Language-Induced Priors for Domain Adaptation

Qiyuan Chen
Industrial and Operations Engineering
University of Michigan
Ann Arbor, MI 48109
cqiyuan@umich.edu

Jiayu Zhou
School of Information
University of Michigan
Ann Arbor, MI 48109
jiayuz@umich.edu

Raed Al Kontar
Industrial and Operations Engineering
University of Michigan
Ann Arbor, MI 48109
alkontar@umich.edu

Abstract

Domain adaptation faces a fundamental paradox in the cold-start regime. When target data is scarce, statistical methods fail to distinguish relevant source domains from irrelevant ones, which often leads to negative transfer. In this paper, we address this challenge by leveraging expert textual descriptions of the target domain, a resource that is often available but overlooked. We propose a probabilistic framework that translates these semantic descriptions into a choice model, namely a Language-Induced Prior (LIP), that learns the preferences from a pretrained Large Language Model (LLM). The LIP is then integrated into an Expectation-Maximization algorithm to identify source relevance. Methodologically, this framework is compatible with any parametric model where a likelihood is available. It allows the LIP to guide the selection of sources when target signals are weak, while gradually refining these choices as samples accumulate. Theoretically, we prove that the estimator roughly matches an oracle cold-start MSE under a correct prior, while remaining asymptotically consistent regardless of the quality of the LIP. Empirically, we validated the framework on a descriptive (Gaussian estimation), a predictive (C-MAPSS dataset), and a prescriptive task (MuJoCo hopper).

1 Introduction

The success of modern machine learning is driven by the abundance of data, but it can be fragile in the cold-start regime. When an agent starts its task in a new environment, whether it is a new machine deployed to an assembly line or a diagnostic model for a new patient, it becomes statistically challenging to estimate an accurate model from the limited observations. The standard remedy for this problem is Domain Adaptation (DA). DA leverages the abundant data from existing agents that potentially operate under different environments. These environments from which data are borrowed are called source domains, while the new environment we care about is called the target domain.

However, the mere existence of source data is not enough. In realistic settings, source domains are usually heterogeneous, including both relevant and irrelevant domains. If an agent indiscriminately pools data from all available machines, it can suffer from negative transfer, where the borrowed knowledge biases the model and results in even worse performance compared to a simple model trained on the limited target data alone. For example, the corrosion of a machine operating in a humid environment can occur faster compared to one operating in a dry environment. If we train a model

Preprint.

for a humid environment with the degradation data generated in a dry one, the remaining life can be significantly longer than its actual value.

Although identifying the relevant sources is important for DA, it is not easier than learning a good model with limited data. In fact, if one can correctly identify the relevant sources, then an accurate model can be trained using the data from these sources. This creates a paradox in the cold-start regime. To reliably borrow data from the source domains, the agent needs a reliable understanding of the target domain, which is unrealistic given the limited data. Fundamentally, regardless of how rich the source datasets are, the performance of the resulting model is bottle-necked by the limited information regarding the target domain, which calls for attention beyond the source datasets.

Fortunately, strictly numerical data is rarely the only information available when deploying a new agent. In many practical applications, we usually have some contextual information about the target domain. This contextual information, usually taking the form of expert textual descriptions, is a rich modality of information that has been overlooked in the DA literature. A maintenance log might note that a machine is "vibrating more than usual," or a clinical chart might describe a patient as "working irregular night shifts". While these descriptions do not contain labeled training samples, they induce a strong prior regarding the underlying physics of the environment.

Humans naturally evade the cold-start paradox by leveraging this context. An engineer need not observe a hundred failure events to know that a vibrating machine fails more often. Semantic descriptions immediately remind us of relevant prior experiences. In this paper, we bridge the gap between semantic reasoning and statistical learning. Our contributions are summarized as follows.

- **Language-Induced Prior (LIP).** To the best of our knowledge, this is the first work to define and leverage LIP for identifying source relevance in DA (see related work in the Appendix A). We use a choice model to quantify the relevance of the sources, which guides DA in a cold-start setting.
- **LIP-aided EM.** LIP is integrated into MDA through a Bayesian hierarchical model with latent source relevance, which is solved by an Expectation-Maximization (EM) algorithm. This allows the model to be guided by expert intuition when the target signals are weak, while automatically refining the source relevance as data accumulates.
- **Theoretical guarantees.** Our theory verifies the methodological claims. Theorem 4.1 shows that EM behaves as if it is trained on all the relevant sources under a correct prior, demonstrating why and how semantic knowledge helps. The fall-back analysis shows that both the E-step (Theorem 4.3) and the M-step (Theorem 4.2) are consistent under any LIP, including the incorrect ones.
- **Empirical validation.** We empirically validate our approach on three tasks: an illustrative example of Gaussian estimation (both 1-d and 2-d); a case study on the C-MAPSS dataset when no approximations are needed; and a deep reinforcement learning example for MuJoCo hopper. In all cases, our LIP-aided EM shows superior performance, especially when the target data is scarce. Our code is available at `https://github.com/Chen-Qiyuan/LIP-EM`.

2 Setting

Problem Formulation Consider $K+1$ independent agents indexed by k, where each agent operates under a corresponding domain k whose behavior is controlled by a set of latent parameters θ_k of dimension d. Without loss of generality, we let domain $k = 0$, where agent $k = 0$ operates, be the target domain, and the rest of the agents $k \in \{1, \dots, K\}$ are considered the source domains. For each domain k, we denote their historical dataset as $\{\mathcal{D}_k\}_{k=1}^{K}$, where each $\mathcal{D}_k$ contains N_k i.i.d. data points generated by parameter θ_k. Note that the source domains are purely for borrowing knowledge, and we care only about the performance in the target domain, which boils down to the estimation of θ_0. We note that this problem setup is a general framework for any parameter estimation problem.

We model the data generating process of a multi-source domain adaptation (MDA) problem as a Bayesian hierarchical model shown in Fig. 1. A prior π_k generates a latent indicator variable $c_k \in \{0, 1\}$, determining whether a source domain k shares a similar distribution with the target, i.e.,

$$\theta_k \sim \begin{cases} \mathcal{N}(\theta_0, \tau^2 I) & c_k = 1 \\ \phi_{\text{null}} & c_k = 0 \end{cases} \tag{1}$$

If $c_k = 1$, then the parameter of domain k is sampled from an isotropic normal distribution centered at θ_0 with per-coordinate variance τ^2, where the hyperparameter τ is a *small* number that defines how

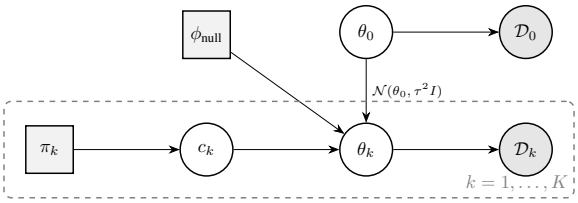


Figure 1: Generative Directed Acyclic Graph

similar the relevant domains are. In the special case where $\tau = 0$, it is enforced that all the relevant source domains should have the exact same parameter $\theta_k = \theta_0$. If $c_k = 0$, the agent operates under a domain-specific noise parameter θ_k generated by the noise component ϕ_{null}.

Motivation of the LIP As highlighted earlier, determining membership c_k in a cold-start regime presents a circular challenge: when N_0 is scarce, identifying relevant sources requires a reliable estimate of θ_0, yet estimating θ_0 requires knowing which sources to borrow from. We break this dilemma by supplying additional information through a *Language-Induced Prior* (LIP) $\pi = (\pi_1, \ldots, \pi_K)$, where $\pi_k = \mathbb{P}(c_k = 1)$. LIP leverages a textual description of the target domain to encode source relevance before any target data is processed. We note that our framework requires a text description *only* for the target environment. This reflects how such information arises in practice: domain experts typically describe the current system of interest (e.g., "the machine is vibrating abnormally"), but equivalent descriptions for historical source datasets are rarely recorded and are infeasible to reconstruct retrospectively. While experts could, in principle, assess each source's relevance from the target description, doing so manually is costly and does not scale with size K. We propose to use pretrained LLMs as a practical alternative, leveraging their reasoning ability and broad pretrained knowledge to parse the description and assess relevance.

3 Methodology

3.1 LIP Construction

To construct the LIP, we use Empirical Bayes to find the prior that maximizes the likelihood of LLM responses. In this step, one can be creative in the elicitation method, and any prior (even one that is not created by an LLM) is compatible with the proposed EM algorithm in the next section. Here, we propose one exemplary method through subgroup selection (see details in Appendix B). Roughly, this consists of two steps (see Fig. 2). First, we elicit a comparison dataset M from an LLM judge. Then, the LIP π is fitted by maximizing its likelihood on M.

Subgroup Preference Elicitation We pose the problem to the LLM as a multiple choice question. For each query $m = 1, \ldots, N_M$, we sample a subgroup $S_m \subseteq \{1, \ldots, K\}$ and ask the LLM to select the most relevant source from S_m given sufficient background knowledge about the data structure,

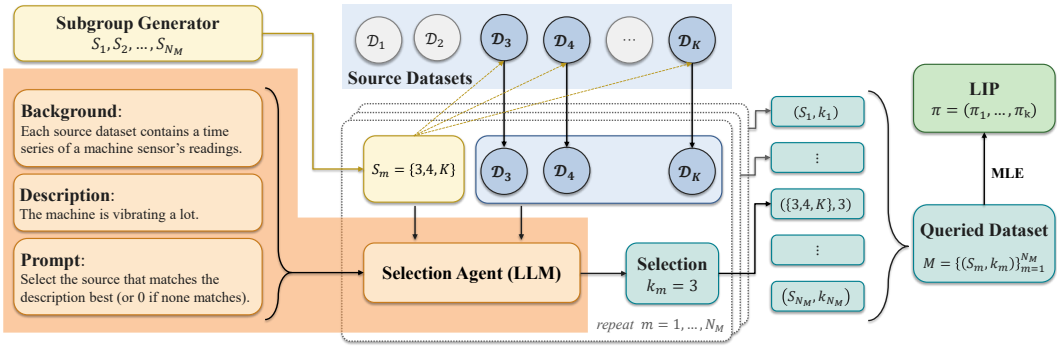


Figure 2: LIP Pipeline

the target description h, and the corresponding datasets $\{D_k\}_{k\in S_m}$, or return $k_m = 0$ if no candidate is deemed relevant. We record the responses as $M = \{(S_m, k_m)\}_{m=1}^{N_M}$.

Empirical Bayes for LIP We now estimate the prior probability distribution by maximizing its likelihood on M. LIP is parameterized as $\pi_k = \sigma(\alpha_k)$, where α_k is a real-valued latent variable associated with the source k. Because the LLM is restricted to selecting at most one candidate per query, we model its choice as a *conditional logit with an outside option* [McFadden, 1974, Luce et al., 1959]. The null option is a virtual alternative whose worth α_0 represents the selection threshold. Concretely, the probability that the LLM returns choice $k_m \in S_m \cup \{0\}$ given subgroup S_m is

$$p(k_m \mid S_m) = \frac{e^{\alpha_{k_m}}}{e^{\alpha_0} + \sum_{j\in S_m} e^{\alpha_j}}, \tag{2}$$

where $\alpha_{k_m} = \alpha_0$ when $k_m = 0$ (the null is selected) and $\alpha_{k_m} = \alpha_k$ when a source k is selected.

To find the most probable LIP, we minimize the regularized negative log-likelihood:

$$\min_{\alpha_0,\alpha_1,\ldots,\alpha_K} \quad -\sum_{m=1}^{N_M} \log p(k_m \mid S_m) + \epsilon \sum_{k=1}^{K} \left(\alpha_k - \log \frac{p_0}{1-p_0}\right)^2, \tag{3}$$

where ϵ is an L_2 regularization coefficient and p_0 is the default probability for a uniform prior.

3.2 Optimization with LIP-aided EM

Recall that the goal in our problem setup is to estimate an accurate θ_0. At this stage, we have both the LIP π and the observed data $\mathbf{O} := \{\mathcal{D}_0, \ldots, \mathcal{D}_K\}$. Under a probabilistic model with latent parameters, a natural choice to infer the memberships is EM. We denote the latent memberships as $\mathbf{c} := \{c_1, \ldots, c_K\}$ and optimize them jointly with θ_0 by maximizing the complete-data likelihood

$$p(\mathbf{O}, \mathbf{c} \mid \theta) = \mathcal{L}(\theta; \mathcal{D}_0) \prod_{k=1}^{K} p(\mathcal{D}_k \mid c_k = 1, \theta)^{c_k}\, p(\mathcal{D}_k \mid c_k = 0)^{1-c_k}\, \pi_k^{c_k} (1-\pi_k)^{1-c_k}. \tag{4}$$

EM alternates between an E-step (updating the posterior of $\mathbf{c}$) and an M-step (maximizing over θ), where each iteration aims to maximize the expectation of the complete-data log-likelihood with respect to the posterior of $\mathbf{c}$, conditional on the current iterate:

$$\theta^{(t+1)} \leftarrow \arg\max_{\theta} \mathbb{E}_{\mathbf{c}\mid\mathbf{O},\theta^{(t)}} \left[\log p(\mathbf{O}, \mathbf{c} \mid \theta)\right]. \tag{5}$$

3.2.1 The E-Step (Expectation)

To optimize (5), we need to first compute the objective. Using the Bernoulli form of c_k ($\mathbb{P}(c_k) = \pi_k^{c_k}(1-\pi_k)^{1-c_k}$) and dropping θ-independent terms (see Appendix D.1), we can reduce (5) to

$$\max_{\theta}\ \log \mathcal{L}(\theta; \mathcal{D}_0) + \sum_{k=1}^{K} w_k^{(t)} \log p(\mathcal{D}_k \mid c_k = 1, \theta), \tag{6}$$

where the objective function in (5) can be written as a weighted sum of likelihoods with weights $w_k^{(t)} := \mathbb{E}[c_k \mid \mathcal{D}_k, \theta^{(t)}] = \mathbb{P}(c_k = 1 \mid \mathcal{D}_k, \theta^{(t)})$. To compute these weights $w_k^{(t)}$, we apply Bayes' rule and write it in a numerically stable sigmoid form:

$$w_k^{(t)} = \sigma\left(\log \frac{p(\mathcal{D}_k \mid c_k = 1, \theta^{(t)})}{p(\mathcal{D}_k \mid c_k = 0)} + \log \frac{\pi_k}{1-\pi_k}\right), \tag{7}$$

where σ is the sigmoid function. In a nutshell, the log-Likelihood Ratio (log-LR) $\log[p(\mathcal{D}_k \mid c_k = 1, \theta^{(t)})/p(\mathcal{D}_k \mid c_k = 0)]$ corrects the prior π_k by the empirical evidence. When the relevant likelihood $p(\mathcal{D}_k \mid c_k = 1, \theta^{(t)})$ is larger than the null likelihood $p(\mathcal{D}_k \mid c_k = 0)$, the ratio is positive and raises the weight and conversely lowers it. Now what remains is computing these two likelihoods.

Relevant Likelihood Given $\theta^{(t)}$, $p(\mathcal{D}_k \mid c_k = 1, \theta^{(t)})$ can be computed by marginalizing out θ_k

$$p(\mathcal{D}_k \mid c_k = 1, \theta^{(t)}) = \int \mathcal{L}(\theta_k; \mathcal{D}_k)\, \mathcal{N}(\theta_k \mid \theta^{(t)}, \tau^2 I)\, d\theta_k,$$

which is generally intractable. For low-dimensional θ_k, one may efficiently approximate the integral using Monte Carlo methods with $\theta_s \sim \mathcal{N}$. However, when d is large, sampling θ_s can be inefficient, so we approximate the likelihood with a second-order Taylor expansion of $\log \mathcal{L}(\,\cdot\,; \mathcal{D}_k)$ at $\theta^{(t)}$.

$$\log \mathcal{L}(\theta; \mathcal{D}_k) \approx \log \mathcal{L}(\theta^{(t)}; \mathcal{D}_k) + g_k^{(t)\intercal}(\theta - \theta^{(t)}) - \tfrac{1}{2}(\theta - \theta^{(t)})^\intercal H_k^{(t)}(\theta - \theta^{(t)}), \tag{8}$$

where $g_k^{(t)} := \nabla_\theta \log \mathcal{L}(\theta; \mathcal{D}_k)|_{\theta=\theta^{(t)}}$ and $H_k^{(t)} := -\nabla_\theta^2 \log \mathcal{L}(\theta; \mathcal{D}_k)|_{\theta=\theta^{(t)}}$. Under this approximation, the relevant likelihood can be approximated by (9) (Proposition 3.1, proved in the Appendix).

