uses $\mathcal{L}^*_{\mathrm{T},\ell} \geq 0$ to upper-bound surrogate excess risk, and applies the nondecreasing multiclass regret transform ψ_ℓ^{-1} to obtain the calibrated 0–1 excess-risk bound. Finally, Corollary 11 combines a deterministic four-term comparison between $\tau_{\mathrm{score}}(g, W)$ and $\widehat{\tau_{\mathrm{score}}}(g, W)$ (Eqn. (15), which accounts for the factor 2 on the source-sample deviation) with upfront zero-envelope checks, two-sided Hoeffding concentration, and a union bound across all $3M$ scalar sample means.

H.1 Auxiliary measurability and envelope lemma

Recall from Appendix B.3 that for measurable $(g, W) \in \mathcal{G} \times \mathcal{W}$, the score-induced source classifier is $\hat{b}_{\mathrm{S}}(u) := \mathrm{lexargmax}_{k \in [K_{\mathrm{S}}]} s_{\mathrm{S},k}(u)$, its one-hot embedding is $c(u) := e_{\hat{b}_{\mathrm{S}}(u)} \in \mathbb{R}^{K_{\mathrm{S}}}$, the pushforward target marginal is $\mu_g := g_{\#}P_{\mathrm{T},X}$, and the population certificate is $\tau_{\mathrm{score}}(g, W) := \mathrm{mis}_{\mathrm{score}}(g, W) + \epsilon_{S,\ell}(W) + \mathrm{disc}_\ell(g, W)$. Before proving Theorem 11 and Corollary 11 (and the uniform Rademacher certificate in Appendix I), we record the measurability and range envelopes of the underlying plug-in and distortion maps across all $B_s, B_h, L_\ell \in [0, \infty)$.

Lemma 5 (Measurability and range envelopes of score-output plug-in quantities)**.** *Let $\mathcal{X}_{\mathrm{S}}$ be a measurable space, let $K_{\mathrm{S}}, K_{\mathrm{T}} \geq 1$ be integers, and let $B_s, B_h, L_\ell \in [0, \infty)$. Suppose $s_{\mathrm{S}} : \mathcal{X}_{\mathrm{S}} \to [-B_s, B_s]^{K_{\mathrm{S}}}$ is measurable, $W \in \mathbb{R}^{K_{\mathrm{T}} \times K_{\mathrm{S}}}$ satisfies $\|W\|_F \leq B_h$, and $\ell : [K_{\mathrm{T}}] \times \mathbb{R}^{K_{\mathrm{T}}} \to [0, \infty)$ is measurable and L_ℓ-Lipschitz in its second argument with respect to $\|\cdot\|_2$ on the closed Euclidean ball $B(0, R_{U1}) \subseteq \mathbb{R}^{K_{\mathrm{T}}}$ of radius $R_{U1} := B_h \max\{B_s\sqrt{K_{\mathrm{S}}}, 1\}$. Set $\hat{b}_{\mathrm{S}}(u) := \mathrm{lexargmax}_{k \in [K_{\mathrm{S}}]} s_{\mathrm{S},k}(u)$, $c(u) := e_{\hat{b}_{\mathrm{S}}(u)} \in \mathbb{R}^{K_{\mathrm{S}}}$, and $\zeta_W(u) := \max_{y \in [K_{\mathrm{T}}]} \big|\ell(y, Ws_{\mathrm{S}}(u)) - \ell(y, Wc(u))\big|$ for $u \in \mathcal{X}_{\mathrm{S}}$, and define*

$$B_\zeta \;:=\; L_\ell B_h\big(B_s\sqrt{K_{\mathrm{S}}} + 1\big), \qquad B_\lambda \;:=\; \sup_{y \in [K_{\mathrm{T}}],\, q \in B(0, R_{U1})} \ell(y, q).$$

Then:

- *(i) $\hat{b}_{\mathrm{S}} : \mathcal{X}_{\mathrm{S}} \to [K_{\mathrm{S}}]$, $c : \mathcal{X}_{\mathrm{S}} \to \mathbb{R}^{K_{\mathrm{S}}}$, and $\zeta_W : \mathcal{X}_{\mathrm{S}} \to [0, \infty)$ are measurable, and $B_\lambda \in [0, \infty)$;*
- *(ii) $Ws_{\mathrm{S}}(u) \in B(0, R_{U1})$ and $Wc(u) \in B(0, R_{U1})$ for every $u \in \mathcal{X}_{\mathrm{S}}$; and*
- *(iii) $\ell(y, Wc(u)) \in [0, B_\lambda]$ and $\zeta_W(u) \in [0, B_\zeta]$ for every $(y, u) \in [K_{\mathrm{T}}] \times \mathcal{X}_{\mathrm{S}}$.*

Proof. For claim (i), deterministic lexicographic tie-breaking (Appendix A) writes each level set of $\hat{b}_{\mathrm{S}}$ as

$$\big\{u \in \mathcal{X}_{\mathrm{S}} : \hat{b}_{\mathrm{S}}(u) = k\big\} \;=\; \bigcap_{r<k} \big\{u \in \mathcal{X}_{\mathrm{S}} : s_{\mathrm{S},k}(u) > s_{\mathrm{S},r}(u)\big\} \cap \bigcap_{r>k} \big\{u \in \mathcal{X}_{\mathrm{S}} : s_{\mathrm{S},k}(u) \geq s_{\mathrm{S},r}(u)\big\}$$

for every $k \in [K_{\mathrm{S}}]$. Because the coordinate functions $s_{\mathrm{S},1}, \ldots, s_{\mathrm{S},K_{\mathrm{S}}}$ are measurable, each level set is a finite intersection of measurable sets, so $\hat{b}_{\mathrm{S}}$ and its one-hot embedding $c(u) = \sum_{k=1}^{K_{\mathrm{S}}} \mathbf{1}\{\hat{b}_{\mathrm{S}}(u) = k\}\, e_k$ are measurable. Composing the continuous linear map $v \mapsto Wv$ and the measurable function $q \mapsto \ell(y, q)$ with s_{S} and c shows that $u \mapsto \ell(y, Ws_{\mathrm{S}}(u))$ and $u \mapsto \ell(y, Wc(u))$ are measurable for every $y \in [K_{\mathrm{T}}]$; taking absolute differences and the pointwise maximum over the finite set $[K_{\mathrm{T}}]$ establishes measurability of ζ_W. Furthermore, for each $y \in [K_{\mathrm{T}}]$, L_ℓ-Lipschitz continuity on $B(0, R_{U1})$ implies continuity of $q \mapsto \ell(y, q)$ on the nonempty compact ball $B(0, R_{U1})$ (including $B(0, 0) = \{0\}$ when $R_{U1} = 0$), so $\sup_{q \in B(0, R_{U1})} \ell(y, q) < \infty$; taking the maximum over the finite alphabet $[K_{\mathrm{T}}]$ and using $\ell \geq 0$ gives $B_\lambda \in [0, \infty)$.

For claim (ii), $s_{\mathrm{S}}(u) \in [-B_s, B_s]^{K_{\mathrm{S}}}$ implies $\|s_{\mathrm{S}}(u)\|_2 = \big(\sum_{k=1}^{K_{\mathrm{S}}} s_{\mathrm{S},k}(u)^2\big)^{1/2} \le B_s\sqrt{K_{\mathrm{S}}}$, while $\|c(u)\|_2 = \|e_{\hat{b}_{\mathrm{S}}(u)}\|_2 = 1$ and $\|W\|_{\mathrm{op}} \le \|W\|_F \le B_h$. For every $u \in \mathcal{X}_{\mathrm{S}}$, submultiplicativity of the operator norm yields

$$\begin{aligned}
\|W s_{\mathrm{S}}(u)\|_2 &\le \|W\|_{\mathrm{op}}\, \|s_{\mathrm{S}}(u)\|_2 && \text{(operator-norm bound)}\\
&\le B_h B_s \sqrt{K_{\mathrm{S}}} && \text{(norm bounds)}\\
&\le B_h \max\big\{B_s\sqrt{K_{\mathrm{S}}},\, 1\big\} \;=\; R_{U1},
\end{aligned}$$

and similarly

$$\begin{aligned}
\|W c(u)\|_2 &\le \|W\|_{\mathrm{op}}\, \|c(u)\|_2 && \text{(operator-norm bound)}\\
&\le B_h && \text{(norm bounds)}\\
&\le B_h \max\big\{B_s\sqrt{K_{\mathrm{S}}},\, 1\big\} \;=\; R_{U1},
\end{aligned}$$

so both $W s_{\mathrm{S}}(u)$ and $W c(u)$ lie in $B(0, R_{U1})$.

For claim (iii), nonnegativity of ℓ and $Wc(u) \in B(0, R_{U1})$ from claim (ii) immediately imply $0 \le \ell(y, Wc(u)) \le B_\lambda$ for every $(y, u) \in [K_{\mathrm{T}}] \times \mathcal{X}_{\mathrm{S}}$. Likewise, because both $W s_{\mathrm{S}}(u)$ and $Wc(u)$ belong to $B(0, R_{U1})$, L_ℓ-Lipschitz continuity of $\ell(y, \cdot)$ on $B(0, R_{U1})$ gives, for every $u \in \mathcal{X}_{\mathrm{S}}$,

$$\begin{aligned}
0 \;\le\; \zeta_W(u) &= \max_{y \in [K_{\mathrm{T}}]} \big|\ell\big(y, W s_{\mathrm{S}}(u)\big) - \ell\big(y, Wc(u)\big)\big|\\
&\le L_\ell \,\big\|W s_{\mathrm{S}}(u) - Wc(u)\big\|_2 \quad (L_\ell\text{-Lipschitz on } B(0, R_{U1}))\\
&\le L_\ell \,\|W\|_{\mathrm{op}}\, \|s_{\mathrm{S}}(u) - c(u)\|_2 \quad \text{(operator-norm bound)}\\
&\le L_\ell \,\|W\|_{\mathrm{op}} \big(\|s_{\mathrm{S}}(u)\|_2 + \|c(u)\|_2\big) \quad \text{(triangle inequality)}\\
&\le L_\ell B_h \big(B_s\sqrt{K_{\mathrm{S}}} + 1\big) \;=\; B_\zeta,
\end{aligned}$$

where the final inequality substitutes $\|W\|_{\mathrm{op}} \le B_h$, $\|s_{\mathrm{S}}(u)\|_2 \le B_s\sqrt{K_{\mathrm{S}}}$, and $\|c(u)\|_2 = 1$. □

H.2 Proof of Theorem 5

Theorem 11 (Source-conditional surrogate certificate; restatement of Theorem 5)**.** *For every target distribution P_{T} over $\mathcal{X}_{\mathrm{T}} \times [K_{\mathrm{T}}]$, every source-input marginal $P_{\mathrm{S},X}$ on $\mathcal{X}_{\mathrm{S}}$, and constants $B_s, B_h, L_\ell \in [0, \infty)$, suppose $s_{\mathrm{S}} : \mathcal{X}_{\mathrm{S}} \to [-B_s, B_s]^{K_{\mathrm{S}}}$ is measurable, every $W \in \mathcal{W}$ satisfies $\|W\|_F \le B_h$, and the measurable surrogate loss $\ell : [K_{\mathrm{T}}] \times \mathbb{R}^{K_{\mathrm{T}}} \to [0, \infty)$ is L_ℓ-Lipschitz in its second argument with respect to $\|\cdot\|_2$ on the closed Euclidean ball $B(0, R_{U1}) \subseteq \mathbb{R}^{K_{\mathrm{T}}}$ of radius $R_{U1} := B_h \max\{B_s\sqrt{K_{\mathrm{S}}}, 1\}$ (`A8–A11` in Table 3). Then for every measurable pair $(g, W) \in \mathcal{G} \times \mathcal{W}$,*

$$\mathcal{L}_{\mathrm{T},\ell}(g, W) - \mathcal{L}^*_{\mathrm{T},\ell} \;\le\; \tau_{\mathrm{score}}(g, W) \;=\; \mathrm{mis}_{\mathrm{score}}(g, W) + \epsilon_{S,\ell}(W) + \mathrm{disc}_\ell(g, W).$$