Proposition 3.1. *Fix $\tau > 0$ and $\theta^{(t)} \in \mathbb{R}^d$, and assume $H_k^{(t)} \succeq 0$. Approximation* (8) *yields*

$$\log p(\mathcal{D}_k \mid c_k = 1, \theta^{(t)}) \approx \log \mathcal{L}(\theta^{(t)}; \mathcal{D}_k) + \tfrac{\tau^2}{2}\, g_k^{(t)\intercal}(I + \tau^2 H_k^{(t)})^{-1} g_k^{(t)} - \tfrac{1}{2} \log\det(I + \tau^2 H_k^{(t)}). \tag{9}$$

Remark 3.2 (Zero-shift limit). At $\tau = 0$, both sides of (9) reduce to $\log \mathcal{L}(\theta^{(t)}; \mathcal{D}_k)$, so the approximation holds with equality.

Remark 3.3 (Exactness for Gaussian likelihood). Suppose $p(x \mid \theta_k) = \mathcal{N}(x;\, \theta_k, \Sigma_k)$ with known $\Sigma_k \succ 0$. Then (8) holds with equality at every $\theta^{(t)}$, and consequently (9) holds with equality.

Remarks 3.2 and 3.3 are proved in the Appendix. When τ is small, the approximation in (9) is justified even for non-Gaussian likelihoods. Note that a larger τ relaxes the relevance criterion and includes more sources, but at the cost of additional estimation error (see the discussion in Sec. 4).

Null Likelihood $p(\mathcal{D}_k \mid c_k = 0)$ is determined by the null prior ϕ_{null}, a user-specified hyperparameter that acts as a background noise level, where a source k is deemed relevant only if its likelihood exceeds this threshold. The precise choice of ϕ_{null} is uncritical when N_k is large, as the relevance weight commits to the correct value regardless (Theorem 4.3). However, a null floor set too low fails to suppress irrelevant sources and triggers negative transfer. It is therefore safer to err on the side of a slightly larger null likelihood. Two practical methods for setting ϕ_{null} are discussed in Appendix D.2.

3.2.2 The M-Step (Maximization)

Having computed the relevance weights $w_k^{(t)}$ in the E-step, we can now solve the maximization problem in (6). The objective Q-function with respect to θ is

$$Q(\theta \mid \theta^{(t)}) = \log \mathcal{L}(\theta; \mathcal{D}_0) + \sum_{k=1}^{K} w_k^{(t)} \log p(\mathcal{D}_k \mid c_k = 1, \theta). \tag{10}$$

Replacing the relevant log-likelihood with (9) gives

$$\begin{aligned} \max_{\theta^{(t+1)}}\ & \log \mathcal{L}(\theta^{(t+1)}; \mathcal{D}_0) + \sum_k w_k^{(t)} \log \mathcal{L}(\theta^{(t+1)}; \mathcal{D}_k) \\ & + \sum_k w_k^{(t)} \left[\frac{\tau^2}{2} g_k^{(t+1)\intercal} \left(I + \tau^2 H_k^{(t+1)}\right)^{-1} g_k^{(t+1)} - \frac{1}{2} \log\det\left(I + \tau^2 H_k^{(t+1)}\right) \right], \end{aligned} \tag{11}$$

which can be interpreted as maximizing the log likelihood regularized by a τ^2-weighted term. Yet, solving (11) exactly requires computing the gradient g_k and inverting a costly $d \times d$ matrix, which is not practical for high-dimensional θ_0. As such, we introduce two approaches for fast computations.

First, if we re-expand each $\log \mathcal{L}(\,\cdot\,; \mathcal{D}_k)$ around the local Maximum Likelihood Estimation (MLE) $\hat{\theta}_k := \arg\max_{\theta_k} \mathcal{L}(\theta_k; \mathcal{D}_k)$ rather than around the current iterate $\theta^{(t)}$. This shift of the expansion center kills the gradient term (since the gradient is 0 at MLE) and replaces the iteration-dependent Hessian $H_k^{(t)}$ with the iteration-*independent* Hessian at the MLE:

$$\log \mathcal{L}(\theta; \mathcal{D}_k) \approx \text{const} - \tfrac{1}{2}(\theta - \hat{\theta}_k)^\intercal H_k^{\text{MLE}}(\theta - \hat{\theta}_k),$$

where $H_k^{\mathrm{MLE}} := -\nabla_\theta^2 \log \mathcal{L}(\theta; \mathcal{D}_k)|_{\theta=\hat{\theta}_k}$. Hereafter, we drop the superscript and write $H_k := H_k^{\mathrm{MLE}}$ when the expansion center is unambiguous. Marginalizing over θ_k as in Proposition 3.1 gives

$$\log p(\mathcal{D}_k \mid c_k = 1, \theta) \approx \text{const} - \tfrac{1}{2}(\theta - \hat{\theta}_k)^\intercal (I + \tau^2 H_k)^{-1} H_k (\theta - \hat{\theta}_k).$$

This expansion makes the objective quadratic, which admits a closed-form update

$$\theta^{(t+1)} = \Big(I + \sum_{k=1}^{K} \Lambda_k^{(t)}\Big)^{-1} \Big(\hat{\theta}_0 + \sum_{k=1}^{K} \Lambda_k^{(t)} \hat{\theta}_k\Big), \quad \Lambda_k^{(t)} := w_k^{(t)} H_0^{-1} (I + \tau^2 H_k)^{-1} H_k. \tag{12}$$

The update is a relative-precision-weighted blend of the source MLEs that requires no gradient-descent inner loop. Because $\hat{\theta}_k, g_k, H_k$ depend only on $\mathcal{D}_k$ and not on the EM iterate, they are computed *only once and reused* across iterations t.

Second, when τ is a small number, its square is numerically negligible in practice. We further approximate $H_k^{\mathrm{MLE}} \sim N_k \mathcal{I}$ and $H_0^{\mathrm{MLE}} \sim N_0 \mathcal{I}$ using the expected Fisher information, which reduces (12) to a clean weighted average of source MLEs (Appendix C):

$$\theta^{(t+1)} = \frac{N_0 \hat{\theta}_0 + \sum_{k=1}^{K} w_k^{(t)} N_k \, \hat{\theta}_k}{N_0 + \sum_{k=1}^{K} w_k^{(t)} N_k}, \tag{13}$$

For over-parameterized models like neural networks, weight-space averaging is meaningless, so we directly perform gradient ascent on (11). Appendices C and D provide the pseudocode for all variants.

Bayesian Tempering in EM The E-step weight (7) can be unstable in early iterations when $\theta^{(t)}$ is a poor estimate of θ_0 because the log-likelihood ratio in (7) can have high variance and overwhelm the LIP correction (see detailed discussion in Appendix C.2). To mitigate this, we introduce a tempering schedule that suppresses the log-likelihood ratio in early iterations and gradually increases it as the EM iterations approach θ_0. This is achieved by multiplying the log-likelihood ratio by a tempering parameter $\beta_k^{(t)}$ that grows with t and shrinks inversely with the standard deviation of the log-likelihood ratio. As a result, EM relies more on the LIP when the estimate is noisy and delegates to data otherwise. The effect of tempering will be discussed in the theoretical analysis in Sec. 4.

4 Theoretical Results

This section provides two complementary guarantees for LIP-aided EM. First, when the LIP is well-calibrated, the MSE of the surrogate update (13) matches an oracle cold-start MSE (Sec. 4.1). Second, even when the LIP is misspecified, the algorithm remains asymptotically consistent (Sec. 4.2). The first result formalizes the role of LIP in cold start, while the second part serves as a safety net.

4.1 Finite-sample cold-start MSE under a correct LIP

We consider the Gaussian-mean estimation problem where each source k has $N_k = N$ i.i.d. samples from $p_k \triangleq \mathcal{N}(\theta_k, \sigma^2 I)$, the null prior ϕ_{null} determines the null density $q(x) \triangleq p(x \mid c_k = 0)$, and $R \triangleq \{k : c_k = 1\}$, $\bar{R} \triangleq \{k : c_k = 0\}$ denote the relevant and irrelevant sets. Define the per-sample log-LR of the relevant model against the null on a single observation, i.e., $\ell(x; \theta) \triangleq \log \frac{p(x|\theta)}{q(x)}$, and its expectation under p_k at the truth θ_0,

$$\rho_k(\theta^{(t)}) \triangleq \mathbb{E}_{x \sim p_k}\Big[\ell(x; \theta^{(t)})\Big] = \mathrm{KL}(p_k \,\|\, q) - \mathrm{KL}(p_k \,\|\, p(\cdot \mid \theta^{(t)})). \tag{14}$$

Arguably, the overarching *oracle* one can aim for in a cold start identifies all the relevant sources:

$$\theta^\star \triangleq \frac{N_0 \hat{\theta}_0 + N \sum_{k \in R} \hat{\theta}_k}{N_0 + N|R|}. \tag{15}$$

Under our model assumption in Fig. 1, Proposition E.6 in Appendix characterizes the oracle MSE as

$$\mathrm{MSE}_\star \triangleq \mathbb{E}\|\theta^\star - \theta_0\|^2 = \frac{d\sigma^2}{N_0 + N|R|} + \frac{d\tau^2 N^2 |R|}{(N_0 + N|R|)^2}, \tag{16}$$

which is much smaller than the target-only MSE $d\sigma^2/N_0$ when $N|R| \gg N_0$. The second term is the irreducible τ-shift contribution from the relevant cluster spread (variance τ^2 per coordinate, $|R|$ relevant sources averaged). Motivated by (16), the rest of the analysis *specialize to* $\tau = 0$. The general $\tau > 0$ case follows the same arguments with the additive $d\tau^2 N^2|R|/(N_0 + N|R|)^2$ correction in the oracle term and an analogous correction term in the resulting theorem.

Theorem 4.1 below rests on four assumptions, deferred to Appendix E.4: E.1 fixes $\tau = 0$ and assumes a null prior ϕ_{null} with a finite second moment $\bar{D}^2 \triangleq \mathbb{E}_{\theta\sim\phi_{\text{null}}}\|\theta - \theta_0\|^2$; E.2 states Fisher-information regularity and V-sub-Gaussian and L-Lipschitz per-sample log-LR; E.3 requires probabilistic separation $\mathbb{P}_{\theta\sim\phi_{\text{null}}}(\|\theta - \theta_0\| < r_{\text{sep}}) \le \alpha$; and E.4 assumes an identifiability margin $\Delta^* > 0$ such that $\min_k |\rho_k| \ge \Delta^*$ on the well-separated event $\mathcal{S} \triangleq \{\|\theta_k - \theta_0\| \ge r_{\text{sep}} \text{ for all } k \in \bar{R}\}$. With the assumptions, we show that LIP-aided EM matches the oracle MSE up to terms that are exponentially small in N. An unconditional result that removes the conditioning on the well-separated event $\mathcal{S}$ (see Corollary E.13) also holds at an extra price of $K\alpha r_{\text{sep}}^2$.

Theorem 4.1 (Cold-start MSE under a correct LIP)**.** *Under Assumptions E.1–E.4, suppose the EM iterate of* (13) *enters and remains in the basin* $\{\|\theta - \theta_0\| \le c_1\Delta^*/L\}$ *across iterations. Let* $B \triangleq \max_k |\log[\pi_k/(1-\pi_k)]|$. *Then, there exist absolute constants* $c_1, c_2, c_3 > 0$ *such that with probability at least* $1 - K\alpha$ *over the prior draw of* $\{\theta_k\}_{k=1}^K$*, the EM fixed point* $\theta^{(\infty)}$ *satisfies*

$$\mathbb{E}\big\|\theta^{(\infty)} - \theta_0\big\|^2 \;\le\; \underbrace{\frac{2d\sigma^2}{N_0 + N|R|}}_{\text{oracle rate}} + \underbrace{C_1\, e^{-c_2\beta N\Delta^*}}_{\substack{\text{weight-error}\\ \text{residual}}} + \underbrace{C_2\, e^{-c_3 N(\Delta^*)^2/V^2}}_{\substack{\text{concentration}\\ \text{failure}}}, \tag{17}$$

where $C_1 = O\big(K^2(N/N_0)^2\,(\bar{D}^2 + d\sigma^2/N)\,e^{2B}\big)$ *and* $C_2 = O\big(K^{5/2}(N/N_0)^2\,(\bar{D}^2 + d\sigma^2/N)\big)$.

At first glance, one may question the necessity of a correct LIP since the theorem is seemingly unrelated to LIP correctness. The connection lies in the basin-entry hypothesis: the EM iterate must enter $\{\|\theta - \theta_0\| \le c_1\Delta^*/L\}$, and the LIP determines whether the very first M-step does. Under the Bayesian-tempering schedule of Sec. C.2, the E-step weights collapse to $w_k^{(0)} = \pi_k$, and therefore

$$\theta^{(1)} - \theta_0 \;=\; \underbrace{\frac{\sum_{k\in\bar{R}}\pi_k\,(\theta_k - \theta_0)}{N_0/N + \sum_k \pi_k}}_{\substack{\text{bias, } \|\cdot\|=O(1) \text{ in } N}} + \underbrace{\frac{N_0(\hat{\theta}_0 - \theta_0) + \sum_k N\pi_k(\hat{\theta}_k - \theta_k)}{N_0 + N\sum_k \pi_k}}_{\substack{\text{noise: mean 0, } \mathbb{E}\|\cdot\|^2=O(d\sigma^2/N)}}. \tag{18}$$

When the LIP is correct, π_k puts higher emphasis on relevant sources, so the bias is small and noise from the irrelevant sources is also suppressed. As such, under a correct LIP, basin-entry failure decays exponentially in the source size N rather than the target size N_0 (per the last two terms of (17)).

4.2 Asymptotic consistency under any LIP

Under the unfortunate scenario where the LIP is misspecified, such that the EM iterates fail to enter the basin of attraction of θ_0, finite-sample MSE bounds become intractable. That said, asymptotic consistency still holds in two regimes that are robust to arbitrary LIP.

Theorem 4.2 (Consistency as $N_0 \to \infty$)**.** *Fix any iterate* t*, sizes* $\{N_k\}$*, and prior* π*. Under the standard regularity conditions of [Van der Vaart, 1998, Theorem 5.39] for the target log-likelihood, both the exact M-step iterate* (12) *and the surrogate iterate* (13) *satisfy* $\theta^{(t+1)} \xrightarrow{p} \theta_0$ *as* $N_0 \to \infty$.

At $\tau = 0$, $p(\mathcal{D}_k \mid c_k = 1, \theta) = \prod_i p(x_i \mid \theta)$ (Remark 3.2), so the E-step weight reduces to a per-sample log-LR.

Theorem 4.3 (Asymptotic dichotomy of weights)**.** *Fix iteration* t*, source* k*, and prior* $\pi_k \in (0,1)$. *Assume* $\mathbb{E}_{x\sim p_k}|\log p(x \mid \theta^{(t)})| < \infty$ *and* $\mathbb{E}_{x\sim p_k}|\log q(x)| < \infty$*, so that* $\rho_k(\theta^{(t)})$ *in* (14) *is finite. If* $\rho_k(\theta^{(t)}) \neq 0$*, the relevance weight satisfies* $w_k^{(t)} \xrightarrow{p} \mathbf{1}\{\rho_k(\theta^{(t)}) > 0\}$ *as* $N_k \to \infty$.

Theorem 4.3 shows that the relevance weight $w_k^{(t)}$ asymptotically commits to 1 if the relevant model better explains the source data than the null and commits to 0 if the null is better. This also echoes our discussion in the null model selection: when source k is relevant, the second KL divergence eventually vanishes, so any null density $q \neq p_k$ yields $\rho_k(\theta^{(t)}) > 0$ and $w_k^{(t)} \to 1$.

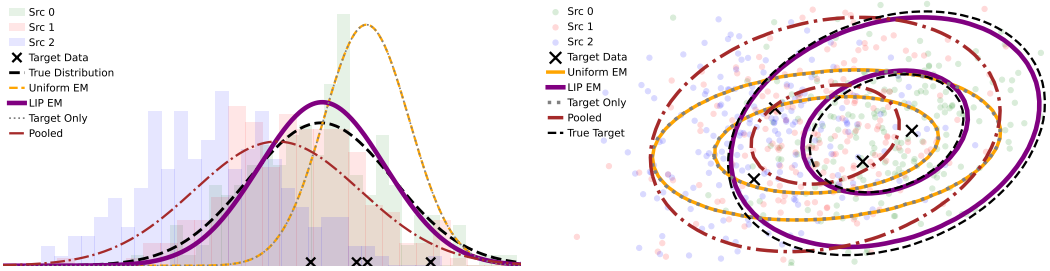
5 Empirical Results

We evaluate our method on three tasks: descriptive, predictive, and prescriptive. Details about benchmark methods and hyperparameter settings are provided in Appendix F.1 and F.2.

5.1 1-d and 2-d Gaussian Estimation

Our first experiment is to estimate Gaussian distributions. Data points are drawn from 1-d (Fig. 3a) and 2-d (Fig. 3b) Gaussian distributions. The target samples are labeled with black crosses. For both experiments, we pick three source domains (plotted as histograms labeled in purple, red, and green), while only the green one matches the target. To simulate the cold-start regime, we take $N_k = 200$ samples from each source and 4 from each target. As a proof of concept, we provide unambiguous descriptions to the LLM, such that any reasonable LLM can label the relevant source. In the 1-d case, we provide the description: "*The target domain has a larger mean*". In the 2-d case, we provide the description: "*The target domain has a larger x mean*".

The results show the significant role of using the LIP. LIP-aided EM (purple curve) effectively identifies the relevant domain with the help of the language prior compared to EM with a uniform prior of p_0 (denoted as Uniform EM, orange curve). As a result, the fitted distribution almost recovers the ground truth (black dashed curve). In comparison, when relying solely on the scarce target data (Target Only, gray curve), the estimator suffers from high variance, yielding a high deviation. Meanwhile, indiscriminately pooling all the data (Pooled, brown curve) suffers from negative transfer and is biased towards irrelevant sources.



(a) 1-d Gaussian mean estimation.

(b) 2-d Gaussian estimation.

Figure 3: Gaussian estimation. LIP-aided EM (in purple) best fits the target (in black dashed curve).

5.2 Case Study on C-MAPSS Dataset

To evaluate our method on real-world physical systems, we apply the LIP-aided EM framework to the NASA Commercial Modular Aero-Propulsion System Simulation (C-MAPSS) dataset FD001 [Saxena et al., 2008]. The dataset consists of 100 run-to-failure time-series trajectories simulating the life cycles of various turbofan engines under heterogeneous operational conditions. The source trajectories vary in length as they terminate when their Remaining Useful Life (RUL) reaches zero. In this case study, the task is to predict the physical core speed (sensor 9) of the High-Pressure Compressor (HPC) over time (measured in flight cycles). Each machine is treated as one domain. Among the 100 machines, we select one engine from the pile that exhibits a fast decay trend as the target domain, with its first N_0 cycles of physical core speed used to predict its physical core speed in the future. The rest of the 99 machines are treated as sources, where we have their full trajectory of physical core speed until RUL reaches zero. We provide the following language prompt that hints at a fast degradation in HPC efficiency: *This aircraft is operating in a desert environment. Routine visual inspections confirm severe aerodynamic degradation of the high-pressure compressor, likely due to continuously ingesting fine, dry abrasives at high velocities.* The original paper [Saxena et al., 2008] that produced this dataset is provided alongside the text description to the LLM to help understand the underlying physics.

The target domains in our 10 replications are labeled in red in Fig. 4. For each selected engine, we test the performance of benchmark algorithms for different percentages of RULs (measured in Root Mean Square Error, RMSE). For example, 70% RUL means the target has only seen the first 30%

of the data, which is a typical cold-start scenario. A Generalized Linear Model (GLM), namely a natural cubic spline, is used as the backbone statistical model for all the benchmark methods [Yue et al., 2024]. GLM naturally admits Gaussian likelihood and does not need approximation, so we use the *no approximation version* (Algorithm 1 in Appendix D).

The experimental results are summarized in Table 1. Here, LIP-G means running EM with LIP generated by Gemini 3 Flash API, and LIP-C means running EM with LIP generated by Claude Opus 4.7 local agent (see Appendix F.4.1). Again, LIP significantly reduced the prediction error during the cold-start phase (RUL$= 90\%, 70\%, 50\%$). We report the mean of the 10 replications, along with their standard error in parentheses. As an example, we plot the predictions of engine 80 at its 70% RUL. At the early stage of a machine RUL, the machine is still operating under relatively healthy conditions, and the degradation trend is statistically hidden by the noise. As such, with only the target data, one might confuse it with a healthy machine and predict relatively stable HPC (see gray, brown, and yellow lines). This indicates the importance of leveraging language as a modality in DA.

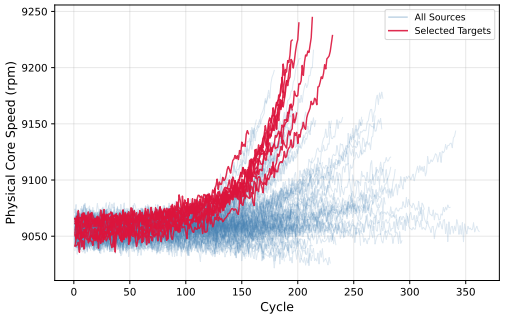


Figure 4: Selected Sources

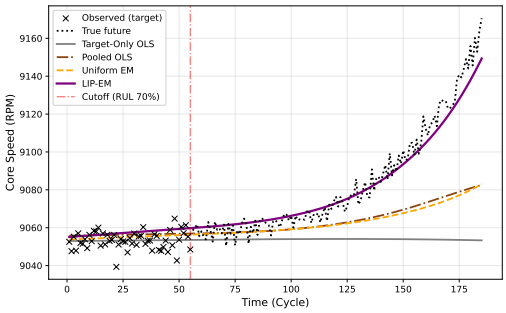


Figure 5: Engine 80 at 70% RUL

Table 1: C-MAPSS core speed RMSE

RUL	LIP-C	LIP-G	Uniform EM	Pooled	Target Only
90%	**14.3 (2.0)**	**15.0 (3.3)**	21.4 (3.0)	33.9 (2.6)	43.7 (2.7)
70%	**15.9 (2.0)**	**15.2 (3.6)**	35.2 (3.1)	38.1 (3.0)	58.7 (10.8)
50%	**17.7 (3.5)**	**17.9 (6.1)**	24.9 (4.4)	44.7 (3.5)	36.6 (5.1)
30%	20.1 (3.2)	**16.5 (3.3)**	**15.2 (2.2)**	56.5 (4.4)	27.0 (5.8)
10%	11.5 (1.6)	9.0 (1.3)	**7.3 (0.8)**	82.9 (6.5)	10.1 (2.7)

Table 2: MuJoCo hopper reward

N_0	LIP-C	LIP-G	Uniform EM	Pooled	Target Only
128	**2670 (57)**	1627 (69)	1283 (34)	1610 (48)	19 (0)
256	**2539 (61)**	1647 (70)	1578 (51)	1298 (32)	24 (4)
512	**2586 (55)**	1760 (76)	**2491 (57)**	1364 (38)	223 (3)
1024	**2599 (57)**	**2636 (56)**	**2604 (55)**	1336 (33)	179 (9)
2048	**2607 (56)**	**2607 (56)**	**2666 (56)**	1496 (42)	164 (13)

5.3 Case Study on MuJoCo Hopper Dynamics

To validate the our framework in deep neural networks, we apply the LIP-aided EM algorithm to a deep reinforcement learning task in the MuJoCo hopper environment [Todorov et al., 2012]. MuJoCo allows users to change the simulation environment of a hopper (a three joint jumping robot). The agent in this task is deployed to a new "planet" (the target environment). It only has a small dataset $\mathcal{D}_0$ collected in the deployment environment and a much larger dataset $\{\mathcal{D}_k\}_{k=1}^{10}$ collected under 10 different gravities $g \in \{1, 2, \ldots, 10\}$ $\mathrm{m/s^2}$. For each source dataset, we collect a *replay* dataset of one million transitions by running soft actor-critic (SAC) [Haarnoja et al., 2018] from scratch and saving the entire training replay buffer. It is worth noting that for the target dataset $\mathcal{D}_0$, not only is the dataset small, but the data quality is also poor in the sense that most of the early explorations of the SAC are falling off, making domain adaptation more critical. The target domain is selected to have gravity similar to that of Venus ($g = 8.87\mathrm{m/s^2}$). We give the target description: *The hopper is deployed to a planet with Venus-like gravity*. We then follow Sec. 3.1 by asking the LLM to select one source from the sampled subgroups with which physics is most consistent with the description.

We chose IQL [Kostrikov et al., 2022] as our backbone algorithm for this experiment due to the well-known fact that model-free RL performs empirically better than model-based RL [Yu et al., 2020] when the offline dataset is abundant. However, model-free RL does not natively provide the likelihood function $\mathcal{L}$ needed for our framework, so we first follow the standard practice of Model-based RL [Yu et al., 2020] and model the dynamics as a Gaussian distribution whose mean and variance are parameterized by a Multi-layer Perceptron, optimizing it using MLE. With the dynamics model, we

run our EM and extract the weights w_k at convergence. Then, in the second step, we run IQL on the w_k-weighted source dataset and the unweighted target dataset to train RL policies. We used Algorithm 3 in Appendix D for the EM iterations. Thanks to the efficient approximation proposed in Appendix C, the EM itself converges within a few minutes on a single GPU (see Appendix F.3).

Table 2 reports the mean episodic rewards over 200 evaluation episodes, along with the standard error. Again, LIP-C achieves a substantially better return in the cold start phase ($N_0 \in \{128, 256\}$). Interestingly, LIP-G uses a false LIP produced due to limited reasoning capabilities (see failure mode in Appendix F.4.2). This false LIP prefers $g = 7\ \mathrm{m/s^2}$ over the closer matches $g \in \{8, 9, 10\}\ \mathrm{m/s^2}$ to Venus's gravity. This false prior helps outperform the uniform prior in the cold start, but at the price of a slower recovery (see $N_0 = 512$). This again emphasizes the quality of the prior, but also showcases the auto-correcting mechanism of EM, echoing our theoretical analysis.

6 Discussion

We have shown that textual descriptions of a target domain carry information that statistical methods cannot recover from scarce target data. A pretrained LLM can translate this information into a Bayesian prior (LIP) over source relevance. Theoretically, a correct LIP yields a finite-sample MSE that matches the oracle precision-weighted rate, while any LIP, even misspecified, still maintains asymptotic consistency via EM. These theoretical claims are verified by our experiments. Nonetheless, the quality of the LIP is crucial for cold-start (see Appendix F.4). Future work includes exploring more sophisticated elicitation methods, relaxing the basin-entry hypothesis in Theorem 4.1, and extending the framework to non-parametric models and to models without explicit likelihood functions.

References

Sivaraman Balakrishnan, Martin J. Wainwright, and Bin Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77–120, 2017.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Alexander Capstick, Rahul Krishnan, and Payam Barnaghi. Autoelicit: Using large language models for expert prior elicitation in predictive modelling. In *Forty-second International Conference on Machine Learning*, 2025. URL `https://openreview.net/forum?id=GekXB58ZS7`.

Paul H Garthwaite, Joseph B Kadane, and Anthony O'Hagan. Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, 100(470):680–701, 2005.

Bowen Gu, Rishi J Desai, Kueiyu Joshua Lin, and Jie Yang. Probabilistic medical predictions of large language models. *npj Digital Medicine*, 7(1):367, 2024.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. Pmlr, 2018.

Judy Hoffman, Brian Kulis, Trevor Darrell, and Kate Saenko. Discovering latent domains for multisource domain adaptation. In *European Conference on Computer Vision (ECCV)*, pages 702–715. Springer, 2012.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213, 2022.

Raed Kontar, Garvesh Raskutti, and Shiyu Zhou. Minimizing negative transfer of knowledge in multivariate gaussian processes: A scalable and regularized approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3508–3522, 2021.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. In *International Conference on Learning Representations*, 2022.

R Duncan Luce et al. *Individual choice behavior*, volume 4. Wiley New York, 1959.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation with multiple sources. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 21, 2008.

Cristina Mata, Kanchana Ranasinghe, and Michael S Ryoo. Copt: Unsupervised domain adaptive segmentation using domain-agnostic text embeddings. In *European conference on computer vision*, pages 424–440. Springer, 2024.

Daniel McFadden. Conditional logit analysis of qualitative choice behavior. In P. Zarembka, editor, *Frontiers in Econometrics*, pages 105–142. Academic Press, New York, 1974.

Anthony O'Hagan, Caitlin E Buck, Alireza Daneshkhah, J Richard Eiser, Paul H Garthwaite, David J Jenkinson, Jeremy E Oakley, and Tim Rakow. *Uncertain judgements: eliciting experts' probabilities*. John Wiley & Sons, 2006.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 2463–2473, 2019.

Abhinav Saxena, Kai Goebel, Don Simon, and Neil Eklund. Damage propagation modeling for aircraft engine run-to-failure simulation. In *2008 international conference on prognostics and health management*, pages 1–9. IEEE, 2008.

Nikola Sekulovski, Meike Waaijers, and Giuseppe Arena. Llm-based prior elicitation for bayesian graphical modeling. *British Journal of Mathematical and Statistical Psychology*, 2026.

Qian Sun, Rita Chattopadhyay, Sethuraman Panchanathan, and Jieping Ye. A two-stage weighting framework for multi-source domain adaptation. *Advances in neural information processing systems*, 24, 2011.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, 2023.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 5026–5033. IEEE, 2012.

Naonori Ueda and Ryohei Nakano. Deterministic annealing EM algorithm. *Neural Networks*, 11(2): 271–282, 1998.

Aad W Van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge University Press, 1998.

Zhenbin Wang, Lei Zhang, Lituan Wang, and Minjuan Zhu. Landa: Language-guided multi-source domain adaptation. *IEEE Transactions on Artificial Intelligence*, 2025.

Zirui Wang, Zihang Dai, Barnabás Póczos, and Jaime Carbonell. Characterizing and avoiding negative transfer. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11293–11302, 2019.

Junfeng Wen, Russell Greiner, and Dale Schuurmans. Domain aggregation networks for multi-source domain adaptation. In *International conference on machine learning*, pages 10214–10224. PMLR, 2020.

Caiming Xiong, Scott McCloskey, Shao-Hang Hsieh, and Jason Corso. Latent domains modeling for visual domain adaptation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 28(1), 2014.

Huaxiu Yao, Xinyu Yang, Xinyi Pan, Shengchao Liu, Pang Wei Koh, and Chelsea Finn. Improving domain generalization with domain relations. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. *Advances in neural information processing systems*, 33:14129–14142, 2020.

Xubo Yue, Raed Al Kontar, and Ana Maria Estrada Gomez. Federated data analytics: A study on linear models. *IISE Transactions*, 56(1):16–28, 2024.

Wen Zhang, Lingfei Deng, Lei Zhang, and Dongrui Wu. A survey on negative transfer. *IEEE/CAA Journal of Automatica Sinica*, 10(2):305–329, 2022.

Han Zhao, Shanghang Zhang, Guanhang Wu, José MF Moura, Joao P Costeira, and Geoffrey J Gordon. Adversarial multiple source domain adaptation. *Advances in neural information processing systems*, 31, 2018.

Sicheng Zhao, Hui Chen, Hu Huang, Pengfei Xu, and Guiguang Ding. More is better: deep domain adaptation with multiple sources. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 8354–8362, 2024.

Yilun Zhu, Naihao Deng, Naichen Shi, Aditya Gangrade, and Clayton Scott. Domain generalization under posterior drift. *arXiv preprint arXiv:2510.04441*, 2026.

A Related Work

Our work sits at an unexplored intersection of three lines, namely multi-source domain adaptation, LLM-based knowledge elicitation, and metadata-aware domain generalization. We are the first to model source relevance as a *binary latent variable* in a Bayesian hierarchical EM, with the prior over this indicator *elicited from a pretrained LLM via subgroup-restricted preference queries*, and using *target-domain context only* as input. The four paragraphs below organize prior work along these three axes of differentiation.