Consequently, whenever ℓ is multiclass classification-calibrated with nondecreasing regret transform $\psi_\ell^{-1} : [0, \infty) \to [0, \infty)$,

$$R_{\mathrm{T}}(f_{g,W}) - R^*_{\mathrm{T}} \;\le\; \psi_\ell^{-1}\big(\tau_{\mathrm{score}}(g, W)\big) \;=\; \psi_\ell^{-1}\big(\mathrm{mis}_{\mathrm{score}}(g, W) + \epsilon_{S,\ell}(W) + \mathrm{disc}_\ell(g, W)\big).$$

Proof. Fix a measurable pair $(g, W) \in \mathcal{G} \times \mathcal{W}$ and write $\mu_g := g_\# P_{\mathrm{T},X}$. By Lemma 5, $\hat{b}_{\mathrm{S}}$, c, and ζ_W are measurable, $W s_{\mathrm{S}}(u) \in B(0, R_{U1})$, $\ell(y, Wc(u)) \in [0, B_\lambda]$, and $\zeta_W(u) \in [0, B_\zeta]$ for all $(y, u) \in [K_{\mathrm{T}}] \times \mathcal{X}_{\mathrm{S}}$. In particular, $\ell(y, W s_{\mathrm{S}}(u)) \in [0, B_\lambda]$ as well, so $\mathcal{L}_{\mathrm{T},\ell}(g, W)$, $\mathrm{mis}_{\mathrm{score}}(g, W)$, $\mathbb{E}_{\mu_g}[\zeta_W]$, and $\epsilon_{S,\ell}(W) = \mathbb{E}_{P_{\mathrm{S},X}}[\zeta_W]$ are all finite.

By definition of ζ_W, every $(y, u) \in [K_{\mathrm{T}}] \times \mathcal{X}_{\mathrm{S}}$ satisfies $\ell(y, W s_{\mathrm{S}}(u)) \le \ell(y, Wc(u)) + \zeta_W(u)$. Moreover, pointwise nonnegativity $\ell \ge 0$ implies $\mathcal{L}^*_{\mathrm{T},\ell} = \inf_q \mathbb{E}_{(X,Y)\sim P_{\mathrm{T}}}[\ell(Y, q(X))] \ge 0$. Evaluating the plug-in bound at $(Y, g(X))$, taking expectation under $(X, Y) \sim P_{\mathrm{T}}$, and transferring ζ_W

from μ_g to $P_{\mathrm{S},X}$ gives

$$\begin{aligned}
\mathcal{L}_{\mathrm{T},\ell}(g,W) - \mathcal{L}^*_{\mathrm{T},\ell} &\le \mathcal{L}_{\mathrm{T},\ell}(g,W) \\
&= \mathbb{E}_{(X,Y)\sim P_{\mathrm{T}}}\big[\ell\big(Y, W s_{\mathrm{S}}(g(X))\big)\big] \\
&\le \mathbb{E}_{(X,Y)\sim P_{\mathrm{T}}}\big[\ell\big(Y, W c(g(X))\big)\big] + \mathbb{E}_{X\sim P_{\mathrm{T},X}}\big[\zeta_W(g(X))\big] \\
&= \mathrm{mis}_{\mathrm{score}}(g,W) + \mathbb{E}_{\mu_g}[\zeta_W] \\
&= \mathrm{mis}_{\mathrm{score}}(g,W) + \mathbb{E}_{P_{\mathrm{S},X}}[\zeta_W] + \big(\mathbb{E}_{\mu_g}[\zeta_W] - \mathbb{E}_{P_{\mathrm{S},X}}[\zeta_W]\big) \\
&\le \mathrm{mis}_{\mathrm{score}}(g,W) + \mathbb{E}_{P_{\mathrm{S},X}}[\zeta_W] + \big|\mathbb{E}_{\mu_g}[\zeta_W] - \mathbb{E}_{P_{\mathrm{S},X}}[\zeta_W]\big| \\
&= \mathrm{mis}_{\mathrm{score}}(g,W) + \epsilon_{S,\ell}(W) + \mathrm{disc}_\ell(g,W) \;=\; \tau_{\mathrm{score}}(g,W),
\end{aligned}$$

where the first inequality uses $\mathcal{L}^*_{\mathrm{T},\ell} \ge 0$, the second uses $\ell(Y, W s_{\mathrm{S}}(g(X))) \le \ell(Y, W c(g(X))) + \zeta_W(g(X))$, the second equality applies the pushforward identity $\mathbb{E}_{X\sim P_{\mathrm{T},X}}[\zeta_W(g(X))] = \mathbb{E}_{\mu_g}[\zeta_W]$, and the third inequality uses $t \le |t|$.

When ℓ is multiclass classification-calibrated [56, 57], the score-induced predictor $f_{g,W}(x) = \mathrm{lexargmax}_{y\in[K_{\mathrm{T}}]}(W s_{\mathrm{S}}(g(x)))_y$ satisfies $R_{\mathrm{T}}(f_{g,W}) - R^*_{\mathrm{T}} \le \psi_\ell^{-1}\big(\mathcal{L}_{\mathrm{T},\ell}(g,W) - \mathcal{L}^*_{\mathrm{T},\ell}\big)$. Since $\psi_\ell^{-1} : [0,\infty) \to [0,\infty)$ is nondecreasing and $0 \le \mathcal{L}_{\mathrm{T},\ell}(g,W) - \mathcal{L}^*_{\mathrm{T},\ell} \le \tau_{\mathrm{score}}(g,W)$, applying ψ_ℓ^{-1} yields $R_{\mathrm{T}}(f_{g,W}) - R^*_{\mathrm{T}} \le \psi_\ell^{-1}\big(\tau_{\mathrm{score}}(g,W)\big)$. □

Remark 11 (One-directional nature and family-level infimum). Because the proof of Theorem 11 lower-bounds $\mathcal{L}^*_{\mathrm{T},\ell}$ by 0 and bounds surrogate differences via the worst-label envelope ζ_W, Theorem 11 is a one-directional upper certificate rather than a two-sided characterization. For nonempty $\mathcal{G}$ and $\mathcal{W}$, setting $\tau_* := \inf_{(g,W)\in\mathcal{G}\times\mathcal{W}} \tau_{\mathrm{score}}(g,W)$ yields $a_{\mathrm{score},\ell} \le \tau_*$ and $a_{\mathrm{score},0/1} \le \inf_{\varepsilon>0} \psi_\ell^{-1}(\tau_* + \varepsilon)$; this right-hand limit equals $\psi_\ell^{-1}(\tau_*)$ when ψ_ℓ^{-1} is right-continuous at τ_*, while $a_{\mathrm{score},0/1} \le \psi_\ell^{-1}(\tau_*)$ also follows directly whenever τ_* is attained at some $(g^\star, W^\star) \in \mathcal{G}\times\mathcal{W}$ by evaluating Theorem 11 at $(g^\star, W^\star)$.

H.3 Proof of Corollary 6

Corollary 11 (Finite-library empirical certificate; restatement of Corollary 6)**.** *Under the assumptions of Theorem 11, set $B_\zeta := L_\ell B_h(B_s\sqrt{K_{\mathrm{S}}}+1)$ and $B_\lambda := \sup_{y\in[K_{\mathrm{T}}],\, q\in B(0,R_{U1})} \ell(y,q) < \infty$. For every integer $M \ge 1$, every finite candidate library $Q = \{(g_r, W_r)\}_{r=1}^M \subseteq \mathcal{G}\times\mathcal{W}$ of measurable pairs, all sample sizes $n_{\mathrm{T}}, n_{\mathrm{S}} \ge 1$, and every $\delta \in (0,1)$, if $D_{\mathrm{T}} \sim P_{\mathrm{T}}^{n_{\mathrm{T}}}$ and $D_{\mathrm{S}} \sim P_{\mathrm{S},X}^{n_{\mathrm{S}}}$ are mutually independent i.i.d. samples drawn independently of Q, then with probability at least $1-\delta$ over $(D_{\mathrm{T}}, D_{\mathrm{S}})$, simultaneously for all $(g,W) \in Q$,*

$$\mathcal{L}_{\mathrm{T},\ell}(g,W) - \mathcal{L}^*_{\mathrm{T},\ell} \;\le\; \widehat{\tau_{\mathrm{score}}}(g,W) + B_\lambda\sqrt{\frac{\log(6M/\delta)}{2n_{\mathrm{T}}}} + B_\zeta\sqrt{\frac{\log(6M/\delta)}{2n_{\mathrm{T}}}} + 2B_\zeta\sqrt{\frac{\log(6M/\delta)}{2n_{\mathrm{S}}}}.$$

Consequently, whenever ℓ is multiclass classification-calibrated with nondecreasing regret transform $\psi_\ell^{-1} : [0,\infty) \to [0,\infty)$, on the same $(1-\delta)$-probability event, simultaneously for all $(g,W) \in Q$,

$$R_{\mathrm{T}}(f_{g,W}) - R^*_{\mathrm{T}} \;\le\; \psi_\ell^{-1}\left(\widehat{\tau_{\mathrm{score}}}(g,W) + (B_\lambda + B_\zeta)\sqrt{\frac{\log(6M/\delta)}{2n_{\mathrm{T}}}} + 2B_\zeta\sqrt{\frac{\log(6M/\delta)}{2n_{\mathrm{S}}}}\right).$$

Proof. For any candidate $(g,W) \in Q$, abbreviate its three population expectations and empirical sample means from Appendix B.4 by

$$\begin{aligned}
&\lambda := \mathrm{mis}_{\mathrm{score}}(g,W), && T := \mathbb{E}_{X\sim P_{\mathrm{T},X}}\big[\zeta_W(g(X))\big], && S := \mathbb{E}_{U\sim P_{\mathrm{S},X}}\big[\zeta_W(U)\big], \\
&\widehat{\lambda} := \widehat{\mathrm{mis}_{\mathrm{score}}}(g,W), && \widehat{T} := \widehat{T}(g,W), && \widehat{S} := \widehat{S}(W),
\end{aligned}$$

so that $\tau_{\mathrm{score}}(g,W) = \lambda + S + |T-S|$ and $\widehat{\tau_{\mathrm{score}}}(g,W) = \widehat{\lambda} + \widehat{S} + |\widehat{T} - \widehat{S}|$. Bounding the difference deterministically via the scalar triangle inequality gives

$$\begin{aligned}
\tau_{\mathrm{score}}(g,W) - \widehat{\tau_{\mathrm{score}}}(g,W) &= (\lambda - \widehat{\lambda}) + (S - \widehat{S}) + \big(|T-S| - |\widehat{T} - \widehat{S}|\big) \\
&\le |\lambda - \widehat{\lambda}| + |S - \widehat{S}| + \big|(T-S) - (\widehat{T} - \widehat{S})\big| \\
&= |\lambda - \widehat{\lambda}| + |S - \widehat{S}| + \big|(T - \widehat{T}) - (S - \widehat{S})\big| \\
&\le |\lambda - \widehat{\lambda}| + |T - \widehat{T}| + 2|S - \widehat{S}|,
\end{aligned}$$

where the first inequality uses $u \le |u|$ and $||a| - |b|| \le |a - b|$, and the second applies the triangle inequality, rearranging to the deterministic certificate comparison

$$\tau_{\mathrm{score}}(g,W) \;\le\; \widehat{\tau_{\mathrm{score}}}(g,W) + |\lambda - \widehat{\lambda}| + |T - \widehat{T}| + 2|S - \widehat{S}|, \tag{15}$$

where the factor 2 on $|S - \widehat{S}|$ arises because $\widehat{S}$ estimates both $\epsilon_{S,\ell}(W) = S$ and the source-side reference inside $\mathrm{disc}_\ell(g,W) = |T - S|$.