Multi-Source Domain Adaptation. Our problem setup is closest to Multi-Source Domain Adaptation (MDA), which leverages labeled data from K distinct source domains to learn a model for a target domain. A fundamental challenge in MDA is *negative transfer*, where the inclusion of irrelevant or adversarial sources degrades target performance [Pan and Yang, 2009, Wang et al., 2019, Kontar et al., 2021, Zhang et al., 2022, Zhao et al., 2024]. Two prior lines distinguish sources by exploiting target features. The first reweights sources by feature proximity [Sun et al., 2011], moment alignment [Peng et al., 2019], adversarial training [Zhao et al., 2018], generalization bounds [Wen et al., 2020], or theoretically optimal convex combinations [Mansour et al., 2008]. The second uses EM-style alternation to *discover* unknown domain structure within source data [Hoffman et al., 2012, Xiong et al., 2014]. Both lines are data-driven and become unreliable when target data is scarce, since the discrepancies and latent clusters they rely on cannot be reliably estimated from a small target sample. We instead model source relevance as a *binary* latent indicator and place a non-data prior on it, which lets source selection proceed in the cold-start regime where existing methods are bottle-necked by the target sample size.

LLMs as Expert Reasoners. Historically, translating qualitative expert intuition into quantitative probability distributions required extensive manual labor [Garthwaite et al., 2005, O'Hagan et al., 2006]. The premise of leveraging natural language for statistical inference is rooted in treating LLMs as implicit knowledge bases and expert reasoners [Petroni et al., 2019]. Because LLMs are pretrained on a vast corpus of digital textual knowledge, they exhibit strong zero-shot reasoning capabilities across specialized domains [Brown et al., 2020, Kojima et al., 2022]. The closest precursor is AutoElicit [Capstick et al., 2025], which queries an LLM directly for the numerical values of a prior over linear-model parameters. Our elicitation primitive is fundamentally different. We ask the LLM to answer a multiple choice question over subgroups of sources rather than directly producing numerical probabilities. We then fit a conditional-logit-with-null choice model to the resulting preference data to recover the prior. The multiple-choice route exploits the well-documented fact that LLMs are more reliable at discrete classification than at numerical estimation [Gu et al., 2024, Tian et al., 2023], and produces a calibrated prior even when the LLM cannot articulate explicit numbers, as similarly observed in concurrent work using LLM-elicited binary judgments for Bayesian graphical modeling [Sekulovski et al., 2026].

Domain Generalization with Metadata. Closely related to our motivation of using contextual information is Domain Generalization (DG) with metadata. Unlike standard DG, which seeks a purely data-driven domain-invariant representation, metadata-driven DG leverages supplementary domain attributes to capture inter-domain relationships. Zhu et al. [2026] demonstrate the necessity of metadata when the physics of the domain changes, showing that pooling-ERM is provably suboptimal under posterior shift, and incorporate domain metadata as an additional input during training and inference. Other methods such as $\mathrm{D}^3\mathrm{G}$ [Yao et al., 2024] establish domain similarity matrices from structured metadata to reweigh domain-specific models. These works share a structural assumption that we relax: metadata must be consistently available across *all* training domains. In engineering practice, such records exist only for the present target system, while historical source datasets predate the metadata convention. Our framework operates on unstructured natural language descriptions available *only* for the target and delegates the relevance reasoning to the LLM rather than to a similarity computation over fixed metadata fields.

Language-Guided Domain Adaptation. The emergence of Vision-Language Models (VLMs) like CLIP has enabled adaptation methods to use natural language as a bridge between domains. Recent works such as LangDA [Wang et al., 2025] and CoPT [Mata et al., 2024] demonstrate the efficacy of using text descriptions to guide visual alignment by treating text as a feature embedding aligned with image features in a shared space. Two structural commitments distinguish these methods from ours. First, like domain generalization with metadata, they also require text descriptions for source domains as well as for the target, which is impractical when the source datasets predate the natural-language pipeline. Second, this field is built upon massive VLM architectures, making the methods task-specific and restricted to modalities that the VLM was pretrained on. We instead use the LLM as a *reasoning agent* producing discrete preferences, require text only for the target, and remain model-agnostic, applicable to general parameter estimation tasks beyond vision-language alignment.

B LIP Construction via Subgroup Preference Elicitation

B.1 Subgroup Preference Elicitation

Directly prompting an LLM for numerical probabilities π_k can sometimes be unreliable due to the context window constraint, just as it is not an easy task for human experts either. To keep the cognitive task as simple and reliable as possible, we constrain the LLM to select *at most one* candidate per query: which source domain, among a set of candidates, is more likely to satisfy the contextual information of the target domain? Effectively, we are asking a prompted LLM to act as a judge in selecting the relevant source domains. It either returns the single most relevant source from the presented subgroup, or it returns nothing if no candidate satisfies the contextual description. To perform this selection, the LLM is provided with sufficient background knowledge h, which includes the text that describes the target domain and background information about the task, including relevant technical reports.

Formally, we denote the LLM judge equipped with contextual knowledge h as $f_{\text{LLM}}(\cdot\,; h)$. For each query $m = 1, \ldots, N_M$, we sample a subgroup $S_m \subseteq [K]$ of source-domain indices and present the corresponding datasets $\{\mathcal{D}_k\}_{k \in S_m}$ to the LLM. The LLM returns a *choice* $k_m \in S_m \cup \{0\}$, where $k_m \in S_m$ identifies the selected source and $k_m = 0$ denotes the null option (no candidate is considered relevant). We collect these responses into a dataset $M = \{(S_m, k_m)\}_{m=1}^{N_M}$. The choice of subgroup size $|S_m|$ and composition can be altered per application and is independent of our proposed method. In practice, $|S_m|$ is bounded only by the LLM's context window and its ability to reliably compare many candidates at once.

B.2 Empirical Bayes for LIP

Empirical Bayes is a category of methods that defines a prior based on data. With these collected LLM preferences M, we now estimate the prior probability distribution by maximizing its likelihood. To avoid numerical issues in the downstream Bayesian model, we parameterize $\pi_k = \sigma(\alpha_k)$ so that $\pi_k \in (0, 1)$, where α_k is a real-valued latent variable associated with source k.

Because the LLM is restricted to selecting at most one candidate per query, we model its choice as a *conditional logit with an outside option* [McFadden, 1974, Luce et al., 1959]. The null option is a virtual alternative whose worth α_0 represents the threshold above which a candidate is considered relevant: a source k is selected only if its worth α_k exceeds both α_0 and the worths of the other candidates in S in the LLM's judgment. Concretely, the probability that the LLM returns choice $k_m \in S_m \cup \{0\}$ given the presented subgroup S_m is the single softmax (2) reproduced here:

$$p(k_m \mid S_m) = \frac{e^{\alpha_{k_m}}}{e^{\alpha_0} + \sum_{j \in S_m} e^{\alpha_j}},$$

where $\alpha_{k_m} = \alpha_0$ when $k_m = 0$ (the null is selected) and $\alpha_{k_m} = \alpha_k$ when a particular source k is selected.

To find the most probable LIP, we minimize the regularized negative log-likelihood (3):

$$\min_{\alpha_0, \alpha_1, \ldots, \alpha_K} \quad -\sum_{m=1}^{N_M} \log p(k_m \mid S_m) + \epsilon \sum_{k=1}^{K} \left(\alpha_k - \log \frac{p_0}{1 - p_0} \right)^2,$$

where ϵ is an L_2 regularization coefficient and p_0 is the default probability for a uniform prior. The objective is the sum of a convex conditional-logit log-likelihood and a strictly convex quadratic regularizer in $\alpha_1, \ldots, \alpha_K$; together with the conditional-logit Hessian's positive-definiteness in the α_0 direction (which holds whenever M contains at least one null choice and one source choice), the full objective is strongly convex in $(\alpha_0, \alpha_1, \ldots, \alpha_K)$, admitting a unique global optimum. The regularization term serves three functions. First, it provides an anchor for the LIP. The conditional-logit likelihood is invariant under a common additive shift of all α_k (including the null α_0), so $\pi_k = \sigma(\alpha_k)$ is, on its own, gauge-dependent — only the differences $\alpha_k - \alpha_0$ enter the choice probabilities. The quadratic regularizer fixes the gauge by anchoring each α_k near $\log[p_0/(1 - p_0)]$, which makes $\pi_k = \sigma(\alpha_k)$ a meaningful probability that recovers the uniform prior p_0 when M is empty. Second, the regularization prevents numerical instability in the optimization. If a source k wins every subgroup in which it appears, the unregularized problem would drive α_k to infinity. Third, the regularization absorbs occasional noise in the dataset M caused by LLM hallucinations. The null-option worth α_0 is left unregularized and serves as the implicit reference level against which the α_k are measured.

C Practical Implementation of the EM Algorithm

C.1 Hessian Approximation

One natural concern of the described method is its computational complexity. Evaluating the exact posterior weights $w_k^{(t)}$ in the E-step and optimizing the global parameter in the M-step requires the explicit computation and inversion of the local Hessian matrices ($H_k^{(t)}$, H_k, and H_0) for every source domain at every iteration. As the parameter dimension d grows, this $\mathcal{O}(Kd^3)$ operation becomes a significant bottleneck. To circumvent this computational burden, we specialize to $\tau = 0$ throughout

this appendix; for sufficiently small $\tau > 0$, the same formulas hold up to $\mathcal{O}(\tau^2)$ corrections that are numerically negligible.

We start with the E-step. By Remark 3.2, at $\tau = 0$ the second and third terms in the marginal log-likelihood approximation (the quadratic gradient and log-determinant penalty) vanish exactly:

$$\log p(\mathcal{D}_k \mid c_k = 1, \theta^{(t)}) = \log \mathcal{L}(\theta^{(t)}; \mathcal{D}_k). \tag{19}$$

Consequently, the relevance weight (7) reduces to a standard likelihood ratio:

$$w_k^{(t)} = \sigma\left(\log \frac{\mathcal{L}(\theta^{(t)}; \mathcal{D}_k)}{p(\mathcal{D}_k \mid c_k = 0)} + \log \frac{\pi_k}{1 - \pi_k}\right). \tag{20}$$

In the M-step, we leverage the approximation $H_k \approx N_k \mathcal{I}$ and $H_0 \approx N_0 \mathcal{I}$, where $\mathcal{I}$ is the expected per-sample Fisher Information matrix. This simplifies the relative precision matrix to $\Lambda_k^{(t)} \approx w_k^{(t)} \frac{N_k}{N_0}(I + \tau^2 N_k \mathcal{I})^{-1}$. For a sufficiently small τ^2 such that the term $\tau^2 N_k \lambda_{\max}(\mathcal{I}) \ll 1$, the matrix inverse $(I + \tau^2 N_k \mathcal{I})^{-1}$ approaches the identity matrix I. Interestingly, since $(I + \tau^2 N_k \mathcal{I})^{-1} \preceq I$, by approximating $(I + \tau^2 N_k \mathcal{I})^{-1}$ with I, the resulting objective function essentially serves as a lower bound of the original likelihood. The EM algorithm itself can be considered a sub-class of the MM (Minorize-Maximize) algorithm that maximizes an evidence lower bound. The resulting approximation further lowers the evidence lower bound, echoing the same logic of EM algorithm.

Finally, by substituting both the E and M step updates, the global parameter estimation reduces to a simple weighted average of the local MLEs:

$$\theta^{(t+1)} = \frac{N_0 \hat{\theta}_0 + \sum_{k=1}^K w_k^{(t)} N_k \hat{\theta}_k}{N_0 + \sum_{k=1}^K w_k^{(t)} N_k}. \tag{21}$$

This formulation reduces the computational complexity to $\mathcal{O}(d)$, where the influence of each source domain is governed by its sample size N_k and its current relevance $w_k^{(t)}$.

C.2 Bayesian Tempering

Although the convergence of EM iterations is well-established, it only guarantees convergence to a locally optimal solution. In a cold-start regime, the initial estimation of the target parameter $\hat{\theta}_0$ is anchored on an extremely small dataset $\mathcal{D}_0$ of size N_0. When the relevance weights $w_k^{(t)}$ are small, the estimate $\hat{\theta}_0$ depends almost exclusively on this target dataset and suffers from high variance. This volatility may sometimes drive the algorithm into an undesired trivial local optimum where all the relevance weights $w_k^{(t)}$ are zero. To avoid being absorbed by this degenerate solution, we apply a Bayesian tempering mechanism. Structurally, this is in the deterministic-annealing-EM tradition [Ueda and Nakano, 1998], which multiplies the log-likelihood term by a temperature parameter that anneals from low (prior-dominated) to high (data-dominated) over iterations. Our specialization is a per-source variance-calibrated schedule $\beta_k^{(t)} \propto 1/\varepsilon_k$, where ε_k is the per-source cold-start log-LR error scale derived below.

To systematically analyze how the estimation error of $\hat{\theta}_0$ affects the posterior likelihood, we approximate $\log \mathcal{L}(\hat{\theta}_0; \mathcal{D}_k) - \log \mathcal{L}(\theta_0; \mathcal{D}_k)$ by a first-order Taylor expansion around the true parameter θ_0:

$$\left| \log \mathcal{L}(\hat{\theta}_0; \mathcal{D}_k) - \log \mathcal{L}(\theta_0; \mathcal{D}_k) \right| \approx \left| (\hat{\theta}_0 - \theta_0)^\intercal \nabla_\theta \log \mathcal{L}(\theta; \mathcal{D}_k) \big|_{\theta=\theta_0} \right|. \tag{22}$$

This error magnitude is governed by the inner product of two independent random vectors. Assuming the asymptotic normality of the MLE estimator, the target parameter error $(\hat{\theta}_0 - \theta_0)$ has a covariance matrix approximated by the inverse target Hessian, H_0^{-1}. Meanwhile, $\nabla_\theta \log \mathcal{L}(\theta; \mathcal{D}_k)|_{\theta=\theta_0}$ is the score function evaluated at the true parameter, whose covariance is approximated by H_k. As such, the variance of this inner product is given by the trace of their coupled covariances: $\varepsilon_k^2 = \mathrm{Tr}(H_0^{-1} H_k)$.

To evaluate this variance, we may apply the same expected Fisher Information approximation established in Sec. C.1, where $H_k \approx N_k \mathcal{I}$. By substituting $H_0 \approx N_0 \mathcal{I}$ and $H_k \approx N_k \mathcal{I}$, the variance simplifies to $\varepsilon_k^2 \approx d \frac{N_k}{N_0}$.

Therefore, we can see that the variance of the estimation error in the evidence term scales strictly with $\mathcal{O}\left(N_k/N_0\right)$. In the cold-start regime, the source domains are typically much richer than the target domain ($N_k \gg N_0$). Consequently, this un-tempered likelihood error can easily overwhelm the effect of the prior, forcing the algorithm to discard all source domains to minimize variance.

To stabilize this variance and prevent degenerate behavior, we explicitly multiply the likelihood term in the E-step by a tempering parameter $\beta_k^{(t)}$:

$$w_k^{(t)} = \sigma\left(\beta_k^{(t)} \cdot \log \frac{p(\mathcal{D}_k \mid c_k = 1, \theta^{(t)})}{p(\mathcal{D}_k \mid c_k = 0)} + \log \frac{\pi_k}{1-\pi_k}\right), \tag{23}$$

where $\beta_k^{(t)} = \mathcal{O}\big((1 - e^{-\nu t})/\varepsilon_k\big)$ and $\nu > 0$ controls how fast $\beta_k^{(t)}$ converges to its asymptote $1/\varepsilon_k$.

This tempering parameter operates via two parallel mechanisms. The scale $1/\varepsilon_k$ brings the variance of the log-likelihood error back to $\mathcal{O}(1)$. When the target dataset is scarce, this ratio remains small, inherently emphasizing the prior over the volatile data-dependent posterior. For a relatively rich target dataset, the temperature increases and naturally shifts the emphasis back to the data.

Meanwhile, the temporal component, $1 - \exp(-\nu t)$, ensures that the effect of the likelihood is released gradually. When t is small, the $\beta_k^{(t)}$ is also small, so the likelihood relies more on the prior. This protects the EM algorithm from being absorbed by the trivial solution. As the EM loop iterates, this temporal component disappears exponentially fast, recovering the scale ε_k for sufficiently large t. Pseudocode for the exact, small-τ approximate, and neural-network variants of the procedure is given in Appendix D.

D LIP-aided EM Algorithms

D.1 EM Derivation

Using the Bernoulli form $\mathbb{P}(c_k) = \pi_k^{c_k}(1-\pi_k)^{1-c_k}$, the complete-data log-likelihood splits as

$$\begin{aligned}\log p(\mathbf{O}, \mathbf{c} \mid \theta) = &\underbrace{\log \mathcal{L}(\theta; \mathcal{D}_0) + \sum_{k=1}^{K} c_k \log p(\mathcal{D}_k \mid c_k = 1, \theta)}_{\text{terms involving } \theta} \\ &+ \underbrace{\sum_{k=1}^{K} (1-c_k) \log p(\mathcal{D}_k \mid c_k = 0) + c_k \log \pi_k + (1-c_k)\log(1-\pi_k)}_{\text{terms not involving } \theta}. \end{aligned} \tag{24}$$

Dropping the θ-independent terms in (24) reduces (5) to (6) (reproduced below for convenience):

$$\max_{\theta} \ \log \mathcal{L}(\theta; \mathcal{D}_0) + \sum_{k=1}^{K} w_k^{(t)} \log p(\mathcal{D}_k \mid c_k = 1, \theta).$$

D.2 Heuristics for Null Likelihood Selection

This analysis in 3.2.2 motivates two practical choices.

Empirical Bayes with Monte Carlo When source domains are abundant and assumed to be drawn from ϕ_{null}, we approximate ϕ_{null} by the empirical distribution of source MLEs and use the corresponding mixture as the null density:

$$p(x \mid c_k = 0) \approx \frac{1}{K-1} \sum_{j \neq k} (1 - w_j)\, \mathcal{L}(\hat{\theta}_j; x), \tag{25}$$

where $\hat{\theta}_k \triangleq \arg\max_{\theta_k} \mathcal{L}(\theta_k; \mathcal{D}_k)$ is the per-source MLE. As a Monte-Carlo estimator, this option is more fragile when the sources are few or contaminated by relevant ones.

Parametric Null When sources are few or potentially contaminated, we fit a parametric pooled model $\theta^{\mathrm{pool}} \triangleq \arg\max_\theta \mathcal{L}(\theta; \bigcup_k \mathcal{D}_k)$ on the combined source data and set $p(x \mid c_k = 0) \triangleq \mathcal{L}(\theta^{\mathrm{pool}}; x)$. The resulting null is a smoothed version of the source empirical distribution, and the log-ratio in (7) then measures how much better $\theta^{(t)}$ explains $\mathcal{D}_k$ than the pooled model does. A source that fails this test adds no value beyond the pooled model and is safely excluded.

D.3 Pseudocode Algorithm

We collect three pseudocode versions of the procedure described in Sec. 3.2 and Sec. C.2: the closed-form Hessian-reuse implementation (Algorithm 1), the small-τ approximation that admits a closed-form M-step (Algorithm 2), and the neural-network variant that replaces the closed form with gradient ascent (Algorithm 3).

Hessian reuse across iterations. For tractability, Algorithm 1 computes Hessians once at the local MLEs $\hat{\theta}_k$ and reuses them across iterations: it identifies the iteration-dependent $H_k^{(t)}$ in (9) with the frozen $H_k = -\nabla^2_\theta \log \mathcal{L}(\theta; \mathcal{D}_k)\big|_{\hat{\theta}_k}$. This identification is exact at $\theta^{(t)} = \hat{\theta}_k$ and an $O(\|\theta^{(t)} - \hat{\theta}_k\|)$ approximation in its neighborhood; the same identification is built into the closed-form M-step (12). A truly Hessian-recomputing variant would redo the Hessian inversion at every iterate at $\mathcal{O}(Kd^3)$ per step.

Algorithm 1 LIP-aided EM with Bayesian Tempering (Hessian-reuse implementation)

Require: Target data $\mathcal{D}_0$, source data $\{\mathcal{D}_k\}_{k=1}^K$, LIP π, null prior ϕ_{null}, prior variance τ, tempering rate ν

Compute target MLE $\hat{\theta}_0 = \arg\max_\theta \mathcal{L}(\theta; \mathcal{D}_0)$ and Hessian $H_0 = -\nabla^2_\theta \log \mathcal{L}(\theta; \mathcal{D}_0)\big|_{\theta=\hat{\theta}_0}$
for $k = 1, \ldots, K$ **do**
  Compute source MLE $\hat{\theta}_k = \arg\max_\theta \mathcal{L}(\theta; \mathcal{D}_k)$
  Compute source Hessian $H_k = -\nabla^2_\theta \log \mathcal{L}(\theta; \mathcal{D}_k)\big|_{\theta=\hat{\theta}_k}$ *// frozen and reused across iterations*
  Compute tempering scale $\varepsilon_k^2 = \mathrm{Tr}(H_0^{-1} H_k)$
end for
Initialize $\theta^{(0)} \leftarrow \vec{0}$, $t \leftarrow 0$ *// at $t = 0$, $\beta_k^{(0)} = 0$ gives $w_k^{(0)} = \pi_k$ regardless of $\theta^{(0)}$*
while not converged **do**
  Tempering schedule: $\beta_k^{(t)} \leftarrow (1 - e^{-\nu t})/\varepsilon_k$ for each k
  ▷ *E-step*
  for $k = 1, \ldots, K$ **do**
    Compute relevant likelihood $p(\mathcal{D}_k \mid c_k = 1, \theta^{(t)})$ via (9)
    Compute null likelihood $p(\mathcal{D}_k \mid c_k = 0)$ via ϕ_{null}
    Compute relevance weight $w_k^{(t)}$ via (23)
  end for
  ▷ *M-step*
  Update the global parameter $\theta^{(t+1)}$ via (12)
  $t \leftarrow t + 1$
end while
return $\theta^{(t)}$

Algorithm 2 LIP-aided EM with Bayesian Tempering and small-τ approximation

Require: Target data $\mathcal{D}_0$, source data $\{\mathcal{D}_k\}_{k=1}^K$, LIP π, null density $p(\cdot \mid c_k = 0)$, tempering rate ν

Compute target MLE $\hat{\theta}_0 = \arg\max_\theta \mathcal{L}(\theta; \mathcal{D}_0)$
for $k = 1, \ldots, K$ **do**
  Compute source MLE $\hat{\theta}_k = \arg\max_\theta \mathcal{L}(\theta; \mathcal{D}_k)$
  Set tempering scale $\varepsilon_k^2 = d\, N_k / N_0$ *// assumes $H_k \approx N_k \mathcal{I}$ and $H_0 \approx N_0 \mathcal{I}$ with a common $\mathcal{I}$*
end for
Initialize $\theta^{(0)} \leftarrow \vec{0}$, $t \leftarrow 0$
while not converged **do**
  Tempering schedule: $\beta_k^{(t)} \leftarrow (1 - e^{-\nu t})/\varepsilon_k$ for each k
  ▷ *E-step*
  for $k = 1, \ldots, K$ **do**
    $w_k^{(t)} \leftarrow \sigma\Big(\beta_k^{(t)} \log \frac{\mathcal{L}(\theta^{(t)}; \mathcal{D}_k)}{p(\mathcal{D}_k \mid c_k = 0)} + \log \frac{\pi_k}{1 - \pi_k}\Big)$
  end for
  ▷ *M-step*
  $$\theta^{(t+1)} \leftarrow \frac{N_0\, \hat{\theta}_0 + \sum_{k=1}^K w_k^{(t)} N_k\, \hat{\theta}_k}{N_0 + \sum_{k=1}^K w_k^{(t)} N_k} \qquad (13)$$
  $t \leftarrow t + 1$
end while
return $\theta^{(t)}$

Algorithm 3 LIP-aided EM for Neural Networks (gradient-based M-step)

Require: Target data $\mathcal{D}_0$, source data $\{\mathcal{D}_k\}_{k=1}^K$, LIP π, tempering rate ν, M-step optimizer SGD

Train pooled model $\theta^{\mathrm{pool}} \leftarrow \arg\max_\theta \mathcal{L}\left(\theta; \bigcup_{k=1}^K \mathcal{D}_k\right)$ on the combined source data
for $k = 1, \ldots, K$ **do**
  Pre-compute null likelihood $p(\mathcal{D}_k \mid c_k = 0) \leftarrow \mathcal{L}(\theta^{\mathrm{pool}}; \mathcal{D}_k)$
  Set tempering scale $\varepsilon_k^2 = d\, N_k / N_0$ *// assumes $H_k \approx N_k \mathcal{I}$ and $H_0 \approx N_0 \mathcal{I}$ with a common $\mathcal{I}$*
end for
Initialize $\theta^{(0)} \leftarrow \theta^{\mathrm{pool}}$, $t \leftarrow 0$
while not converged **do**
  Tempering schedule: $\beta_k^{(t)} \leftarrow (1 - e^{-\nu t})/\varepsilon_k$ for each k
  ▷ *E-step*
  for $k = 1, \ldots, K$ **do**
    $w_k^{(t)} \leftarrow \sigma\Big(\beta_k^{(t)} \log \frac{\mathcal{L}(\theta^{(t)}; \mathcal{D}_k)}{p(\mathcal{D}_k \mid c_k = 0)} + \log \frac{\pi_k}{1 - \pi_k}\Big)$
  end for
  ▷ *M-step (gradient ascent on the weighted log-likelihood)*
  $\theta^{(t+1)} \leftarrow \mathrm{SGD}\left(\theta^{(t)};\ \log \mathcal{L}(\theta; \mathcal{D}_0) + \sum_{k=1}^K w_k^{(t)} \log \mathcal{L}(\theta; \mathcal{D}_k)\right)$
  $t \leftarrow t + 1$
end while
return $\theta^{(t)}$

E Proof of Theorems

E.1 Proof of Proposition 3.1

Proof. The marginal-likelihood integral is

$$p(\mathcal{D}_k \mid c_k = 1, \theta^{(t)}) = \int p(\mathcal{D}_k \mid \theta_k)\, \mathcal{N}(\theta_k \mid \theta^{(t)}, \tau^2 I)\, d\theta_k. \qquad (26)$$

Exponentiating (8),

$$p(\mathcal{D}_k \mid \theta_k) \approx \mathcal{L}(\theta^{(t)}; \mathcal{D}_k) \exp\left(g_k^{(t)\intercal}(\theta_k - \theta^{(t)}) - \tfrac{1}{2}(\theta_k - \theta^{(t)})^\intercal H_k^{(t)}(\theta_k - \theta^{(t)})\right),$$

substituting into (26), and pulling the constant $\mathcal{L}(\theta^{(t)}; \mathcal{D}_k)$ outside the integral gives

$$p(\mathcal{D}_k \mid c_k = 1, \theta^{(t)}) \approx \mathcal{L}(\theta^{(t)}; \mathcal{D}_k) \int \exp\left(g_k^{(t)\intercal}(\theta_k - \theta^{(t)}) - \tfrac{1}{2}(\theta_k - \theta^{(t)})^\intercal H_k^{(t)}(\theta_k - \theta^{(t)})\right) \mathcal{N}(\theta_k \mid \theta^{(t)}, \tau^2 I)\, d\theta_k. \tag{27}$$

Let $u \triangleq \theta_k - \theta^{(t)}$, which has Jacobian 1. Writing the Gaussian density explicitly as $\mathcal{N}(\theta_k \mid \theta^{(t)}, \tau^2 I) = (2\pi\tau^2)^{-d/2} \exp(-\frac{1}{2\tau^2} u^\intercal u)$, the integrand of (27) becomes

$$(2\pi\tau^2)^{-d/2} \exp\left(g_k^{(t)\intercal} u - \tfrac{1}{2} u^\intercal H_k^{(t)} u - \tfrac{1}{2\tau^2} u^\intercal u\right).$$

The two quadratic terms combine into a single quadratic form via the algebraic identity $u^\intercal H_k^{(t)} u + \tau^{-2} u^\intercal u = u^\intercal A\, u$ with

$$A \triangleq H_k^{(t)} + \tau^{-2} I.$$

Since $H_k^{(t)} \succeq 0$ by hypothesis and $\tau^{-2} I \succ 0$, the sum $A \succ 0$. Equation (27) therefore becomes

$$p(\mathcal{D}_k \mid c_k = 1, \theta^{(t)}) \approx \mathcal{L}(\theta^{(t)}; \mathcal{D}_k)\, (2\pi\tau^2)^{-d/2} \int_{\mathbb{R}^d} \exp\left(g_k^{(t)\intercal} u - \tfrac{1}{2} u^\intercal A\, u\right) du. \tag{28}$$

For any $A \succ 0$ and $b \in \mathbb{R}^d$, completing the square in $u \mapsto u - A^{-1}b$ gives

$$\int_{\mathbb{R}^d} \exp\left(b^\intercal u - \tfrac{1}{2} u^\intercal A u\right) du = (2\pi)^{d/2} (\det A)^{-1/2} \exp\left(\tfrac{1}{2} b^\intercal A^{-1} b\right), \tag{29}$$

where the prefactor $(2\pi)^{d/2}(\det A)^{-1/2}$ is the normalizing constant of the density $\mathcal{N}(A^{-1}b, A^{-1})$. Applying (29) with $b = g_k^{(t)}$ to (28),

$$p(\mathcal{D}_k \mid c_k = 1, \theta^{(t)}) \approx \mathcal{L}(\theta^{(t)}; \mathcal{D}_k)\, (\tau^2)^{-d/2} (\det A)^{-1/2} \exp\left(\tfrac{1}{2} g_k^{(t)\intercal} A^{-1} g_k^{(t)}\right), \tag{30}$$

where the $(2\pi)^{d/2}$ factors from the prefactor in (28) and from (29) cancel.

Using $\det(\alpha M) = \alpha^d \det M$ for a scalar α and $d \times d$ matrix M,

$$(\tau^2)^{-d/2} (\det A)^{-1/2} = \det\left(\tau^2 A\right)^{-1/2} = \det\left(\tau^2 H_k^{(t)} + I\right)^{-1/2}, \tag{31}$$

where the last equality uses $\tau^2 A = \tau^2 H_k^{(t)} + I$.

Factoring τ^{-2} out of A,

$$A = \tau^{-2}(I + \tau^2 H_k^{(t)}),$$

so by $(\alpha M)^{-1} = \alpha^{-1} M^{-1}$,

$$A^{-1} = \tau^2 \left(I + \tau^2 H_k^{(t)}\right)^{-1}. \tag{32}$$

Inserting (31) and (32) into (30),

$$p(\mathcal{D}_k \mid c_k = 1, \theta^{(t)}) \approx \mathcal{L}(\theta^{(t)}; \mathcal{D}_k)\, \det\left(I + \tau^2 H_k^{(t)}\right)^{-1/2} \exp\left(\tfrac{\tau^2}{2}\, g_k^{(t)\intercal} \left(I + \tau^2 H_k^{(t)}\right)^{-1} g_k^{(t)}\right).$$

Taking the natural logarithm of both sides yields (9).

The only approximation enters through (8). If that holds with equality, so does (9). □

E.2 Proof of Remark 3.2

Proof. We evaluate both sides of (9) at $\tau = 0$.

The prior $\mathcal{N}(\theta^{(t)}, \tau^2 I)$ at $\tau = 0$ is the Dirac measure $\delta_{\theta^{(t)}}$. Substituting into the marginal-likelihood integral and using the defining property of the Dirac measure,

$$p(\mathcal{D}_k \mid c_k = 1, \theta^{(t)}) = \int p(\mathcal{D}_k \mid \theta_k)\, \delta_{\theta^{(t)}}(\theta_k)\, d\theta_k = p(\mathcal{D}_k \mid \theta^{(t)}) = \mathcal{L}(\theta^{(t)}; \mathcal{D}_k).$$

Taking the log gives $\log \mathcal{L}(\theta^{(t)}; \mathcal{D}_k)$.