Now condition on Q if Q is constructed on an independent holdout split, so that every $(g_r, W_r) \in Q$ ($r \in [M]$) is fixed and independent of $(D_{\mathrm{T}}, D_{\mathrm{S}})$. By Lemma 5, for each fixed $(g_r, W_r) \in Q$, the n_{T} summands $\ell(Y_i, W_r c(g_r(X_i)))$ in $\widehat{\mathrm{mis}_{\mathrm{score}}}(g_r, W_r)$ are i.i.d. in $[0, B_\lambda]$, the n_{T} summands $\zeta_{W_r}(g_r(X_i))$ in $\widehat{T}(g_r, W_r)$ are i.i.d. in $[0, B_\zeta]$, and the n_{S} summands $\zeta_{W_r}(U_j)$ in $\widehat{S}(W_r)$ are i.i.d. in $[0, B_\zeta]$. For any integer $n \ge 1$ and envelope $B \ge 0$, set $t_n(B) := B\sqrt{\log(6M/\delta)/(2n)}$, and let $V_1, \dots, V_n \in [0, B]$ be i.i.d. with mean $\mu = \mathbb{E}[V_1]$ and empirical mean $\widehat{\mu} = n^{-1}\sum_{i=1}^n V_i$. In the degenerate zero-envelope case $B = 0$ (namely $B_\lambda = 0$ for $\widehat{\mathrm{mis}_{\mathrm{score}}}$, or $B_\zeta = 0$ for $\widehat{T}$ and $\widehat{S}$), every V_i vanishes identically, so $|\mu - \widehat{\mu}| = 0$ almost surely and $\mathbb{P}(|\mu - \widehat{\mu}| > t_n(0)) = \mathbb{P}(0 > 0) = 0 \le \delta/(3M)$. When $B > 0$, two-sided Hoeffding's inequality on $[0, B]$ yields

$$\begin{aligned}
\mathbb{P}\big(|\mu - \widehat{\mu}| > t_n(B)\big) &\le 2\exp\!\left(-\frac{2n\, t_n(B)^2}{B^2}\right) && \text{(Hoeffding's inequality)} \\
&= 2\exp\big(-\log(6M/\delta)\big) \;=\; \frac{\delta}{3M}.
\end{aligned}$$

Writing $T(g_r, W_r) := \mathbb{E}_{X \sim P_{\mathrm{T},X}}[\zeta_{W_r}(g_r(X))]$ and $S(W_r) := \mathbb{E}_{U \sim P_{\mathrm{S},X}}[\zeta_{W_r}(U)]$, define for each $r \in [M]$ the three deviation events

$$\begin{aligned}
E_{r,1} &:= \big\{\big|\mathrm{mis}_{\mathrm{score}}(g_r, W_r) - \widehat{\mathrm{mis}_{\mathrm{score}}}(g_r, W_r)\big| > t_{n_{\mathrm{T}}}(B_\lambda)\big\}, \\
E_{r,2} &:= \big\{\big|T(g_r, W_r) - \widehat{T}(g_r, W_r)\big| > t_{n_{\mathrm{T}}}(B_\zeta)\big\}, \qquad E_{r,3} := \big\{\big|S(W_r) - \widehat{S}(W_r)\big| > t_{n_{\mathrm{S}}}(B_\zeta)\big\}.
\end{aligned}$$

In both zero- and positive-envelope regimes, $\mathbb{P}(E_{r,j}) \le \delta/(3M)$ for all $r \in [M]$ and $j \in \{1,2,3\}$, so a union bound across all $3M$ scalar means gives

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{r=1}^M \big(E_{r,1} \cup E_{r,2} \cup E_{r,3}\big)\right) &\le \sum_{r=1}^M \Big(\mathbb{P}(E_{r,1}) + \mathbb{P}(E_{r,2}) + \mathbb{P}(E_{r,3})\Big) && \text{(union bound)} \\
&\le \sum_{r=1}^M 3 \cdot \frac{\delta}{3M} \;=\; \delta.
\end{aligned}$$

On the complementary event of probability at least $1 - \delta$, combining Theorem 11 with Eqn. (15) simultaneously for every $(g, W) \in Q$ gives

$$\begin{aligned}
\mathcal{L}_{\mathrm{T},\ell}(g,W) - \mathcal{L}^*_{\mathrm{T},\ell} &\le \tau_{\mathrm{score}}(g,W) \quad \text{(by Theorem 11)} \\
&\le \widehat{\tau_{\mathrm{score}}}(g,W) + \big|\mathrm{mis}_{\mathrm{score}}(g,W) - \widehat{\mathrm{mis}_{\mathrm{score}}}(g,W)\big| \\
&\quad + \big|T(g,W) - \widehat{T}(g,W)\big| + 2\big|S(W) - \widehat{S}(W)\big| \quad \text{(by Eqn. (15))} \\
&\le \widehat{\tau_{\mathrm{score}}}(g,W) + t_{n_{\mathrm{T}}}(B_\lambda) + t_{n_{\mathrm{T}}}(B_\zeta) + 2\, t_{n_{\mathrm{S}}}(B_\zeta),
\end{aligned}$$

which establishes the simultaneous surrogate certificate. When ℓ is multiclass classification-calibrated, applying the calibrated excess-risk bound of Theorem 11 together with monotonicity of $\psi_\ell^{-1} : [0,\infty) \to [0,\infty)$ on the same $(1-\delta)$-probability event yields the simultaneous 0–1 excess-risk bound. □

Remark 12 (Evaluation-independence requirement). Simultaneity in Corollary 11 holds over the evaluation-independent library Q; whenever candidate wrappers are trained or selected using data, D_{T} and D_{S} must be held out from that selection step (for example via sample splitting or cross-fitting).

I Uniform Empirical Score-Output Certificate over a Fixed Candidate Class

Proof roadmap. This section proves the uniform empirical score-output certificate over a fixed (possibly infinite) candidate set (Theorem 12, restating Theorem 6), replacing the $\log(6M/\delta)$ finite-library union bound of Corollary 11 with empirical Rademacher complexities of three induced scalar classes. Appendix I.1 records the pointwise-measurability, range-envelope, and evaluation-independence conditions on $Q \subseteq \mathcal{G}\times\mathcal{W}$ and $(D_{\mathrm{T}}, D_{\mathrm{S}})$. Appendix I.2 first establishes a self-contained two-sided empirical Rademacher uniform-deviation lemma (Lemma 6), then applies it at failure parameter $\delta' = \delta/6$ across $(\mathcal{L}_Q, \mathcal{Z}_{\mathrm{T},Q}, \mathcal{Z}_{\mathrm{S},Q})$ together with Lemma 5, the deterministic comparison Eqn. (15), and Theorem 11 to prove Theorem 12, and concludes with optional vector-contraction refinements via Maurer [55].

Let $Q \subseteq \mathcal{G} \times \mathcal{W}$ be a nonempty set of measurable candidate pairs fixed independently of the certification samples (or constructed on an independent holdout split so that it is independent of the certification folds), and denote its head projection by

$$\mathcal{W}_Q \;:=\; \big\{W \in \mathcal{W} : \exists\, g \in \mathcal{G} \text{ with } (g, W) \in Q\big\}.$$

Let $D_{\mathrm{T}} = \{(X_i, Y_i)\}_{i=1}^{n_{\mathrm{T}}} \sim P_{\mathrm{T}}^{n_{\mathrm{T}}}$ (with target-input tuple $X_{1:n_{\mathrm{T}}} := (X_1, \dots, X_{n_{\mathrm{T}}})$) and $D_{\mathrm{S}} = \{U_j\}_{j=1}^{n_{\mathrm{S}}} \sim P_{\mathrm{S},X}^{n_{\mathrm{S}}}$ $(n_{\mathrm{T}}, n_{\mathrm{S}} \geq 1)$ be mutually independent i.i.d. samples drawn independently of Q. As in Appendix B.4, define the three induced scalar function classes

$$\begin{aligned}
\mathcal{L}_Q &:= \big\{(x, y) \mapsto \ell\big(y, Wc(g(x))\big) : (g, W) \in Q\big\},\\
\mathcal{Z}_{\mathrm{T},Q} &:= \big\{x \mapsto \zeta_W\big(g(x)\big) : (g, W) \in Q\big\},\\
\mathcal{Z}_{\mathrm{S},Q} &:= \big\{u \mapsto \zeta_W(u) : (g, W) \in Q\big\} \;=\; \big\{u \mapsto \zeta_W(u) : W \in \mathcal{W}_Q\big\},
\end{aligned}$$

where $\hat{b}_{\mathrm{S}}(u) := \mathrm{lexargmax}_{k\in[K_{\mathrm{S}}]}\, s_{\mathrm{S},k}(u)$ and $c(u) := e_{\hat{b}_{\mathrm{S}}(u)}$. For a scalar function class $\mathcal{F}$ and sample $Z_{1:n} = (Z_1, \dots, Z_n)$, write its empirical Rademacher complexity as

$$\widehat{\mathfrak{R}}_{Z_{1:n}}(\mathcal{F}) \;:=\; \frac{1}{n}\,\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}} \sum_{i=1}^{n} \sigma_i f(Z_i) \;\middle|\; Z_{1:n}\right],$$

where $\sigma_1, \dots, \sigma_n \in \{-1, +1\}$ are independent uniform Rademacher signs independent of $Z_{1:n}$.

I.1 Assumptions for the uniform empirical certificate

Assume $\mathcal{L}_Q$, $\mathcal{Z}_{\mathrm{T},Q}$, and $\mathcal{Z}_{\mathrm{S},Q}$ are pointwise measurable (i.e., admitting a countable subset dense under pointwise convergence, as in Appendix A) so that their empirical suprema and Rademacher averages are measurable and symmetrization applies. Assume the boundedness and surrogate-regularity hypotheses of Theorem 11 (Theorem 5): for constants $B_s, B_h, L_\ell \in [0, \infty)$, $s_{\mathrm{S}} : \mathcal{X}_{\mathrm{S}} \to [-B_s, B_s]^{K_{\mathrm{S}}}$ is measurable, every $W \in \mathcal{W}$ (and hence every $W \in \mathcal{W}_Q$) satisfies $\|W\|_F \leq B_h$, and $\ell : [K_{\mathrm{T}}] \times \mathbb{R}^{K_{\mathrm{T}}} \to [0, \infty)$ is measurable and L_ℓ-Lipschitz in its second argument with respect to $\|\cdot\|_2$ on $B(0, R_{U1})$ with $R_{U1} := B_h \max\{B_s\sqrt{K_{\mathrm{S}}}, 1\}$, giving the finite envelopes $B_\lambda := \sup_{y\in[K_{\mathrm{T}}],\, q\in B(0,R_{U1})} \ell(y, q)$ and $B_\zeta := L_\ell B_h(B_s\sqrt{K_{\mathrm{S}}} + 1)$ by Lemma 5. Assume $(D_{\mathrm{T}}, D_{\mathrm{S}})$ and Q satisfy the mutual- and evaluation-independence conditions stated at the start of Appendix I.

I.2 Uniform empirical certificate

We first establish the two-sided empirical Rademacher concentration inequality invoked for each of the three induced scalar classes.