The two correction terms in (9) both vanish at $\tau = 0$:

- The quadratic term carries a multiplicative factor τ^2: substituting $\tau = 0$ algebraically,

$$\tfrac{\tau^2}{2}\, g_k^{(t)\intercal} \left(I + \tau^2 H_k^{(t)}\right)^{-1} g_k^{(t)} \Big|_{\tau=0} = 0 \cdot g_k^{(t)\intercal} I^{-1} g_k^{(t)} = 0.$$

- The log-determinant evaluates to

$$-\tfrac{1}{2} \log\det\left(I + \tau^2 H_k^{(t)}\right) \Big|_{\tau=0} = -\tfrac{1}{2} \log\det(I) = -\tfrac{1}{2} \log 1 = 0.$$

The right-hand side therefore reduces to $\log \mathcal{L}(\theta^{(t)}; \mathcal{D}_k)$, which equals the left-hand side. Hence, (9) holds with equality. □

E.3 Proof of Remark 3.3

Proof. Under the Gaussian likelihood model $p(x \mid \theta_k) = \mathcal{N}(x;\, \theta_k, \Sigma_k)$ with fixed $\Sigma_k \succ 0$, the per-sample log-density is

$$\log p(x_i \mid \theta_k) = -\tfrac{1}{2}(x_i - \theta_k)^\intercal \Sigma_k^{-1} (x_i - \theta_k) - \tfrac{1}{2} \log\det(2\pi\Sigma_k).$$

The first term is a quadratic polynomial in θ_k and the second is constant in θ_k. Summing over the N_k samples preserves the quadratic structure:

$$\log \mathcal{L}(\theta_k;\, \mathcal{D}_k) = -\tfrac{1}{2} \sum_{i=1}^{N_k} (x_i - \theta_k)^\intercal \Sigma_k^{-1} (x_i - \theta_k) + C_k,$$

where C_k is a constant independent of θ_k.

The function $\theta_k \mapsto \log \mathcal{L}(\theta_k;\, \mathcal{D}_k)$ is a polynomial of total degree at most 2. By Taylor's theorem with the Lagrange remainder, the second-order Taylor expansion of a C^∞ function at any expansion point $\theta^{(t)}$ has remainder

$$R_2(\theta_k) = \tfrac{1}{6} \sum_{|\alpha|=3} \partial^\alpha \log \mathcal{L}(\xi;\, \mathcal{D}_k)\, (\theta_k - \theta^{(t)})^\alpha$$

for some ξ on the segment between $\theta^{(t)}$ and θ_k. Since $\log \mathcal{L}(\theta_k;\, \mathcal{D}_k)$ is a polynomial of degree ≤ 2, all third- and higher-order partial derivatives vanish identically, so $R_2 \equiv 0$. The expansion (8) therefore holds with equality at every $\theta^{(t)}$. □

E.4 Setup, Statements, and Proofs for the Finite-Sample Analysis

This appendix collects the supporting setup, assumptions, lemmas, and proofs for the finite-sample bound in Sec. 4.1. Throughout, each source k has equal sample size $N_k = N$ and data $\mathcal{D}_k = \{x_i^{(k)}\}_{i=1}^N$ drawn i.i.d. from $p_k \triangleq \mathcal{N}(\theta_k, \sigma^2 I)$. The null density $q(x) \triangleq p(x \mid c_k = 0) = \int p(x \mid \theta_k, c_k = 0)\, \phi_{\text{null}}(\theta_k)\, d\theta_k$ is determined by ϕ_{null} and treated as a known function of x. Let σ^2 be the per-sample noise variance, $R \triangleq \{k : c_k = 1\}$ the relevant set, $\bar{R} \triangleq \{k : c_k = 0\}$ the irrelevant set, $\Delta_k \triangleq \theta_k - \theta_0$ for $k \in \bar{R}$, and $\Delta_{\max} \triangleq \max_{k \in \bar{R}} \|\Delta_k\|$.

E.4.1 Assumptions

Assumption E.1 (Generative model with $\tau = 0$)**.** The source parameters satisfy $\theta_k = \theta_0$ for $k \in R$ ($c_k = 1$) and $\theta_k \mid c_k = 0 \sim \phi_{\text{null}}$ for a known density ϕ_{null} on $\mathbb{R}^d$ with finite second moment around θ_0:

$$\bar{D}^2 \;\triangleq\; \mathbb{E}_{\theta\sim\phi_{\text{null}}} \|\theta - \theta_0\|^2 \;<\; \infty. \tag{33}$$

Conditional on θ_k, the data $\mathcal{D}_k = \{x_i^{(k)}\}_{i=1}^N$ is i.i.d. from $p_k = \mathcal{N}(\theta_k, \sigma^2 I)$. (This is the $\tau = 0$ specialization of the general $\mathcal{N}(\theta_0, \tau^2 I)$ relevant-prior model, justified by the oracle Mean Squared Error (MSE) motivation in Sec. 4.1; the general case adds an additive $d\tau^2 N^2|R|/(N_0 + N|R|)^2$ correction throughout.)

Assumption E.2 (Regularity)**.** The Fisher information $\mathcal{I}(\theta_0)$ is positive definite with eigenvalues bounded by $0 < \lambda_{\min}(\mathcal{I}) \le \lambda_{\max}(\mathcal{I}) < \infty$. The per-sample log-likelihood ratio $\ell(x;\theta) \triangleq \log[p(x \mid \theta)/q(x)]$ satisfies, on a neighborhood of θ_0: (i) for each k and each θ, $\ell(x;\theta) - \rho_k(\theta)$ is sub-Gaussian under p_k with parameter V^2; (ii) $\theta \mapsto \rho_k(\theta) \triangleq \mathbb{E}_{x\sim p_k}[\ell(x;\theta)]$ is L-Lipschitz for every k.

Assumption E.3 (Probabilistic separation of the null prior)**.** There exist a separation radius $r_{\text{sep}} > 0$ and a tail probability $\alpha \in [0, 1]$ such that

$$\mathbb{P}_{\theta\sim\phi_{\text{null}}}\big(\|\theta - \theta_0\| < r_{\text{sep}}\big) \;\le\; \alpha. \tag{34}$$

Assumption E.4 (Identifiability margin)**.** There exists a margin $\Delta^* > 0$ such that, on the well-separated event $\mathcal{S} \triangleq \{\|\theta_k - \theta_0\| \ge r_{\text{sep}} \text{ for all } k \in \bar{R}\}$, the per-source expected log-LR satisfies

$$\min_{k=1,\dots,K} |\rho_k(\theta_0)| \;\ge\; \Delta^*. \tag{35}$$

Remark E.5 (Sufficient condition via Fisher expansion). For the Gaussian-mean model with $p(\cdot|\theta_0) = \mathcal{N}(\theta_0, \sigma^2 I)$, the Fisher-information expansion gives $\mathrm{KL}(p_k\|p(\cdot|\theta_0)) = \frac{1}{2}(\theta_k - \theta_0)^\intercal \mathcal{I}(\theta_0)(\theta_k - \theta_0) + o(\|\theta_k - \theta_0\|^2)$. Assume two-sided KL bounds on the null density that distinguish relevant and irrelevant sources:

- (*Relevant lower bound*) For $k \in R$ (where $\theta_k = \theta_0$), $\mathrm{KL}(p_k\|q) \ge \kappa_R$ for some $\kappa_R > 0$.
- (*Irrelevant upper bound*) For $k \in \bar{R}$ on $\mathcal{S}$, $\mathrm{KL}(p_k\|q) \le \kappa_{\bar{R}}$ for some $\kappa_{\bar{R}} \ge 0$.

The relevant lower bound says q does not match the relevant likelihood (i.e., q is far from $p(\cdot|\theta_0)$ in KL); the irrelevant upper bound says q does fit the dispersed irrelevant sources at least to within $\kappa_{\bar{R}}$. Then $\rho_k(\theta_0) \ge \kappa_R$ for $k \in R$ (since $\mathrm{KL}(p_k\|p(\cdot|\theta_0)) = 0$), and for $k \in \bar{R}$ on $\mathcal{S}$:

$$\rho_k(\theta_0) \;\le\; \kappa_{\bar{R}} - \tfrac{1}{2} r_{\text{sep}}^2 \, \lambda_{\min}(\mathcal{I}).$$

The identifiability margin $\min_k |\rho_k(\theta_0)| \ge \Delta^*$ therefore holds with

$$\Delta^* \;=\; \min\Big(\kappa_R,\; \tfrac{1}{2} r_{\text{sep}}^2 \, \lambda_{\min}(\mathcal{I}) - \kappa_{\bar{R}}\Big) \;>\; 0$$

whenever $r_{\text{sep}}^2 \, \lambda_{\min}(\mathcal{I}) > 2\kappa_{\bar{R}}$. This requires the separation radius to be large relative to the null's worst-case fit on irrelevant sources. For $\tau > 0$ the relevant-side bound becomes $\kappa_R - \frac{1}{2}\tau^2 d\lambda_{\max}(\mathcal{I})$, with the additional requirement $\kappa_R > \frac{1}{2}\tau^2 d\lambda_{\max}(\mathcal{I})$.

E.4.2 Notation for the analysis

Define the per-sample log-likelihood ratio and its empirical and population averages:

$$\ell(x;\theta) \triangleq \log \frac{p(x \mid \theta)}{q(x)}, \qquad \bar{s}_k(\theta) \triangleq \frac{1}{N}\sum_{i=1}^N \ell(x_i^{(k)};\theta), \qquad \rho_k(\theta) \triangleq \mathbb{E}_{x\sim p_k}[\ell(x;\theta)].$$

We abuse notation and set $\rho_k \triangleq \rho_k(\theta_0)$, the population per-sample log-LR at the truth, which matches (14). Under Assumption E.1 ($\tau = 0$), the joint marginal $p(\cdot \mid c_k = 1, \theta)$ equals the per-sample density $p(x \mid \theta)$ exactly by Remark 3.2.

Tempering scalar. The body uses a per-source per-iteration tempering schedule $\beta_k^{(t)}$ (Sec. C.2). For the analysis below, we treat β as a single scalar equal to the asymptotic limit of $\beta_k^{(t)}$ to be consistent with the asymptotic analysis of basin entry and weight concentration. Per-source variations $\beta_k^{(t)} \to \beta$ enter only through model-dependent constants C_1, C_2.

The total log-LR is $s_k(\theta) \triangleq N\bar{s}_k(\theta) = \log[p(\mathcal{D}_k \mid \theta)/q(\mathcal{D}_k)]$. Under the Gaussian relevant likelihood, $s_k(\theta) = -\frac{N}{2\sigma^2}\|\hat{\theta}_k - \theta\|^2 + \xi_k$ for some θ-independent ξ_k depending on $\mathcal{D}_k$ and q. The surrogate update (13) specializes to

$$w_k^{(t)} = \sigma\Big(\beta\, s_k(\theta^{(t)}) + \log \tfrac{\pi_k}{1-\pi_k}\Big), \tag{36}$$

$$\theta^{(t+1)} = \frac{N_0\hat{\theta}_0 + N\sum_k w_k^{(t)}\hat{\theta}_k}{T^{(t)}}, \qquad T^{(t)} \triangleq N_0 + N\sum_k w_k^{(t)}, \tag{37}$$

and the oracle iterate $\theta^\star$ is defined in (15).

E.4.3 Oracle MSE $\mathrm{MSE}_\star$: a finite-sample bias–variance identity

Proposition E.6 (Conditional MSE of the surrogate update with fixed weights)**.** *For deterministic weights $w \in [0,1]^K$ and any source parameters $\{\theta_k\}_{k=1}^K$, the surrogate update* (37) *satisfies*

$$\mathbb{E}\Big[\|\theta^{(1)} - \theta_0\|^2 \mid \{\theta_k\}\Big] = \frac{d\sigma^2\, N_{\mathrm{eff}}}{T^2} + \Big\|\frac{\sum_{k=1}^K w_k\, N\,(\theta_k - \theta_0)}{T}\Big\|^2, \tag{38}$$

where $T \triangleq N_0 + N\sum_k w_k$ and $N_{\mathrm{eff}} \triangleq N_0 + N\sum_k w_k^2$.

Proof. Decompose $\theta^{(1)} - \theta_0$ as

$$\theta^{(1)} - \theta_0 = \frac{1}{T}\Big[N_0(\hat{\theta}_0 - \theta_0) + \sum_{k=1}^K w_k\, N\,(\hat{\theta}_k - \theta_0)\Big].$$

Conditional on θ_k, $\hat{\theta}_k - \theta_0 = (\hat{\theta}_k - \theta_k) + (\theta_k - \theta_0)$ has mean $\theta_k - \theta_0$ and covariance $\sigma^2 I/N$; the target term has mean zero and covariance $\sigma^2 I/N_0$. Independence across sources (and target) yields

$$\mathbb{E}\Big[\theta^{(1)} - \theta_0 \mid \{\theta_k\}\Big] = \frac{1}{T}\sum_{k=1}^K w_k\, N\,(\theta_k - \theta_0),$$

$$\mathrm{Cov}\Big(\theta^{(1)} \mid \{\theta_k\}\Big) = \frac{1}{T^2}\Big[N_0^2 \cdot \tfrac{\sigma^2 I}{N_0} + \sum_k (w_k N)^2 \cdot \tfrac{\sigma^2 I}{N}\Big] = \frac{\sigma^2\, N_{\mathrm{eff}}}{T^2}\, I.$$

Summing trace and squared conditional bias yields (38). □

Specializing to oracle weights $w_k = \mathbf{1}_R(k)$ gives $N_{\mathrm{eff}} = T = N_0 + N|R|$. Under Assumption E.1 ($\tau = 0$), $\theta_k = \theta_0$ for $k \in R$, so the bias term vanishes and the oracle MSE collapses to the standard precision-weighted rate:

$$\mathrm{MSE}_\star = \frac{d\sigma^2}{N_0 + N|R|}. \tag{39}$$

For the general $\tau > 0$ case, the prior expectation $\mathbb{E}\|\sum_{k\in R}(\theta_k - \theta_0)\|^2 = |R|\tau^2 d$ adds a $d\tau^2 N^2|R|/(N_0 + N|R|)^2$ correction, recovering (16).

E.4.4 Per-step decomposition with random weights

Let $\gamma_k^{(t)} \triangleq w_k^{(t)} - \mathbf{1}_R(k)$ denote the deviation of the EM weights from the oracle weights. The following identity reduces the analysis of the EM iterate to controlling $\mathbb{E}[(\gamma_k^{(t)})^2]$.

Lemma E.7 (Oracle decomposition)**.** *For every $t \geq 0$,*

$$\theta^{(t+1)} - \theta^\star = \frac{N}{T^{(t)}} \sum_{k=1}^{K} \gamma_k^{(t)} \left(\hat{\theta}_k - \theta^\star\right). \tag{40}$$

Consequently, with $T^{(t)} \geq N_0$,

$$\mathbb{E}\left\|\theta^{(t+1)} - \theta_0\right\|^2 \;\leq\; 2\,\mathrm{MSE}_\star \;+\; 2K \left(\frac{N}{N_0}\right)^2 \sum_{k=1}^{K} \mathbb{E}\left[(\gamma_k^{(t)})^2 \left\|\hat{\theta}_k - \theta^\star\right\|^2\right]. \tag{41}$$

Proof. Let $\gamma_k \triangleq w_k^{(t)} - \mathbf{1}_R(k)$ and $T_R \triangleq N_0 + N|R|$, so $T^{(t)} = T_R + N\sum_k \gamma_k$. Multiply (37) by $T^{(t)}$ and (15) by T_R:

$$T^{(t)}\theta^{(t+1)} = N_0\hat{\theta}_0 + N\sum_k w_k^{(t)}\hat{\theta}_k, \qquad T_R\theta^\star = N_0\hat{\theta}_0 + N\sum_{k\in R}\hat{\theta}_k.$$

Subtract:

$$T^{(t)}\theta^{(t+1)} - T_R\theta^\star = N\sum_k \gamma_k\,\hat{\theta}_k.$$

Since $T^{(t)} = T_R + N\sum_k \gamma_k$, the left-hand side equals $T_R(\theta^{(t+1)} - \theta^\star) + N(\sum_k \gamma_k)\theta^{(t+1)}$. Substituting and rearranging,

$$T_R(\theta^{(t+1)} - \theta^\star) = N\sum_k \gamma_k(\hat{\theta}_k - \theta^{(t+1)}).$$

Repeating the same step starting from $T^{(t)}\theta^{(t+1)} = T^{(t)}\theta^\star + (T^{(t)}\theta^{(t+1)} - T^{(t)}\theta^\star)$ and using $T^{(t)}\theta^\star = T_R\theta^\star + N(\sum\gamma_k)\theta^\star$ yields the symmetric identity

$$T^{(t)}(\theta^{(t+1)} - \theta^\star) = N\sum_k \gamma_k(\hat{\theta}_k - \theta^\star), \tag{42}$$

which is (40). For the MSE bound, $(a+b)^2 \leq 2a^2 + 2b^2$ applied to $\theta^{(t+1)} - \theta_0 = (\theta^{(t+1)} - \theta^\star) + (\theta^\star - \theta_0)$ gives

$$\mathbb{E}\|\theta^{(t+1)} - \theta_0\|^2 \leq 2\mathrm{MSE}_\star + 2\mathbb{E}\|\theta^{(t+1)} - \theta^\star\|^2.$$

By (42) and $T^{(t)} \geq N_0$,

$$\|\theta^{(t+1)} - \theta^\star\|^2 \leq \left(\tfrac{N}{N_0}\right)^2 \Big(\sum_k |\gamma_k|\,\|\hat{\theta}_k - \theta^\star\|\Big)^2 \leq K\Big(\tfrac{N}{N_0}\Big)^2 \sum_k \gamma_k^2\,\|\hat{\theta}_k - \theta^\star\|^2$$

by Cauchy–Schwarz. Take expectations to conclude. □

E.4.5 Margin condition and weight concentration

Definition E.8 (Classification margin on the well-separated event)**.** Let $\rho_k = \rho_k(\theta_0)$ denote the per-source expected log-LR (14), and let $\mathcal{S} \triangleq \{\|\theta_k - \theta_0\| \geq r_{\mathrm{sep}} \text{ for all } k \in \bar{R}\}$ be the well-separated event under Assumption E.3. On $\mathcal{S}$, the population classification margin is $\min_k |\rho_k|$. Assumption E.4 posits the existence of $\Delta^* > 0$ that lower-bounds this margin on $\mathcal{S}$ and the high-probability relevant-cluster ball; Remark E.5 gives a sufficient Fisher-information condition for the Gaussian-mean model.

The next theorem bounds $\mathbb{E}[(\gamma_k^{(t)})^2]$ under the margin and a basin-entry condition on $\theta^{(t)}$.

Theorem E.9 (Weight concentration)**.** *Let $B \triangleq \max_k |\log\frac{\pi_k}{1-\pi_k}|$ and assume $\Delta^* > 0$. Set $c_1 = 1/4$, $c_2 = 1/2$, $c_3 = 1/32$. On the event*

$$\mathcal{E} \triangleq \Big\{\|\theta^{(t)} - \theta_0\| \leq c_1\Delta^*/L\Big\} \;\cap\; \bigcap_{k=1}^{K} \Big\{\big|\bar{s}_k(\theta^{(t)}) - \rho_k(\theta^{(t)})\big| \leq \Delta^*/4\Big\}, \tag{43}$$

the sigmoid input in (36) *has the correct sign with magnitude*

$$\big(2\mathbf{1}_R(k) - 1\big)\big[\beta s_k(\theta^{(t)}) + \log\tfrac{\pi_k}{1-\pi_k}\big] \;\geq\; c_2\,\beta N\,\Delta^* - B, \tag{44}$$

and consequently $|\gamma_k^{(t)}| \le \exp\big(-c_2\beta N\Delta^* + B\big)$ *provided* $c_2\beta N\Delta^* > B$. *Combined with sub-Gaussian concentration of the per-sample log-LR sums under Assumption E.2,*

$$\mathbb{E}\big[(\gamma_k^{(t)})^2\big] \;\le\; \exp\big(-2c_2\beta N\Delta^* + 2B\big) \;+\; 2K\,\exp\big(-c_3 N(\Delta^*)^2/V^2\big), \tag{45}$$

provided $\|\theta^{(t)} - \theta_0\| \le c_1\Delta^*/L$.

The first term in (45) is the residual sigmoid mass when the input has correct sign, and the second is the failure probability of the concentration event $\mathcal{E}$. We prove Theorem E.9 by combining a Lipschitz bound on $\rho_k(\theta)$ with sub-Gaussian concentration of the empirical log-LR, following the population-to-sample EM analysis template of Balakrishnan et al. [2017], specialized to our sigmoid-weighted setting with per-source binary latent indicators.

Lemma E.10 (Lipschitz drift of ρ_k in the basin)**.** *Under the Lipschitz clause of Assumption E.2, for any* θ *with* $\|\theta - \theta_0\| \le \Delta^*/(4L)$,

$$\big|\rho_k(\theta) - \rho_k\big| \le L\|\theta - \theta_0\| \le \Delta^*/4.$$

In particular, $\rho_k(\theta)\cdot\operatorname{sign}(\rho_k) \ge |\rho_k| - \Delta^*/4 \ge 3\Delta^*/4$ *for every* k *(using* $|\rho_k| \ge \Delta^*$*).*

Proof. Direct application of the Lipschitz clause in Assumption E.2 and the definition of Δ^*. □

Lemma E.11 (Sub-Gaussian concentration of the empirical log-LR)**.** *Under the sub-Gaussian clause of Assumption E.2, for any fixed* θ *in the basin and* $t > 0$,

$$\mathbb{P}_{\mathcal{D}_k\sim p_k^{\otimes N}}\big(\,|\bar{s}_k(\theta) - \rho_k(\theta)| > t\,\big) \le 2\exp\big(-Nt^2/(2V^2)\big).$$

Proof. $\bar{s}_k(\theta) = N^{-1}\sum_i \ell(x_i^{(k)};\theta)$ is the sample mean of i.i.d. sub-Gaussian random variables with parameter V^2; this is the standard Hoeffding-type sub-Gaussian tail bound. □

Proof of Theorem E.9. On $\mathcal{E}$, decompose

$$\bar{s}_k(\theta^{(t)}) = \rho_k + \big[\rho_k(\theta^{(t)}) - \rho_k\big] + \big[\bar{s}_k(\theta^{(t)}) - \rho_k(\theta^{(t)})\big].$$

By Lemma E.10 the first bracket is at most $\Delta^*/4$ in absolute value, and by the second clause of $\mathcal{E}$ the second bracket is at most $\Delta^*/4$. Hence

$$\big|\bar{s}_k(\theta^{(t)}) - \rho_k\big| \le \Delta^*/2,$$

and since $|\rho_k| \ge \Delta^*$, $\bar{s}_k(\theta^{(t)})$ has the same sign as ρ_k with magnitude $\ge \Delta^*/2$. Multiplying by N,

$$\big(2\mathbf{1}_R(k) - 1\big)\, s_k(\theta^{(t)}) \ge N\Delta^*/2 = c_2 N\Delta^*.$$

Multiplying by β and adding the bounded prior shift $\log[\pi_k/(1-\pi_k)]$ (whose absolute value is at most B, so $|(2\mathbf{1}_R(k) - 1)\log[\pi_k/(1-\pi_k)]| \le B$) gives (44). The sigmoid satisfies $\sigma(z) \le e^{-|z|}$ on the side of decay, so for $k \in R$ (where $\mathbf{1}_R(k) = 1$ and $w_k^* = 1$),

$$|\gamma_k^{(t)}| = 1 - w_k^{(t)} = \sigma\big(-(\beta s_k(\theta^{(t)}) + \log\tfrac{\pi_k}{1-\pi_k})\big) \le \exp\big(-c_2\beta N\Delta^* + B\big),$$

and analogously for $k \in \bar{R}$. Squaring gives the first term of (45).

For the failure probability, Lemma E.11 applied with $t = \Delta^*/4$ and a union bound over the K source concentration events gives

$$\mathbb{P}(\mathcal{E}^c) \le 2K\exp\big(-N(\Delta^*/4)^2/(2V^2)\big) = 2K\exp\big(-c_3 N(\Delta^*)^2/V^2\big)$$

with $c_3 = 1/32$ (the basin condition on $\theta^{(t)}$ is part of the standing assumption, not an additional event). On $\mathcal{E}^c$, $|\gamma_k| \le 1$, contributing at most $\mathbb{P}(\mathcal{E}^c)$ to $\mathbb{E}[\gamma_k^2]$. Adding gives the second term of (45). □

E.4.6 Proof of Theorem 4.1 and Corollary E.13

Proof of Theorem 4.1. Throughout this proof, "with probability at least $1 - K\alpha$ over the prior draw" refers to the well-separated event $\mathcal{S} \triangleq \{\|\theta_k - \theta_0\| \geq r_{\text{sep}} \text{ for all } k \in \bar{R}\}$, which by Assumption E.3 and a union bound satisfies $\mathbb{P}(\mathcal{S}) \geq 1 - K\alpha$. The expectation on the left-hand side of (17) is taken over the source data given the prior realization $\{\theta_k\}_{k=1}^K$. We work on the event $\mathcal{S}$ throughout.

On $\mathcal{S}$, Assumption E.4 gives $\min_k |\rho_k| \geq \Delta^* > 0$ (Definition E.8), so the basin and margin hypotheses of Theorem E.9 hold. Lemma E.7 yields, in conditional expectation given $\{\theta_k\} \in \mathcal{S}$,

$$\mathbb{E}\|\theta^{(t+1)} - \theta_0\|^2 \;\leq\; 2\,\mathrm{MSE}_\star + 2K\Big(\tfrac{N}{N_0}\Big)^2 \sum_{k=1}^{K} \mathbb{E}\big[\gamma_k^2\,\|\hat{\theta}_k - \theta^\star\|^2\big], \tag{46}$$

where $\mathrm{MSE}_\star$ is the oracle MSE (39) and the conditioning on $\{\theta_k\} \in \mathcal{S}$ is implicit on both sides; the conditional/unconditional MSE differ by at most a factor $1/\mathbb{P}(\mathcal{S}) \leq 1/(1 - K\alpha) \leq 2$ for $K\alpha \leq 1/2$, absorbed into the constant 2. The crucial step is to bound the cross-moment $\mathbb{E}[\gamma_k^2\,\|\hat{\theta}_k - \theta^\star\|^2]$ without splitting it as a product of marginals — γ_k and $\hat{\theta}_k$ both depend on $\mathcal{D}_k$, so they are not independent.

Let $\mathcal{E}$ be the concentration event of Theorem E.9. On $\mathcal{E} \cap \{\text{basin}\}$, that theorem yields the deterministic bound $|\gamma_k| \leq G_{\mathcal{E}} \triangleq e^{-c_2\beta N\Delta^* + B}$. On $\mathcal{E}^c$, $|\gamma_k| \leq 1$ trivially. Under Assumption E.1 ($\tau = 0$), the in-expectation moment bound

$$M_k^2 \;\triangleq\; \mathbb{E}\|\hat{\theta}_k - \theta^\star\|^2 \;\leq\; 2\,\mathbb{E}\|\hat{\theta}_k - \theta_0\|^2 + 2\,\mathrm{MSE}_\star \;\leq\; 2\big(\bar{D}^2 + d\sigma^2/N\big) + 2\,\mathrm{MSE}_\star$$

holds (using $\mathbb{E}\|\hat{\theta}_k - \theta_0\|^2 \leq \bar{D}^2 + d\sigma^2/N$ for $k \in \bar{R}$ and $\leq d\sigma^2/N$ for $k \in R$). Splitting the cross moment on $\mathcal{E}$ vs. $\mathcal{E}^c$:

$$\begin{aligned}\mathbb{E}\big[\gamma_k^2\,\|\hat{\theta}_k - \theta^\star\|^2\big] &= \mathbb{E}\big[\gamma_k^2\,\|\hat{\theta}_k - \theta^\star\|^2 \mathbf{1}_{\mathcal{E}}\big] + \mathbb{E}\big[\gamma_k^2\,\|\hat{\theta}_k - \theta^\star\|^2 \mathbf{1}_{\mathcal{E}^c}\big] \\ &\leq G_{\mathcal{E}}^2 \cdot M_k^2 + \sqrt{\mathbb{E}\|\hat{\theta}_k - \theta^\star\|^4 \cdot \mathbb{P}(\mathcal{E}^c)},\end{aligned}$$

where the first inequality uses $|\gamma_k| \leq G_{\mathcal{E}}$ deterministically on $\mathcal{E}$ and the second uses Cauchy–Schwarz on $\mathcal{E}^c$. The fourth moment $\mathbb{E}\|\hat{\theta}_k - \theta^\star\|^4$ is bounded by $O((\bar{D}^2 + d\sigma^2/N)^2)$ using Gaussian moments and Assumption E.1. By Theorem E.9, $\mathbb{P}(\mathcal{E}^c) \leq 2K\,e^{-c_3 N(\Delta^*)^2/V^2}$, contributing a $\sqrt{K}$ factor through $\sqrt{\mathbb{P}(\mathcal{E}^c)}$. Summing over the K sources picks up another factor of K, and the leading K from the per-step decomposition (46) gives a total K^2 on the first exponential and $K^{5/2}$ on the second. Absorbing $\bar{D}^2$, $(N/N_0)^2$, and the e^{2B} factor into the model-dependent constants C_1, C_2:

$$2K\Big(\tfrac{N}{N_0}\Big)^2 \sum_{k} \mathbb{E}\big[\gamma_k^2\,\|\hat{\theta}_k - \theta^\star\|^2\big] \;\leq\; C_1\,e^{-c_2\beta N\Delta^*} + C_2\,e^{-c_3 N(\Delta^*)^2/(2V^2)},$$

with $C_1 = O(K^2(N/N_0)^2(\bar{D}^2 + d\sigma^2/N)e^{2B})$ and $C_2 = O(K^{5/2}(N/N_0)^2(\bar{D}^2 + d\sigma^2/N))$. This gives the second and third bracketed terms in (17) (after redefining $c_3 \leftarrow c_3/2$). Substituting back into (46) yields (17) on $\mathcal{S}$. Since $\mathbb{P}(\mathcal{S}) \geq 1 - K\alpha$, the bound holds with probability at least $1 - K\alpha$ over the prior. □

Remark E.12. Basin-invariance The basin-invariance assumption (the iterate stays in $\{\|\theta - \theta_0\| \leq c_1\Delta^*/L\}$ across all iterations) is critical to our analysis. The in-expectation MSE bound at any fixed t does not, by itself, propagate to $t + 1$ without an additional concentration argument that would close the loop. We treat basin invariance as a standing hypothesis, with the LIP-correctness condition characterizing when basin entry occurs at $t = 0$.

Corollary E.13 (Unconditional MSE)**.** *Under the same assumptions as Theorem 4.1, integrating* (17) *over the prior gives the unconditional bound*

$$\mathbb{E}\big\|\theta^{(\infty)} - \theta_0\big\|^2 \;\leq\; \frac{2d\sigma^2}{N_0 + N|R|} \;+\; 3K\alpha\,r_{\text{sep}}^2 \;+\; C_1\,e^{-c_2\beta N\Delta^*} \;+\; C_2\,e^{-c_3 N(\Delta^*)^2/V^2}. \tag{47}$$

Proof of Corollary E.13 (integrating over the prior). Let $\mathcal{V} \triangleq \{k \in \bar{R} : \|\theta_k - \theta_0\| < r_{\text{sep}}\}$ be the violating subset, so $\mathcal{S}^c = \{\mathcal{V} \neq \emptyset\}$. By Assumption E.3 and a union bound, $\mathbb{P}(\mathcal{S}^c) \leq K\alpha$. The law of total expectation gives

$$\mathbb{E}\|\theta^{(\infty)} - \theta_0\|^2 \;=\; \mathbb{E}\Big[\|\theta^{(\infty)} - \theta_0\|^2 \,\Big|\, \mathcal{S}\Big]\;\mathbb{P}(\mathcal{S}) + \mathbb{E}\Big[\|\theta^{(\infty)} - \theta_0\|^2 \,\Big|\, \mathcal{S}^c\Big]\;\mathbb{P}(\mathcal{S}^c). \tag{48}$$

The first term is bounded by Theorem 4.1: that theorem's high-probability statement (with probability $\geq 1 - K\alpha$ over the prior, $\mathbb{E}\|\theta^{(\infty)} - \theta_0\|^2 \leq B$) is equivalent, by averaging over $\{\theta_k\} \in \mathcal{S}$, to the conditional MSE bound $\mathbb{E}[\|\theta^{(\infty)} - \theta_0\|^2 \mid \mathcal{S}] \leq B$, where $B = 2d\sigma^2/(N_0 + N|R|) + C_1 e^{-c_2\beta N\Delta^*} + C_2 e^{-c_3 N(\Delta^*)^2/V^2}$. Multiplied by $\mathbb{P}(\mathcal{S}) \leq 1$, this yields the same upper bound for the first term of (48).