Lemma 6 (Two-sided empirical Rademacher uniform deviation)**.** *Let $\mathcal{Z}$ be a measurable space, let P be a probability distribution on $\mathcal{Z}$, let $n \geq 1$ be an integer, and let $B \in [0, \infty)$. Let $\mathcal{F}$ be a nonempty pointwise-measurable class of functions $f : \mathcal{Z} \to [0, B]$, and let $Z_{1:n} = (Z_1, \dots, Z_n) \sim P^n$ be an i.i.d. sample with $\mathbb{E}[f] := \mathbb{E}_{Z\sim P}[f(Z)]$ and $\widehat{\mathbb{E}}_{Z_{1:n}}[f] := n^{-1}\sum_{i=1}^n f(Z_i)$. Then for every $\delta' \in (0, 1/2)$,*

$$\mathbb{P}_{Z_{1:n}\sim P^n}\left(\sup_{f\in\mathcal{F}}\big|\mathbb{E}[f] - \widehat{\mathbb{E}}_{Z_{1:n}}[f]\big| \;>\; 2\,\widehat{\mathfrak{R}}_{Z_{1:n}}(\mathcal{F}) + 3B\sqrt{\frac{\log(2/\delta')}{2n}}\right) \;\leq\; 2\delta'.$$

Proof. If $B = 0$, then every $f \in \mathcal{F}$ vanishes identically on $\mathcal{Z}$, so $\mathbb{E}[f] = \widehat{\mathbb{E}}_{Z_{1:n}}[f] = 0$ for all $f \in \mathcal{F}$ and $\widehat{\mathfrak{R}}_{Z_{1:n}}(\mathcal{F}) = 0$, whence the probability of exceeding 0 is $0 \le 2\delta'$. Assume henceforth that $B > 0$, and set $t := B\sqrt{\log(2/\delta')/(2n)}$ and $\mathfrak{R}_n(\mathcal{F}) := \mathbb{E}_{Z_{1:n}\sim P^n}\big[\widehat{\mathfrak{R}}_{Z_{1:n}}(\mathcal{F})\big]$.

Define the upper one-sided supremum $\Phi_+(Z_{1:n}) := \sup_{f\in\mathcal{F}}\big(\mathbb{E}[f] - \widehat{\mathbb{E}}_{Z_{1:n}}[f]\big)$, which is measurable along with $\widehat{\mathfrak{R}}_{Z_{1:n}}(\mathcal{F})$ by pointwise measurability of $\mathcal{F}$ [24, 54]. For any two tuples $z_{1:n}, z'_{1:n} \in \mathcal{Z}^n$ differing only at coordinate $i \in [n]$, range containment $f(z_i), f(z'_i) \in [0, B]$ implies $|f(z_i) - f(z'_i)| \le B$, so replacing z_i by z'_i changes $\Phi_+(z_{1:n})$ by at most $\sup_{f\in\mathcal{F}} n^{-1}|f(z'_i) - f(z_i)| \le B/n$ and changes $\widehat{\mathfrak{R}}_{z_{1:n}}(\mathcal{F})$ by at most $\mathbb{E}_\sigma\big[\sup_{f\in\mathcal{F}} n^{-1}|\sigma_i(f(z_i) - f(z'_i))|\big] \le B/n$. Because $\sum_{i=1}^n (B/n)^2 = B^2/n$ and $\exp(-2nt^2/B^2) = \delta'/2$, McDiarmid's bounded-differences inequality yields $\mathbb{P}\big(\Phi_+(Z_{1:n}) - \mathbb{E}[\Phi_+(Z_{1:n})] > t\big) \le \delta'/2$ and $\mathbb{P}\big(\mathfrak{R}_n(\mathcal{F}) - \widehat{\mathfrak{R}}_{Z_{1:n}}(\mathcal{F}) > t\big) \le \delta'/2$.

Introducing an independent ghost sample $Z'_{1:n} = (Z'_1, \dots, Z'_n) \sim P^n$ and independent uniform signs $\sigma_1, \dots, \sigma_n \in \{-1, +1\}$, standard Rademacher symmetrization bounds $\mathbb{E}[\Phi_+(Z_{1:n})]$ via

$$
\begin{aligned}
\mathbb{E}_{Z_{1:n}}\big[\Phi_+(Z_{1:n})\big] &= \mathbb{E}_{Z_{1:n}}\left[\sup_{f\in\mathcal{F}}\Big(\mathbb{E}_{Z'_{1:n}}\big[\widehat{\mathbb{E}}_{Z'_{1:n}}[f]\big] - \widehat{\mathbb{E}}_{Z_{1:n}}[f]\Big)\right] \\
&\le \mathbb{E}_{Z_{1:n},Z'_{1:n}}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n\big(f(Z'_i) - f(Z_i)\big)\right] \quad \text{(Jensen's inequality)} \\
&= \mathbb{E}_{Z_{1:n},Z'_{1:n},\sigma}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n\sigma_i\big(f(Z'_i) - f(Z_i)\big)\right] \quad \text{(exchangeability)} \\
&\le 2\,\mathbb{E}_{Z_{1:n},\sigma}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n\sigma_i f(Z_i)\right] \;=\; 2\,\mathfrak{R}_n(\mathcal{F}), \quad \text{(subadditivity of sup)}
\end{aligned}
$$

where the final inequality splits the supremum and uses $Z'_{1:n} \overset{d}{=} Z_{1:n}$ and $-\sigma \overset{d}{=} \sigma$. On the intersection of the two McDiarmid events (which has failure probability at most $\delta'/2 + \delta'/2 = \delta'$ by the union bound), combining these three bounds gives

$$
\begin{aligned}
\Phi_+(Z_{1:n}) &\le \mathbb{E}_{Z_{1:n}}\big[\Phi_+(Z_{1:n})\big] + B\sqrt{\frac{\log(2/\delta')}{2n}} && \text{(McDiarmid on } \Phi_+\text{)} \\
&\le 2\,\mathfrak{R}_n(\mathcal{F}) + B\sqrt{\frac{\log(2/\delta')}{2n}} && \text{(symmetrization)} \\
&\le 2\,\widehat{\mathfrak{R}}_{Z_{1:n}}(\mathcal{F}) + 3B\sqrt{\frac{\log(2/\delta')}{2n}}, && \text{(McDiarmid on } \widehat{\mathfrak{R}}_{Z_{1:n}}\text{)}
\end{aligned}
$$

where the third step substitutes $2\,\mathfrak{R}_n(\mathcal{F}) \le 2\,\widehat{\mathfrak{R}}_{Z_{1:n}}(\mathcal{F}) + 2t$.

For the lower tail $\Phi_-(Z_{1:n}) := \sup_{f\in\mathcal{F}}\big(\widehat{\mathbb{E}}_{Z_{1:n}}[f] - \mathbb{E}[f]\big) = \sup_{h\in-\mathcal{F}}\big(\mathbb{E}[h] - \widehat{\mathbb{E}}_{Z_{1:n}}[h]\big)$, every $h = -f \in -\mathcal{F}$ takes values in $[-B, 0]$ of diameter B (equivalently, $B - f \in [0, B]$ has $\mathbb{E}[B - f] - \widehat{\mathbb{E}}_{Z_{1:n}}[B - f] = \widehat{\mathbb{E}}_{Z_{1:n}}[f] - \mathbb{E}[f]$), so the B/n bounded-differences property and symmetrization apply identically, while sign symmetry $-\sigma \overset{d}{=} \sigma$ gives $\widehat{\mathfrak{R}}_{Z_{1:n}}(-\mathcal{F}) = \widehat{\mathfrak{R}}_{Z_{1:n}}(\mathcal{F})$. Hence $\mathbb{P}\big(\Phi_-(Z_{1:n}) > 2\,\widehat{\mathfrak{R}}_{Z_{1:n}}(\mathcal{F}) + 3t\big) \le \delta'$ as well. Since $\sup_{f\in\mathcal{F}}|\mathbb{E}[f] - \widehat{\mathbb{E}}_{Z_{1:n}}[f]| = \max\{\Phi_+(Z_{1:n}), \Phi_-(Z_{1:n})\}$, a union bound over the two tails yields total failure probability at most $\delta' + \delta' = 2\delta'$. □

The next theorem proves Theorem 6, upper-bounding the surrogate excess risk $\mathcal{L}_{\mathrm{T},\ell}(g, W) - \mathcal{L}^*_{\mathrm{T},\ell}$ (and, when ℓ is calibrated, the 0–1 excess risk $R_{\mathrm{T}}(f_{g,W}) - R^*_{\mathrm{T}}$) simultaneously over $(g, W) \in Q$ via the empirical plug-in certificate $\widehat{\tau_{\mathrm{score}}}(g, W)$ and the concentration slack Γ_Q.

Theorem 12 (Uniform empirical certificate over a fixed candidate set; restatement of Theorem 6)**.** *Under the assumptions of Theorem 11 (Theorem 5), let B_λ, B_ζ be defined as in Corollary 11 (Corollary 6), and let $Q \subseteq \mathcal{G} \times \mathcal{W}$ be a nonempty set of measurable candidate pairs whose induced scalar classes $\mathcal{L}_Q$, $\mathcal{Z}_{\mathrm{T},Q}$, and $\mathcal{Z}_{\mathrm{S},Q}$ satisfy the measurability conditions of Appendix I.1. For all sample sizes $n_\mathrm{T}, n_\mathrm{S} \geq 1$ and every $\delta \in (0,1)$, if $D_\mathrm{T} = \{(X_i, Y_i)\}_{i=1}^{n_\mathrm{T}} \sim P_\mathrm{T}^{n_\mathrm{T}}$ (with input tuple $X_{1:n_\mathrm{T}}$) and $D_\mathrm{S} = \{U_j\}_{j=1}^{n_\mathrm{S}} \sim P_{\mathrm{S},X}^{n_\mathrm{S}}$ are mutually independent i.i.d. samples drawn independently of Q, then with probability at least $1-\delta$ over $(D_\mathrm{T}, D_\mathrm{S})$, simultaneously for all $(g, W) \in Q$,*

$$\mathcal{L}_{\mathrm{T},\ell}(g,W) - \mathcal{L}^*_{\mathrm{T},\ell} \;\leq\; \widehat{\tau_{\mathrm{score}}}(g,W) + \Gamma_Q, \tag{16}$$

where the uniform concentration slack Γ_Q is

$$\begin{aligned}\Gamma_Q \;:=\; & 2\widehat{\mathfrak{R}}_{D_\mathrm{T}}(\mathcal{L}_Q) + 3B_\lambda\sqrt{\frac{\log(12/\delta)}{2n_\mathrm{T}}} \\ & + 2\widehat{\mathfrak{R}}_{X_{1:n_\mathrm{T}}}(\mathcal{Z}_{\mathrm{T},Q}) + 3B_\zeta\sqrt{\frac{\log(12/\delta)}{2n_\mathrm{T}}} \\ & + 4\widehat{\mathfrak{R}}_{D_\mathrm{S}}(\mathcal{Z}_{\mathrm{S},Q}) + 6B_\zeta\sqrt{\frac{\log(12/\delta)}{2n_\mathrm{S}}}.\end{aligned}$$

Consequently, whenever ℓ is multiclass classification-calibrated with nondecreasing regret transform $\psi_\ell^{-1} : [0,\infty) \to [0,\infty)$, on the same $(1-\delta)$-probability event, simultaneously for all $(g, W) \in Q$,

$$R_\mathrm{T}(f_{g,W}) - R^*_\mathrm{T} \;\leq\; \psi_\ell^{-1}\big(\widehat{\tau_{\mathrm{score}}}(g,W) + \Gamma_Q\big). \tag{17}$$

Proof. Condition on Q whenever Q is constructed on an independent holdout split, so that Q is fixed and independent of $(D_\mathrm{T}, D_\mathrm{S})$. Set the three component slacks

$$\begin{aligned}\Delta_\lambda &:= 2\widehat{\mathfrak{R}}_{D_\mathrm{T}}(\mathcal{L}_Q) + 3B_\lambda\sqrt{\frac{\log(12/\delta)}{2n_\mathrm{T}}}, \\ \Delta_\mathrm{T} &:= 2\widehat{\mathfrak{R}}_{X_{1:n_\mathrm{T}}}(\mathcal{Z}_{\mathrm{T},Q}) + 3B_\zeta\sqrt{\frac{\log(12/\delta)}{2n_\mathrm{T}}}, \\ \Delta_\mathrm{S} &:= 2\widehat{\mathfrak{R}}_{D_\mathrm{S}}(\mathcal{Z}_{\mathrm{S},Q}) + 3B_\zeta\sqrt{\frac{\log(12/\delta)}{2n_\mathrm{S}}},\end{aligned}$$

so that $\Gamma_Q = \Delta_\lambda + \Delta_\mathrm{T} + 2\Delta_\mathrm{S}$.

By Lemma 5, every function in $\mathcal{L}_Q$ is measurable and takes values in $[0, B_\lambda]$, while every function in $\mathcal{Z}_{\mathrm{T},Q}$ and $\mathcal{Z}_{\mathrm{S},Q}$ is measurable and takes values in $[0, B_\zeta]$. Since $\delta \in (0,1)$, setting $\delta' := \delta/6 \in (0, 1/6) \subset (0, 1/2)$ gives $2/\delta' = 12/\delta$ and $2\delta' = \delta/3$. Applying Lemma 6 at $\delta' = \delta/6$ separately to $\mathcal{L}_Q$ on $D_\mathrm{T} \sim P_\mathrm{T}^{n_\mathrm{T}}$ (envelope B_λ), to $\mathcal{Z}_{\mathrm{T},Q}$ on $X_{1:n_\mathrm{T}} \sim P_{\mathrm{T},X}^{n_\mathrm{T}}$ (envelope B_ζ), and to $\mathcal{Z}_{\mathrm{S},Q}$ on $D_\mathrm{S} \sim P_{\mathrm{S},X}^{n_\mathrm{S}}$ (envelope B_ζ) yields three bad events of probability at most $\delta/3$ each. By the union bound, their complements hold simultaneously with probability at least $1 - 3(\delta/3) = 1 - \delta$, on which

$$\begin{aligned}\sup_{(g,W)\in Q}\big|\mathrm{mis}_{\mathrm{score}}(g,W) - \widehat{\mathrm{mis}_{\mathrm{score}}}(g,W)\big| &\leq \Delta_\lambda, \\ \sup_{(g,W)\in Q}\big|T(g,W) - \widehat{T}(g,W)\big| &\leq \Delta_\mathrm{T}, \qquad T(g,W) := \mathbb{E}_{X\sim P_{\mathrm{T},X}}\big[\zeta_W(g(X))\big], \\ \sup_{W\in\mathcal{W}_Q}\big|S(W) - \widehat{S}(W)\big| &\leq \Delta_\mathrm{S}, \qquad S(W) := \mathbb{E}_{U\sim P_{\mathrm{S},X}}\big[\zeta_W(U)\big].\end{aligned}$$

On this $(1-\delta)$-probability event, combining Theorem 11 with the deterministic comparison Eqn. (15) simultaneously for every $(g, W) \in Q$ (noting $W \in \mathcal{W}_Q$) yields

$$\begin{aligned}
\mathcal{L}_{\mathrm{T},\ell}(g,W) - \mathcal{L}^*_{\mathrm{T},\ell} &\le \tau_{\mathrm{score}}(g,W) \quad \text{(by Theorem 11)} \\
&\le \widehat{\tau_{\mathrm{score}}}(g,W) + \left|\mathrm{mis}_{\mathrm{score}}(g,W) - \widehat{\mathrm{mis}_{\mathrm{score}}}(g,W)\right| \\
&\quad + \left|T(g,W) - \widehat{T}(g,W)\right| + 2\left|S(W) - \widehat{S}(W)\right| \quad \text{(by Eqn. (15))} \\
&\le \widehat{\tau_{\mathrm{score}}}(g,W) + \Delta_\lambda + \Delta_{\mathrm{T}} + 2\Delta_{\mathrm{S}} \;=\; \widehat{\tau_{\mathrm{score}}}(g,W) + \Gamma_Q,
\end{aligned}$$

establishing Eqn. (16). When ℓ is multiclass classification-calibrated, applying the calibrated excess-risk bound of Theorem 11 together with monotonicity of $\psi_\ell^{-1} : [0,\infty) \to [0,\infty)$ to Eqn. (16) on the same event yields Eqn. (17) simultaneously for all $(g, W) \in Q$. □

Vector-contraction reductions. When the associated vector-valued classes are measurable on the sample, the scalar empirical Rademacher complexities in Theorem 12 admit further upper bounds via the vector-contraction inequality of Maurer [55]. For a vector class $\mathcal{V} \subseteq (\mathbb{R}^K)^{\mathcal{Z}}$ on $Z_{1:n}$, write

$$\widehat{\mathfrak{V}}_{Z_{1:n}}(\mathcal{V}) \;:=\; \frac{1}{n}\,\mathbb{E}_\sigma\left[\sup_{v\in\mathcal{V}} \sum_{i=1}^{n}\sum_{k=1}^{K} \sigma_{i,k}\, v_k(Z_i) \;\middle|\; Z_{1:n}\right]$$

with independent uniform Rademacher signs $\sigma_{i,k} \in \{-1,+1\}$, under which sample-wise composition $\varphi \circ \mathcal{V} := \{z_i \mapsto \varphi_i(v(z_i)) : v \in \mathcal{V}\}$ for L-Lipschitz scalar maps $\varphi_i : \mathbb{R}^K \to \mathbb{R}$ (with respect to $\|\cdot\|_2$) satisfies $\widehat{\mathfrak{R}}_{Z_{1:n}}(\varphi \circ \mathcal{V}) \le \sqrt{2}\,L\,\widehat{\mathfrak{V}}_{Z_{1:n}}(\mathcal{V})$ [55, Corollary 4]. First, by Lemma 5, every map in $\mathcal{C}_Q := \{x \mapsto Wc(g(x)) : (g,W) \in Q\}$ takes values in $B(0, R_{U1})$, where $q \mapsto \ell(Y_i, q)$ is L_ℓ-Lipschitz with respect to $\|\cdot\|_2$; contracting $\mathcal{L}_Q$ conditioned on $Y_{1:n_{\mathrm{T}}}$ therefore gives

$$\widehat{\mathfrak{R}}_{D_{\mathrm{T}}}(\mathcal{L}_Q) \;\le\; \sqrt{2}\,L_\ell\,\widehat{\mathfrak{V}}_{X_{1:n_{\mathrm{T}}}}(\mathcal{C}_Q).$$

Second, consider $\Psi(a,b) := \max_{y\in[K_{\mathrm{T}}]} |\ell(y,a) - \ell(y,b)|$ on $B(0,R_{U1}) \times B(0,R_{U1}) \subseteq \mathbb{R}^{2K_{\mathrm{T}}}$ equipped with the Euclidean product norm $\|(a,b)\|_2 = (\|a\|_2^2 + \|b\|_2^2)^{1/2}$. For any $(a,b),(a',b') \in B(0,R_{U1})^2$, expanding $|\Psi(a,b) - \Psi(a',b')|$ one step per line yields

$$\begin{aligned}
\left|\Psi(a,b) - \Psi(a',b')\right| &= \Big|\max_{y\in[K_{\mathrm{T}}]}\left|\ell(y,a) - \ell(y,b)\right| - \max_{y\in[K_{\mathrm{T}}]}\left|\ell(y,a') - \ell(y,b')\right|\Big| \\
&\le \max_{y\in[K_{\mathrm{T}}]}\Big|\left|\ell(y,a) - \ell(y,b)\right| - \left|\ell(y,a') - \ell(y,b')\right|\Big| \\
&\le \max_{y\in[K_{\mathrm{T}}]}\Big|\big(\ell(y,a) - \ell(y,a')\big) - \big(\ell(y,b) - \ell(y,b')\big)\Big| \\
&\le \max_{y\in[K_{\mathrm{T}}]}\Big(\left|\ell(y,a) - \ell(y,a')\right| + \left|\ell(y,b) - \ell(y,b')\right|\Big) \\
&\le L_\ell\big(\|a-a'\|_2 + \|b-b'\|_2\big) \quad (L_\ell\text{-Lipschitz on } B(0,R_{U1})) \\
&\le \sqrt{2}\,L_\ell\sqrt{\|a-a'\|_2^2 + \|b-b'\|_2^2} \quad \text{(Cauchy–Schwarz)} \\
&= \sqrt{2}\,L_\ell\,\big\|(a,b) - (a',b')\big\|_2,
\end{aligned}$$

where the first three inequalities apply $|\max_y u_y - \max_y v_y| \le \max_y |u_y - v_y|$, the reverse triangle inequality $||r| - |s|| \le |r - s|$, and the triangle inequality, so Ψ is $\sqrt{2}\,L_\ell$-Lipschitz on $B(0,R_{U1})^2$. Because $\mathcal{D}_{\mathrm{T},Q} := \{x \mapsto (Ws_{\mathrm{S}}(g(x)), Wc(g(x))) : (g,W) \in Q\}$ and $\mathcal{D}_{\mathrm{S},Q} := \{u \mapsto (Ws_{\mathrm{S}}(u), Wc(u)) : W \in \mathcal{W}_Q\}$ take values in $B(0,R_{U1})^2$ by Lemma 5, applying Maurer [55, Corollary 4] with factor $\sqrt{2}\cdot(\sqrt{2}\,L_\ell) = 2L_\ell$ gives

$$\widehat{\mathfrak{R}}_{X_{1:n_{\mathrm{T}}}}(\mathcal{Z}_{\mathrm{T},Q}) \;\le\; 2L_\ell\,\widehat{\mathfrak{V}}_{X_{1:n_{\mathrm{T}}}}(\mathcal{D}_{\mathrm{T},Q}), \qquad \widehat{\mathfrak{R}}_{D_{\mathrm{S}}}(\mathcal{Z}_{\mathrm{S},Q}) \;\le\; 2L_\ell\,\widehat{\mathfrak{V}}_{D_{\mathrm{S}}}(\mathcal{D}_{\mathrm{S},Q}),$$

providing optional vector-complexity refinements of Theorem 12.

J Worked Examples

Section overview. This section provides the complete combinatorial derivation behind Example 1 in Section 3.3. Computing the conditional-relabeling growth $\Pi_{\mathcal{H}|\mathcal{R}}(m)$ and the structure-oblivious proxy $\Pi^{\mathrm{all}}_{\mathcal{H}}(m)$ in closed form for the unrestricted output-mapping class $\mathcal{H} = [K_{\mathrm{T}}]^{[K_{\mathrm{S}}]}$ shows that exact alphabet-exponent identities hold for all $K_{\mathrm{T}} \geq 1$, while for $K_{\mathrm{T}} \geq 2$ conditioning on reachable source-label patterns eliminates an artificial $O(n_{\mathrm{T}} \log K_{\mathrm{T}})$ log-growth penalty in Theorem 1 (Theorem 7) whenever K_{S} is large relative to the reached source alphabet.

Example 2 (Strict improvement on small reachable alphabets; derivation of Example 1)**.** Under the wrapper setup of Section 2.1 (fixed measurable $f_{\mathrm{S}} : \mathcal{X}_{\mathrm{S}} \to [K_{\mathrm{S}}]$, nonempty $\mathcal{X}_{\mathrm{T}}$, and nonempty $\mathcal{G}$), let $K_{\mathrm{T}} \geq 1$ and $K_{\mathrm{S}} \geq 1$, and consider the unrestricted output-mapping class

$$\mathcal{H} \;=\; [K_{\mathrm{T}}]^{[K_{\mathrm{S}}]}. \tag{18}$$

For every integer $m \geq 1$ and every source-label vector $z = (z_1, \ldots, z_m) \in [K_{\mathrm{S}}]^m$, recall from Example 1 and Appendix A that $T(z) = \{z_1, \ldots, z_m\} \subseteq [K_{\mathrm{S}}]$ denotes the set of distinct source labels appearing in z, with cardinality $|z|_{\#} = |T(z)| \in \{1, \ldots, \min\{m, K_{\mathrm{S}}\}\}$. Two maps $h, h' \in \mathcal{H}$ induce the same target-label vector $(h(z_1), \ldots, h(z_m)) = (h'(z_1), \ldots, h'(z_m))$ if and only if their restrictions $h|_{T(z)}$ and $h'|_{T(z)}$ coincide, and under Eqn. (18) every map in $[K_{\mathrm{T}}]^{T(z)}$ extends to $\mathcal{H}$ (for instance by assigning label $1 \in [K_{\mathrm{T}}]$ outside $T(z)$). The restriction map $h \mapsto h|_{T(z)}$ therefore induces a bijection between $\mathrm{Rel}_{\mathcal{H}}(z)$ and $[K_{\mathrm{T}}]^{T(z)}$, giving the exact pattern count for every $K_{\mathrm{T}} \geq 1$:

$$|\mathrm{Rel}_{\mathcal{H}}(z)| \;=\; \left|[K_{\mathrm{T}}]^{T(z)}\right| \;=\; K_{\mathrm{T}}^{|T(z)|} \;=\; K_{\mathrm{T}}^{|z|_{\#}}. \tag{19}$$

To evaluate both growth functions from Eqn. (19), note that $t \mapsto K_{\mathrm{T}}^t$ is nondecreasing on $\mathbb{R}$ for every $K_{\mathrm{T}} \geq 1$. For the structure-oblivious proxy, every $z \in [K_{\mathrm{S}}]^m$ satisfies $|z|_{\#} \leq \min\{m, K_{\mathrm{S}}\}$, with equality attained at $z^\star := (1, 2, \ldots, q, 1, \ldots, 1) \in [K_{\mathrm{S}}]^m$ where $q := \min\{m, K_{\mathrm{S}}\}$, whence

$$\begin{aligned}
\Pi^{\mathrm{all}}_{\mathcal{H}}(m) &= \sup_{z \in [K_{\mathrm{S}}]^m} |\mathrm{Rel}_{\mathcal{H}}(z)| && \text{(by definition)} \\
&= \sup_{z \in [K_{\mathrm{S}}]^m} K_{\mathrm{T}}^{|z|_{\#}} && \text{(by Eqn. (19))} \\
&= K_{\mathrm{T}}^{\max_{z \in [K_{\mathrm{S}}]^m} |z|_{\#}} && (t \mapsto K_{\mathrm{T}}^t \text{ nondecreasing}) \\
&= K_{\mathrm{T}}^{\min\{m, K_{\mathrm{S}}\}}. && \text{(attained at } z^\star)
\end{aligned}$$

For the conditional-relabeling growth, because $\mathcal{X}_{\mathrm{T}}$ and $\mathcal{G}$ are nonempty, the integer $|z|_{\#} \in \{1, \ldots, \min\{m, K_{\mathrm{S}}\}\}$ attains its maximum $t_{\mathcal{R}}(m) = \sup_{x_{1:m} \in \mathcal{X}_{\mathrm{T}}^m} \sup_{z \in \mathrm{Rea}_{\mathcal{G}}(x_{1:m})} |z|_{\#}$ over reachable patterns, giving

$$\begin{aligned}
\Pi_{\mathcal{H}|\mathcal{R}}(m) &= \sup_{x_{1:m} \in \mathcal{X}_{\mathrm{T}}^m} \sup_{z \in \mathrm{Rea}_{\mathcal{G}}(x_{1:m})} |\mathrm{Rel}_{\mathcal{H}}(z)| && \text{(by Definition 3)} \\
&= \sup_{x_{1:m} \in \mathcal{X}_{\mathrm{T}}^m} \sup_{z \in \mathrm{Rea}_{\mathcal{G}}(x_{1:m})} K_{\mathrm{T}}^{|z|_{\#}} && \text{(by Eqn. (19))} \\
&= K_{\mathrm{T}}^{t_{\mathcal{R}}(m)}. && (t \mapsto K_{\mathrm{T}}^t \text{ nondecreasing})
\end{aligned}$$

Consequently, for every $K_{\mathrm{T}} \geq 1$ and every integer $m \geq 1$, matching Example 1,

$$\Pi^{\mathrm{all}}_{\mathcal{H}}(m) \;=\; K_{\mathrm{T}}^{\min\{m, K_{\mathrm{S}}\}}, \qquad \Pi_{\mathcal{H}|\mathcal{R}}(m) \;=\; K_{\mathrm{T}}^{t_{\mathcal{R}}(m)}. \tag{20}$$

When $K_{\mathrm{T}} = 1$, $\log K_{\mathrm{T}} = 0$ and both $\log \Pi^{\mathrm{all}}_{\mathcal{H}}(2n_{\mathrm{T}})$ and $\log \Pi_{\mathcal{H}|\mathcal{R}}(2n_{\mathrm{T}})$ vanish identically. When $K_{\mathrm{T}} \geq 2$, we have $\log K_{\mathrm{T}} > 0$, and evaluating Eqn. (20) at $m = 2n_{\mathrm{T}}$ in the large-source-alphabet regime $K_{\mathrm{S}} \geq 2n_{\mathrm{T}}$ gives

$$\begin{aligned}
\log \Pi^{\mathrm{all}}_{\mathcal{H}}(2n_{\mathrm{T}}) = \min\{2n_{\mathrm{T}}, K_{\mathrm{S}}\} \log K_{\mathrm{T}} \;&=\; 2n_{\mathrm{T}} \log K_{\mathrm{T}}, \\
\log \Pi_{\mathcal{H}|\mathcal{R}}(2n_{\mathrm{T}}) = t_{\mathcal{R}}(2n_{\mathrm{T}}) \log K_{\mathrm{T}}&,
\end{aligned}$$

which yields the strict separation $\log \Pi_{\mathcal{H}|\mathcal{R}}(2n_{\mathrm{T}}) < \log \Pi^{\mathrm{all}}_{\mathcal{H}}(2n_{\mathrm{T}})$ whenever $t_{\mathcal{R}}(2n_{\mathrm{T}}) < 2n_{\mathrm{T}}$. In particular, when $K_{\mathrm{T}} \geq 2$, $K_{\mathrm{S}} \geq 2n_{\mathrm{T}}$, and $t_{\mathcal{R}}(2n_{\mathrm{T}}) = O(1)$, substituting these logarithmic counts

into the statistical term of Theorem 1 (Theorem 7) shows that the structure-oblivious proxy contributes $\sqrt{\log \Pi_{\mathcal{H}}^{\mathrm{all}}(2n_{\mathrm{T}})/n_{\mathrm{T}}} = \sqrt{2\log K_{\mathrm{T}}}$, yielding a vacuous $\Omega(1)$ penalty, whereas the conditional-relabeling contribution $\sqrt{\log \Pi_{\mathcal{H}|\mathcal{R}}(2n_{\mathrm{T}})/n_{\mathrm{T}}} = \sqrt{t_{\mathcal{R}}(2n_{\mathrm{T}})(\log K_{\mathrm{T}})/n_{\mathrm{T}}} = O\big(\sqrt{(\log K_{\mathrm{T}})/n_{\mathrm{T}}}\big)$ vanishes at rate $n_{\mathrm{T}}^{-1/2}$.

K Controlled Empirical Verification: Protocols and Complete Results

This section provides the complete experimental protocols, formal definitions of the hypothesis-class interventions, sample-split and evaluation governance, full numerical tables across all sample sizes and checkpoints, control comparisons, and compute disclosures for the two controlled empirical studies summarized in Section 7. Code, frozen pretrained checkpoints, and reproduction scripts are publicly available at https://github.com/zyecs/repgm-slt.

K.1 Study 1: Synthetic Exact-ERM Study with Analytical Population Risks

Objective and design rationale. The synthetic study isolates the two core structural mechanisms of our theory under conditions free of optimization slack ($\eta_n = 0$) and Monte Carlo evaluation noise:

1. **Estimation invariance and capacity penalty (Theorem 1):** Since Theorem 1 bounds finite-sample estimation error through the sum $\log \Pi_{\mathcal{R}}(2n_{\mathrm{T}}) + \log \Pi_{\mathcal{H}|\mathcal{R}}(2n_{\mathrm{T}})$, shifting the exact same hypothesis-class capacity between input-side reachability and output-side relabeling while preserving the induced target predictor class $\mathcal{F}_{\mathrm{MR}}$ should leave exact ERM behavior invariant. Simultaneously, enlarging active capacity should increase small-sample estimation error and vanish as n_{T} grows.
2. **Approximation mechanism attribution (Proposition 4):** When two wrapper designs share the exact same suboptimal in-class oracle risk $R_{\mathrm{T}}^{\mathcal{F}_{\mathrm{MR}},*}$, total risk alone cannot identify why reprogramming fails, whereas Proposition 4 separates source-label information loss from output-mapping restriction.

Data-generating distribution and frozen source classifier. The target input space $\mathcal{X}_{\mathrm{T}} = \{e_1, \ldots, e_{16}\} \subset \mathbb{R}^{16}$ consists of the 16 standard basis vectors, grouped into 8 ordered pairs $P_j = (e_{2j-1}, e_{2j})$ for $j \in [8]$. The target marginal $P_{\mathrm{T},X}$ is uniform over $\mathcal{X}_{\mathrm{T}}$ ($P_{\mathrm{T},X}(X = e_i) = 1/16$). For each pair P_j, the noiseless Bayes label is 0 on the first element e_{2j-1} and 1 on the second element e_{2j}. The observed target label $Y \in \{0, 1\}$ is corrupted by independent symmetric flip noise of probability 0.1, yielding a target Bayes risk of $R_{\mathrm{T}}^* = 0.1$ exactly. The baseline frozen source classifier $f_{\mathrm{S}} : \mathbb{R}^{16} \to [16]$ assigns each basis vector e_i its unique coordinate index $f_{\mathrm{S}}(e_i) = i$ with multiclass score margin 1, preserving full input distinguishability prior to any controlled bottleneck.

Part A protocol: Input-side reachability vs. output-side relabeling. For each capacity parameter $r \in \{0, 2, 4, 6, 8\}$ and sample size $n_{\mathrm{T}} \in \{8, 16, 32, 64, 128, 256, 512\}$, we construct two matched arms of cardinality 2^r:

- **Input-side reachability arm (A-R):** The output mapping is fixed to the canonical parity head $h_{\mathrm{can}}(k) = \mathbf{1}\{k \text{ is even}\}$. The input transformation class $\mathcal{G}_r$ consists of 2^r coordinate permutations that independently either preserve or swap the two basis vectors within each of the first r pairs $P_1, \ldots, P_r$ while leaving pairs $P_{r+1}, \ldots, P_8$ fixed. Here $|\mathrm{Rea}_{\mathcal{G}}(x_{1:m})| = 2^{q(x_{1:m})}$ where $q(x_{1:m}) \le r$ is the number of active pairs observed in $x_{1:m}$, while $|\mathrm{Rel}_{\mathcal{H}}(z)| = 1$.
- **Output-side relabeling arm (A-H):** The input transformation is fixed to the identity $g_{\mathrm{id}}(x) = x$. The output mapping class $\mathcal{H}_r$ consists of 2^r relabelings that independently either preserve h_{can} or flip the binary target assignments on both source labels $\{2j-1, 2j\}$ of each of the first r pairs. Here $|\mathrm{Rea}_{\mathcal{G}}(x_{1:m})| = 1$, while $|\mathrm{Rel}_{\mathcal{H}}(z)| = 2^{q(x_{1:m})}$.

Both arms induce the exact same composite predictor class $\mathcal{F}_{\mathrm{MR}}$ of size 2^r and contain the target Bayes rule ($R_{\mathrm{T}}^{\mathcal{F}_{\mathrm{MR}},*} = R_{\mathrm{T}}^* = 0.1$, so $a_{\mathrm{MR}} = 0$). For each (r, n_{T}) cell, we run 200 paired repetitions from root seed 20260725. In each repetition, exhaustive finite-class ERM selects $\widehat{f}_n \in \mathcal{F}_{\mathrm{MR}}$ using a shared label-independent canonical tie-breaking order. We evaluate the exact analytical population estimation excess $\mathcal{E}_{\mathrm{T}}^{\mathrm{est}}(\widehat{f}_n) = R_{\mathrm{T}}(\widehat{f}_n) - 0.1$, alongside descriptive checks on a 20,000-example held-out sample and 10,000 paired whole-repetition bootstrap draws.

Part B protocol: Source-label information loss vs. output-mapping restriction. For each severity level $m \in \{0, 2, 4, 6, 8\}$ (number of affected pairs) and sample size $n_{\mathrm{T}} \in \{32, 128, 512, 2048, 8192\}$,

Table 4: Part A complete results across all capacity levels r and target sample sizes n_{T} (200 paired repetitions). Shown are mean exact population estimation excess $\mathcal{E}_{\mathrm{T}}^{\mathrm{est}}(\widehat{f}_n)$ with descriptive 95% paired-bootstrap intervals, and mean realized $\log_2$ active growth $\mathbb{E}[q(X_{1:n_{\mathrm{T}}})]$. In every (r, n_{T}) cell, the paired difference between A-R and A-H is identically 0.0000 with bootstrap interval $[0.0000, 0.0000]$.

Capacity r	$n_{\mathrm{T}} = 8$	$n_{\mathrm{T}} = 16$	$n_{\mathrm{T}} = 32$	$n_{\mathrm{T}} = 64$	$n_{\mathrm{T}} = 128$	$n_{\mathrm{T}} \in \{256, 512\}$
$r = 0$ ($\mathcal{E}_{\mathrm{T}}^{\mathrm{est}}$; $\mathbb{E}[q] = 0$)	0.0000 [0.000, 0.000]	0.0000 [0.000, 0.000]	0.0000 [0.000, 0.000]	0.0000 [0.000, 0.000]	0.0000 [0.000, 0.000]	0.0000 [0.000, 0.000]
$r = 2$ ($\mathcal{E}_{\mathrm{T}}^{\mathrm{est}}$)	0.0475 [0.040, 0.056]	0.0220 [0.017, 0.028]	0.0080 [0.004, 0.012]	0.0020 [0.001, 0.004]	0.0000 [0.000, 0.000]	0.0000 [0.000, 0.000]
$r = 4$ ($\mathcal{E}_{\mathrm{T}}^{\mathrm{est}}$)	0.0995 [0.088, 0.111]	0.0455 [0.037, 0.053]	0.0155 [0.011, 0.021]	0.0035 [0.001, 0.006]	0.0000 [0.000, 0.000]	0.0000 [0.000, 0.000]
$r = 6$ ($\mathcal{E}_{\mathrm{T}}^{\mathrm{est}}$)	0.1340 [0.120, 0.147]	0.0705 [0.060, 0.081]	0.0255 [0.019, 0.033]	0.0045 [0.002, 0.007]	0.0000 [0.000, 0.000]	0.0000 [0.000, 0.000]
$r = 8$ ($\mathcal{E}_{\mathrm{T}}^{\mathrm{est}}$)	0.1765 [0.162, 0.192]	0.0990 [0.087, 0.111]	0.0295 [0.022, 0.037]	0.0050 [0.002, 0.008]	0.0005 [0.000, 0.002]	0.0000 [0.000, 0.000]
$\mathbb{E}[q]$ ($r = 2/4/6/8$)	1.32/2.57/3.89/5.17	1.78/3.58/5.34/7.10	1.97/3.95/5.92/7.90	2.00/4.00/6.00/8.00	2.00/4.00/6.00/8.00	2.00/4.00/6.00/8.00

we compare two interventions whose in-class oracle risk is identically $R_{\mathrm{T}}^{\mathcal{F}_{\mathrm{MR}},*} = 0.1 + 0.05m$ (so $a_{\mathrm{MR}} = 0.05m$):

- **Information-loss intervention** (B-I)**:** The source channel merges both inputs of each affected pair P_j ($1 \le j \le m$) onto a single shared source label, while the output mapping class $\mathcal{H} = \{0,1\}^{[16]}$ is unrestricted. Because each merged cell pools equal probability mass from Bayes-label 0 and Bayes-label 1 points, no relabeling can do better than random guessing on the first m pairs. Analytically, the source-label information loss is $I := R_{\mathrm{all}}^*(g) - R_{\mathrm{T}}^* = 0.05m$, and the output-mapping restriction is $H := R_{\mathcal{H}}^*(g) - R_{\mathrm{all}}^*(g) = 0$.
- **Head-restriction intervention** (B-H)**:** The input transformation is the identity and all 16 source labels remain distinct ($R_{\mathrm{all}}^*(g) = R_{\mathrm{T}}^* = 0.1$, so $I = 0$), but every $h \in \mathcal{H}$ is constrained to output the incorrect target label 1 on the first member e_{2j-1} (whose Bayes label is 0) of each affected pair P_j ($1 \le j \le m$) while leaving the remaining $16 - m$ source labels free in $\{0,1\}$, forcing one Bayes disagreement per affected pair ($|\mathcal{H}| = 2^{16-m}$, matching $|\mathcal{F}_{\mathrm{MR}}| = 2^{16-m}$ of B-I) and yielding $H = 0.05m$.

In addition to exact population evaluation, we compute empirical plug-in estimates $(\widehat{I}, \widehat{H})$ on an independent 20,000-example diagnostic sample per repetition by comparing empirical risk minima over the nested reference classes $\mathcal{C}_{\mathrm{actual}} \subseteq \mathcal{C}_{\mathrm{unrestricted\text{-}head}} \subseteq \mathcal{C}_{\mathrm{unrestricted\text{-}channel}}$.

Table 5: Part B results across severity levels $m \in \{0, 2, 4, 6, 8\}$ (200 repetitions; bracketed ranges denote descriptive 95% bootstrap intervals for $R_{\mathrm{T}}(\widehat{f}_n)$, which have width 0.000 for all $n_{\mathrm{T}} \in \{512, 2048, 8192\}$). Both arms share identical in-class oracle risk $R_{\mathrm{T}}^{\mathcal{F}_{\mathrm{MR}},*} = 0.1 + 0.05m$, while Proposition 4 and $(\widehat{I}, \widehat{H})$ isolate the underlying mechanism.

m	Arm	Oracle $R_{\mathrm{T}}^{\mathcal{F}_{\mathrm{MR}},*}$	Risk ($n_{\mathrm{T}} = 32$) [95% CI]	Risk ($n_{\mathrm{T}} = 128$) [95% CI]	Risk ($n_{\mathrm{T}} \in \{512, 2048, 8192\}$) [95% CI]	Exact (I, H)	Plug-in $(\widehat{I}, \widehat{H})$
0	B-I / B-H	0.1000	0.1988 [0.190, 0.208]	0.1037 [0.102, 0.106]	0.1000 [0.100, 0.100]	(0.0000, 0.0000)	(0.0000, 0.0000)
2	B-I (info. loss)	0.2000	0.2725 [0.266, 0.280]	0.2025 [0.201, 0.204]	0.2000 [0.200, 0.200]	(0.1000, 0.0000)	(0.0979, 0.0000)
	B-H (head restr.)	0.2000	0.2823 [0.275, 0.290]	0.2032 [0.202, 0.205]	0.2000 [0.200, 0.200]	(0.0000, 0.1000)	(0.0000, 0.1000)
4	B-I (info. loss)	0.3000	0.3520 [0.346, 0.358]	0.3018 [0.301, 0.303]	0.3000 [0.300, 0.300]	(0.2000, 0.0000)	(0.1958, 0.0000)
	B-H (head restr.)	0.3000	0.3723 [0.365, 0.380]	0.3028 [0.301, 0.304]	0.3000 [0.300, 0.300]	(0.0000, 0.2000)	(0.0000, 0.1999)
6	B-I (info. loss)	0.4000	0.4258 [0.421, 0.430]	0.4005 [0.400, 0.401]	0.4000 [0.400, 0.400]	(0.3000, 0.0000)	(0.2941, 0.0000)
	B-H (head restr.)	0.4000	0.4578 [0.451, 0.465]	0.4022 [0.401, 0.404]	0.4000 [0.400, 0.400]	(0.0000, 0.3000)	(0.0000, 0.2998)
8	B-I (info. loss)	0.5000	0.5000 [0.500, 0.500]	0.5000 [0.500, 0.500]	0.5000 [0.500, 0.500]	(0.4000, 0.0000)	(0.3923, 0.0000)
	B-H (head restr.)	0.5000	0.5435 [0.538, 0.550]	0.5020 [0.501, 0.504]	0.5000 [0.500, 0.500]	(0.0000, 0.4000)	(0.0000, 0.4000)

K.2 Study 2: Pretrained MLP and LeNet Sources on Rotated-MNIST Parity Prediction

Source-model pretraining and hard-label interface. We pretrain three independent checkpoints of a two-hidden-layer MLP ($784 \to 256 \to \mathrm{ReLU} \to 128 \to \mathrm{ReLU} \to 10$) and three independent checkpoints of a minimal convolutional network (LeNet: $\mathrm{Conv}(1, 6, 5) \to \mathrm{ReLU} \to \mathrm{AvgPool}(2) \to \mathrm{Conv}(6, 16, 5) \to \mathrm{ReLU} \to \mathrm{AvgPool}(2) \to 256 \to 120 \to \mathrm{ReLU} \to 84 \to \mathrm{ReLU} \to 10$) on upright 10-class MNIST [49] in PyTorch [60]. The official 60,000-image training set is split into 55,000 training images and a digit-balanced 5,000-image validation split (normalized with mean 0.1307 and standard deviation 0.3081). Each network is trained for up to 20 epochs using Adam (learning rate 10^{-3}, batch size 128, cross-entropy loss) without data augmentation, selecting the checkpoint with highest validation accuracy (all six checkpoints exceed 0.976 validation accuracy: MLP checkpoints attain 0.9790, 0.9782, 0.9768; LeNet checkpoints attain 0.9864, 0.9886, 0.9890).

Every checkpoint is then frozen, and its output interface is restricted to the single hard source class ID $f_{\mathrm{S}}(u) \in \{0, \ldots, 9\}$; logits, probabilities, internal activations, and gradients are never exposed to the downstream learner.

Target task, sample partitions, and wrapper/head hypothesis classes. The downstream pool consists of the unseen 10,000-image official MNIST test set. Every target image is rotated counterclockwise by 90° ($\mathrm{rot}_{90}(x)$), and the target label is binary digit parity ($Y = \text{digit} \bmod 2 \in \{0, 1\}$). Using deterministic stratified allocation, the 10,000 target images are partitioned into a 4,000-image target training pool, a 2,000-image diagnostic split, and a 4,000-image held-out final evaluation split. Across 20 paired target seeds per source checkpoint, we draw nested, digit-balanced training subsets of sizes $n_{\mathrm{T}} \in \{50, 200, 800, 3200\}$ from the training pool. We evaluate a 2×2 factorial design crossing two input-transformation classes and two output-mapping classes:

- **Input transformations:** $\mathcal{G}_{\mathrm{id}} = \{\mathrm{rot}_0\}$ ($|\mathcal{G}_{\mathrm{id}}| = 1$, leaving the 90° rotation uncorrected) versus $\mathcal{G}_{\mathrm{rot}} = \{\mathrm{rot}_0, \mathrm{rot}_{90}, \mathrm{rot}_{180}, \mathrm{rot}_{270}\}$ ($|\mathcal{G}_{\mathrm{rot}}| = 4$, which includes the exact inverse rotation rot_{270}).
- **Output mappings:** $\mathcal{H}_{\mathrm{thr}}$ ($|\mathcal{H}_{\mathrm{thr}}| = 20$, comprising the 2 constant maps and both orientations $\mathbf{1}\{k \geq t\}$, $\mathbf{1}\{k < t\}$ of single-threshold maps for $t \in \{1, \ldots, 9\}$, which cannot represent alternating even/odd parity) versus $\mathcal{H}_{\mathrm{full}} = \{0, 1\}^{\{0,\ldots,9\}}$ ($|\mathcal{H}_{\mathrm{full}}| = 2^{10} = 1{,}024$, containing all binary relabelings of the 10 source digit IDs).

For every checkpoint, target seed, sample size n_{T}, and factorial cell, exact finite-class ERM enumerates all candidate wrapper-head pairs on the target training subset (breaking ties lexicographically by canonical wrapper and head indices). All 960 predictor selections per architecture were locked on the 2,000-image diagnostic split before evaluating them once, without reselection, on the 4,000-image final evaluation split.

Estimands, controls, and complete learning curves. At $n_{\mathrm{T}} = 3200$, we measure the input-side reachability effect $\Delta_{\mathrm{in}} := R_{\mathrm{T}}(\mathcal{G}_{\mathrm{id}}, \mathcal{H}_{\mathrm{full}}) - R_{\mathrm{T}}(\mathcal{G}_{\mathrm{rot}}, \mathcal{H}_{\mathrm{full}})$, the output-side relabeling effect $\Delta_{\mathrm{out}} := R_{\mathrm{T}}(\mathcal{G}_{\mathrm{rot}}, \mathcal{H}_{\mathrm{thr}}) - R_{\mathrm{T}}(\mathcal{G}_{\mathrm{rot}}, \mathcal{H}_{\mathrm{full}})$, their interaction $\Delta_{\mathrm{int}} := [R_{\mathrm{T}}(\mathcal{G}_{\mathrm{id}}, \mathcal{H}_{\mathrm{thr}}) - R_{\mathrm{T}}(\mathcal{G}_{\mathrm{id}}, \mathcal{H}_{\mathrm{full}})] - \Delta_{\mathrm{out}}$, and the contrast $\Delta_{\mathrm{train}} := R_{\mathrm{T}}(\text{untrained}, \mathcal{G}_{\mathrm{rot}}, \mathcal{H}_{\mathrm{full}}) - R_{\mathrm{T}}(\text{trained}, \mathcal{G}_{\mathrm{rot}}, \mathcal{H}_{\mathrm{full}})$ against one prespecified randomly initialized untrained checkpoint of the same architecture. Target-seed averages are computed within each checkpoint and equally weighted across the three checkpoints.

Table 6: Complete final-split learning curves and $n_{\mathrm{T}} = 3200$ mechanism contrasts on rotated-MNIST parity prediction (equally weighted across 3 pretrained checkpoints $\times$ 20 paired target seeds per architecture; bracketed ranges denote descriptive 2.5%–97.5% target-seed quantiles).

Model	Sample size	$(\mathcal{G}_{\mathrm{id}}, \mathcal{H}_{\mathrm{thr}})$	$(\mathcal{G}_{\mathrm{id}}, \mathcal{H}_{\mathrm{full}})$	$(\mathcal{G}_{\mathrm{rot}}, \mathcal{H}_{\mathrm{thr}})$	$(\mathcal{G}_{\mathrm{rot}}, \mathcal{H}_{\mathrm{full}})$
MLP	$n_{\mathrm{T}} = 50$	0.4701 [0.443, 0.520]	0.4416 [0.407, 0.494]	0.3967 [0.361, 0.458]	0.0098 [0.010, 0.010]
	$n_{\mathrm{T}} = 200$	0.4515 [0.435, 0.476]	0.4069 [0.396, 0.430]	0.3718 [0.362, 0.391]	0.0098 [0.010, 0.010]
	$n_{\mathrm{T}} = 800$	0.4439 [0.436, 0.472]	0.4006 [0.395, 0.408]	0.3636 [0.359, 0.381]	0.0098 [0.010, 0.010]
	$n_{\mathrm{T}} = 3200$	0.4385 [0.435, 0.440]	0.3980 [0.394, 0.400]	0.3613 [0.359, 0.368]	0.0098 [0.010, 0.010]
LeNet	$n_{\mathrm{T}} = 50$	0.4206 [0.401, 0.453]	0.4006 [0.362, 0.449]	0.3621 [0.335, 0.438]	0.0058 [0.006, 0.006]
	$n_{\mathrm{T}} = 200$	0.4029 [0.396, 0.416]	0.3634 [0.354, 0.382]	0.3460 [0.335, 0.371]	0.0058 [0.006, 0.006]
	$n_{\mathrm{T}} = 800$	0.3979 [0.396, 0.403]	0.3552 [0.354, 0.363]	0.3374 [0.335, 0.345]	0.0058 [0.006, 0.006]
	$n_{\mathrm{T}} = 3200$	0.3960 [0.396, 0.396]	0.3538 [0.354, 0.354]	0.3352 [0.335, 0.335]	0.0058 [0.006, 0.006]

Model ($n_{\mathrm{T}} = 3200$)	Input effect Δ_{in}	Output effect Δ_{out}	Interaction Δ_{int}	Untrained contrast Δ_{train}
MLP (3-ckpt mean)	0.3881 [0.384, 0.390]	0.3514 [0.349, 0.358]	−0.3109 [−0.317, −0.308]	0.4257 [0.426, 0.426]
LeNet (3-ckpt mean)	0.3480 [0.348, 0.348]	0.3293 [0.329, 0.329]	−0.2872 [−0.287, −0.287]	0.5017 [0.502, 0.502]

Control checks, empirical growth accounting, and compute resources. Across all 60 $(\mathcal{G}_{\mathrm{rot}}, \mathcal{H}_{\mathrm{full}})$ runs per architecture and all $n_{\mathrm{T}} \in \{50, 200, 800, 3200\}$, exact ERM selected the true inverse rotation rot_{270} in 100% of runs, matching the risk of the mathematical oracle $(\mathrm{rot}_{270}, k \mapsto k \bmod 2)$ (0.0098 on MLP and 0.0058 on LeNet). By contrast, majority-class prediction on the final split has risk 0.4925, and the single untrained source control under $(\mathcal{G}_{\mathrm{rot}}, \mathcal{H}_{\mathrm{full}})$ at $n_{\mathrm{T}} = 3200$ attains mean risk 0.4355 on MLP and 0.5075 on LeNet, confirming that low target risk requires both learned source structure and a wrapper-head pair where neither reachability nor relabeling is bottlenecked.

Direct enumeration of prediction patterns on each target sample $x_{1:n_{\mathrm{T}}}$ verified the sample-wise factorization bound $|\{(f(x_1), \ldots, f(x_{n_{\mathrm{T}}})) : f \in \mathcal{F}_{\mathrm{MR}}\}| \leq |\mathrm{Rea}_{\mathcal{G}}(x_{1:n_{\mathrm{T}}})| \max_{z \in \mathrm{Rea}_{\mathcal{G}}(x_{1:n_{\mathrm{T}}})} |\mathrm{Rel}_{\mathcal{H}}(z)|$ in all $2{,}400$ audited configurations per architecture (at $n_{\mathrm{T}} = 3200$, $|\mathrm{Rea}_{\mathcal{G}}(x_{1:n_{\mathrm{T}}})|$ is 1 under $\mathcal{G}_{\mathrm{id}}$ and 4 under $\mathcal{G}_{\mathrm{rot}}$, yielding 20, $1{,}024$, 74, and $4{,}090$ composed prediction vectors across the four cells). The synthetic study required 18.46 CPU seconds (139.2 MB peak RSS), and the MNIST study required 377.1 seconds for MLP and 397.4 seconds for LeNet on a single NVIDIA A100-80GB MIG (1g.10gb) slice.

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 paper's scope as a wrapper-side finite-sample learnability analysis for hard-label model reprogramming, together with approximation analysis, a logit-exposing extension, and controlled empirical mechanism checks. The stated claims correspond to the main results in Sections 3 to 8 and Appendices B–K.

   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: Section 8 explicitly discusses the theoretical and empirical boundaries of the results, including source-architecture coverage of Condition A, continuous-interface extensions, the one-directional nature of the approximation certificate, and the controlled mechanism-check scope of the empirical studies.

   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: The paper states the model, measurability, sample, growth-function, additive-transformation, source-score, output-mapping, and boundedness assumptions in the relevant theorem statements and in Appendix A. Complete formal proofs are provided in Appendices C–I for the estimation bound, additive specialization, lower bound, approximation identities, hard-label and score-output certificates, and uniform empirical certificate.

   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: Section 7 and Appendix K disclose the complete data-generating distributions, source-model architectures, pretraining splits, wrapper/head hypothesis classes, tie-breaking rules, random seed counts, and evaluation protocols for both controlled studies.

   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: Complete mathematical protocols, architectures, seed schedules, and evaluation rules for both controlled studies are provided in Section 7 and Appendix K, and the standalone reproduction code, frozen pretrained checkpoints, and result tables are publicly available at https://github.com/zyecs/repgm-slt.

   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: Section 7 and Appendix K specify all pretraining and target data splits ($55{,}000/5{,}000$ source train/validation; $4{,}000/2{,}000/4{,}000$ target train/diagnostic/final evaluation), network architectures, optimizer hyperparameters, candidate wrapper/head classes, and exact ERM selection rules.

   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: Appendix K reports 95% paired-bootstrap intervals across 200 paired repetitions for the synthetic study (Table 4) and 2.5%–97.5% target-seed quantile intervals across 3 source checkpoints × 20 paired target seeds for the MNIST study (Table 6).

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 discloses the compute hardware, memory footprint, and execution wall times (18.46 CPU seconds with 139.2 MB peak RSS for the synthetic study; 377.1 s for MLP and 397.4 s for LeNet on a single NVIDIA A100-80GB MIG 1g.10gb slice).

   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: To the best of the authors' knowledge, the work conforms to the NeurIPS Code of Ethics. The paper is theoretical, uses no human subjects or private/sensitive data, and does not release models, datasets, or systems requiring special ethical safeguards.

   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: [No].

    Justification: The paper discusses how the theory helps understand model reprogramming practice, but it does not include a dedicated discussion of positive and negative societal impacts. Any societal impact is indirect, through downstream use of model-reuse and adaptation methods rather than through a deployed system, released model, or dataset introduced by this work.

    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 does not release data, pretrained models, image generators, scraped datasets, or other high-risk assets requiring release safeguards.

    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: The controlled real-image study uses the standard MNIST dataset [49] (available under the Creative Commons Attribution-ShareAlike 3.0 license) and PyTorch [60] (BSD-style license), as cited in Appendix K.

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: [N/A].

Justification: The paper is a theoretical study and does not introduce a new dataset, benchmark, or pretrained model asset; the controlled empirical verification uses the existing MNIST dataset and standard synthetic distributions fully specified in Appendix K.

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: The paper does not involve crowdsourcing, user studies, data collection from people, or other human-subjects research.

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 human subjects, so IRB approval or equivalent review is not applicable.

    Guidelines:

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

16. **Declaration of LLM usage**

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

    Answer: [N/A].

    Justification: The core method development and theoretical analysis in this research do not involve LLMs as any important, original, or non-standard components; LLMs were used only for language polishing and formatting.

    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.