On $\mathcal{S}^c$, the M-step is the convex combination

$$\theta^{(\infty)} - \theta_0 = \frac{N_0(\hat{\theta}_0 - \theta_0) + \sum_{k=1}^K w_k^{(\infty)} N(\hat{\theta}_k - \theta_0)}{N_0 + \sum_k w_k^{(\infty)} N},$$

which lies in the convex hull of $\{\hat{\theta}_0 - \theta_0\} \cup \{\hat{\theta}_k - \theta_0\}_{k=1}^K$ with weights $\propto N_0, N w_k^{(\infty)}$. Partitioning the source contributions by membership in $\mathcal{V}$, R, and $\bar{R} \setminus \mathcal{V}$:

- For $k \in R$: $\theta_k = \theta_0$ exactly under Assumption E.1, so $\|\hat{\theta}_k - \theta_0\| = \|\hat{\theta}_k - \theta_k\| = O_p(\sigma\sqrt{d/N})$.
- For $k \in \mathcal{V}$: $\|\theta_k - \theta_0\| < r_{\text{sep}}$ by definition of violation.
- For $k \in \bar{R} \setminus \mathcal{V}$: $\|\theta_k - \theta_0\| \geq r_{\text{sep}}$, so the per-source margin still holds and Theorem E.9 applied per source gives $\mathbb{E}[w_k^{(\infty)}] \leq C\, e^{-c_2\beta N\Delta^*}$, exponentially small in N. Their contribution to the convex combination is at most $\sum_{k\in\bar{R}\setminus\mathcal{V}} \mathbb{E}[w_k^{(\infty)}\, \|\hat{\theta}_k - \theta_0\|]$, whose Cauchy–Schwarz bound is $KC\, e^{-c_2\beta N\Delta^*}\sqrt{\bar{D}^2 + d\sigma^2/N}$.

Combining via the triangle inequality, taking the maximum across the three contributions, and using $\|\hat{\theta}_0 - \theta_0\|, \|\hat{\theta}_k - \theta_k\| = O_p(\sigma\sqrt{d/\min(N_0, N)})$ on Gaussian noise,

$$\|\theta^{(\infty)} - \theta_0\| \;\leq\; r_{\text{sep}} + O_p\big(\sigma\sqrt{d/N_0}\big) + KC\sqrt{\bar{D}^2 + d\sigma^2/N}\, e^{-c_2\beta N\Delta^*}.$$

Squaring (using $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$) and conditioning on $\mathcal{S}^c$:

$$\mathbb{E}\left[\|\theta^{(\infty)} - \theta_0\|^2 \,\middle|\, \mathcal{S}^c\right] \mathbb{P}(\mathcal{S}^c) \;\leq\; 3K\alpha\, r_{\text{sep}}^2 \;+\; (\text{lower-order}), \tag{49}$$

where the lower-order $O(\sigma^2 dK\alpha/N_0)$ term is absorbed into the leading $3K\alpha r_{\text{sep}}^2$ penalty (since $r_{\text{sep}}^2 \gg \sigma^2 d/N_0$ under Assumption E.3 on the working scale of r_{sep}), and the $K^2C^2(\bar{D}^2 + d\sigma^2/N)\, e^{-2c_2\beta N\Delta^*}$ term is absorbed into the weight-error residual $C_1 e^{-c_2\beta N\Delta^*}$.

Substituting the conditional bound (17) (Theorem 4.1) and the violation bound (49) into (48) yields (47). □

E.5 Proof of Theorem 4.2

We prove convergence of both update rules by showing each is a continuous function of inputs that converge to limits that recover θ_0.

Exact M-step (12). The regularity conditions give $H_0/N_0 \xrightarrow{p} \mathcal{I}(\theta_0)$ with $\mathcal{I}(\theta_0) \succ 0$. Matrix inversion is continuous on the open set of positive-definite symmetric matrices, so by the continuous mapping theorem $(H_0/N_0)^{-1} \xrightarrow{p} \mathcal{I}(\theta_0)^{-1}$. The deterministic identity $H_0^{-1} = N_0^{-1}(H_0/N_0)^{-1}$ combined with Slutsky's theorem (a deterministic null sequence times a sequence converging in probability to a finite limit converges in probability to zero) gives $H_0^{-1} \xrightarrow{p} \mathbf{0}$.

By assumption, $\mathcal{D}_k$ has fixed size N_k independent of N_0, so $\hat{\theta}_k$ and H_k are random objects whose distributions do not depend on N_0. The relevance weight satisfies $w_k^{(t)} \in [0,1]$ by construction, so $C_k \triangleq (I + \tau^2 H_k)^{-1} H_k$ has $\|C_k\|_{\text{op}}$ finite a.s. and independent of N_0. By submultiplicativity of the operator norm,

$$\big\|\Lambda_k^{(t)}\big\|_{\text{op}} = \big\|w_k^{(t)} H_0^{-1} C_k\big\|_{\text{op}} \leq \|H_0^{-1}\|_{\text{op}} \cdot \|C_k\|_{\text{op}}.$$

By Slutsky's theorem the right-hand side converges to 0 in probability, so $\Lambda_k^{(t)} \xrightarrow{p} \mathbf{0}$.

The source MLEs $\hat{\theta}_k$ for $k \geq 1$ have fixed sample size N_k and so are bounded in probability with distributions independent of N_0. Combined with $\Lambda_k^{(t)} \xrightarrow{p} \mathbf{0}$ from Step 2, Slutsky's theorem yields $\Lambda_k^{(t)}\hat{\theta}_k \xrightarrow{p} \mathbf{0}$ and $I + \sum_k \Lambda_k^{(t)} \xrightarrow{p} I$. The mapping $A \mapsto A^{-1}$ is continuous at $A = I$, so by the continuous mapping theorem $(I + \sum_k \Lambda_k^{(t)})^{-1} \xrightarrow{p} I$. Substituting these limits and $\hat{\theta}_0 \xrightarrow{p} \theta_0$ into the M-step (12):

$$\theta^{(t+1)} = \Big(I + \sum_{k=1}^{K} \Lambda_k^{(t)}\Big)^{-1} \Big(\hat{\theta}_0 + \sum_{k=1}^{K} \Lambda_k^{(t)}\hat{\theta}_k\Big) \xrightarrow{p} I^{-1}\big(\theta_0 + 0\big) = \theta_0. \qquad \square$$

Surrogate update (13). Define $r_k^{(t)} \triangleq w_k^{(t)} N_k/N_0$ for each k. Dividing the numerator and denominator of (13) by N_0 gives

$$\theta^{(t+1)} = \frac{\hat{\theta}_0 + \sum_{k=1}^{K} r_k^{(t)}\,\hat{\theta}_k}{1 + \sum_{k=1}^{K} r_k^{(t)}}. \tag{50}$$

Since $w_k^{(t)} \in [0,1]$ a.s. and N_k is fixed,

$$0 \leq r_k^{(t)} \leq N_k/N_0 \xrightarrow{N_0\to\infty} 0,$$

so the squeeze theorem gives $r_k^{(t)} \xrightarrow{a.s.} 0$, hence in probability. Combined with $\hat{\theta}_0 \xrightarrow{p} \theta_0$ and the constancy of $\hat{\theta}_k$ in N_0, the continuous mapping theorem applied to (50)'s numerator and denominator gives, with the denominator limit 1 bounded away from zero,

$$\theta^{(t+1)} \xrightarrow{p} \frac{\theta_0 + \sum_k 0 \cdot \hat{\theta}_k}{1 + \sum_k 0} = \theta_0. \qquad \square$$

E.6 Proof of Theorem 4.3

Proof. The relevance weight (20) reads

$$w_k^{(t)} = \sigma\left(\log \frac{\mathcal{L}(\theta^{(t)}; \mathcal{D}_k)}{p(\mathcal{D}_k \mid c_k = 0)} + \log \frac{\pi_k}{1-\pi_k}\right),$$

where $\mathcal{L}(\theta; \mathcal{D}_k) = \prod_i p(x_i \mid \theta)$ is the per-sample relevant likelihood (by Remark 3.2 at $\tau = 0$). The null factor $p(\mathcal{D}_k \mid c_k = 0) = \prod_i q(x_i)$ factorizes because q is a fixed density. Hence the log-ratio factorizes as $\sum_{i=1}^{N_k} \log[p(x_i \mid \theta^{(t)})/q(x_i)]$. Define

$$A_{N_k} \triangleq \sum_{i=1}^{N_k} \log \frac{p(x_i \mid \theta^{(t)})}{q(x_i)} + \log \frac{\pi_k}{1-\pi_k}, \tag{51}$$

so $w_k^{(t)} = \sigma(A_{N_k})$.

The triangle inequality and the hypothesis give

$$\mathbb{E}_{p_k}\big|\log[p(\cdot \mid \theta^{(t)})/q]\big| \leq \mathbb{E}_{p_k}|\log p(\cdot \mid \theta^{(t)})| + \mathbb{E}_{p_k}|\log q| < \infty.$$

Linearity of expectation gives $\mathbb{E}_{p_k}\big[\log[p(\cdot \mid \theta^{(t)})/q]\big] = \rho_k(\theta^{(t)})$, the quantity in (14).

Conditional on $\theta^{(t)}$ (a function of past iterates and data), the samples $\{x_i^{(k)}\}_{i=1}^{N_k}$ are i.i.d. under p_k and the per-sample log-ratio is integrable; the unconditional almost-sure convergence below follows by the law of total expectation. The strong law of large numbers gives

$$\sum_{i=1}^{N_k} \log \frac{p(x_i \mid \theta^{(t)})}{q(x_i)} = N_k\big(\rho_k(\theta^{(t)}) + \delta_{N_k}\big) \quad \text{a.s.}, \tag{52}$$

where $\delta_{N_k} \xrightarrow{a.s.} 0$.

Substituting (52) into (51),

$$A_{N_k} = N_k\big(\rho_k(\theta^{(t)}) + \delta_{N_k}\big) + \log \frac{\pi_k}{1-\pi_k} \quad \text{a.s.}$$

The last term is a finite constant under $\pi_k \in (0,1)$. Eventually $|\delta_{N_k}| < |\rho_k(\theta^{(t)})|/2$ a.s., so the leading term dominates: $A_{N_k} \to +\infty$ a.s. when $\rho_k(\theta^{(t)}) > 0$, and $A_{N_k} \to -\infty$ a.s. when $\rho_k(\theta^{(t)}) < 0$.

The sigmoid extends continuously to the extended real line via $\sigma(\pm\infty) \in \{0,1\}$, so

$$w_k^{(t)} = \sigma(A_{N_k}) \xrightarrow{a.s.} \mathbf{1}\{\rho_k(\theta^{(t)}) > 0\}.$$

Almost-sure convergence implies convergence in probability. □

F Experimental Results

F.1 Benchmark Methods

In all the experiments, we compare the following estimators:

- **Target-Only:** An estimator trained exclusively on the limited target dataset $\mathcal{D}_0$;
- **Naive Pooling:** An estimator trained indiscriminately on the union of the target dataset and all available source datasets $\mathcal{D}_k$;
- **LIP-G (Gemini 3 Flash):** Our proposed method (LIP estimation and EM) leverages a Gemini 3 Flash API to generate a prior over source relevance using contextual descriptions;
- **LIP-C (Claude Opus 4.7):** The same LIP estimation and EM as LIP-G but with the LIP elicited from the Anthropic Claude Code local agent mode instead of calling Google's Gemini API;
- **EM with a Non-informative Prior:** An EM estimator that assumes a uniform initial prior across all source domains, acting as a purely data-driven approach without textual information.

F.2 Detailed Experimental Setup

All three experiments share the same LIP construction protocol and EM convergence criterion. We use the Claude Code `Opus 4.7` local agent and Google `gemini-3-flash-preview` model (temperature 0, JSON-constrained outputs) as the elicitation oracle. The conditional-logit-with-null likelihood (2) is fit by L-BFGS with strong-Wolfe line search. EM iterates until $\|w^{(t)} - w^{(t-1)}\|_\infty \le 10^{-3}$ for five consecutive iterations or a hard cap is reached. All experiments use random seed 42 unless otherwise stated.

F.2.1 C-MAPSS

Data. We use the FD001 subset [Saxena et al., 2008], treating each of the 100 engines as one domain. Ten target engines are sampled (machines $\{4, 9, 18, 26, 27, 48, 51, 53, 55, 80\}$) and the remaining 99 engines serve as the source pool for each target. We predict the high-pressure-compressor physical core speed (sensor 9) from the cycle index. The backbone is a generalized linear model with a natural cubic regression spline basis: 5 knots placed uniformly on $[0, 300]$ in cycle space and the truncated-power form constrained to remain linear outside the boundary knots. We sweep nine cold-start cutoffs corresponding to RUL levels $\{90\%, 80\%, \ldots, 10\%\}$.

LIP construction. A NASA technical report on engine damage propagation [Saxena et al., 2008] is uploaded once as the LLM context, and the target description is the dust-ingestion paragraph reproduced verbatim in our open-source release. We issue 200 subgroup queries with $|S_m| \in \{3, 4, 5\}$ sampled uniformly, executed with 10 parallel workers and a 429-aware retry policy. The fitted LIP uses $p_0 = 0.01$ and $\varepsilon = 0.1$; the same response set is reused across all ten target engines by dropping queries that contain the chosen target and re-indexing. We report two LIPs constructed identically but elicited from `gemini-3-flash-preview` and Anthropic Claude.

EM. EM uses the closed-form M-step on the GLM coefficients with $\tau = 10^{-3}$, $\nu = 0.05$, the exact τ^2-corrected E-step, and the empirical-Bayes null with self-exclusion. The outer loop is capped at 1000 iterations; the convergence criterion follows the shared protocol. Pooled-OLS and Target-Only baselines use ridge regression ($\lambda = 10^4$, intercept un-penalized) so that the truncated-power basis remains numerically stable at intermediate cutoffs.

F.2.2 MuJoCo Hopper

Data. The source pool is $K = 10$ replay buffers collected by training Soft Actor-Critic [Haarnoja et al., 2018] in Hopper-v5 with gravity $g \in \{1, 2, \ldots, 10\}$ $\mathrm{m/s^2}$ (one buffer per gravity, $\sim 10^6$ transitions each). The target environment uses Venus-like gravity $g = 8.87$ $\mathrm{m/s^2}$, and the target dataset is a separate SAC replay buffer of $\sim 10^6$ transitions from which we draw $N_0 \in \{128, 256, 512, 1024, 2048, 4096\}$ samples for each cell of the table.

Pool dynamics (θ_{pool}). The dynamics model is a multivariate Gaussian distribution parametrized with hidden width 512, 4 residual blocks (Linear + LayerNorm + SiLU), input/output normalization buffers, and learned softplus clamps on the log-variance head. We pretrain θ_{pool} on the union of all 10 source replay buffers for 1000 epochs (batch 1024, AdamW with weight decay 10^{-4}, learning rate 3×10^{-4} with cosine decay to 10^{-5}, gradient clipping at 1.0), holding out 20,000 transitions per source for diagnostics. The terminal learning rate 10^{-5} matches the EM M-step's constant rate, so the M-step continues training without a fresh LR peak.

LIP construction. The technical report passed to the LLM describes the hopper environment and gravity-driven dynamics; the target description is "the hopper is deployed to a planet with Venus-like gravity." Source datasets are anonymized as `source_NN.csv` (state-summary excerpts) under a random source-to-gravity mapping, so the LLM does not see the gravity value. We issue 50 subgroup queries with concurrency 3 and a 120-second fire interval. The fitted LIP uses $p_0 = 0.01$ and $\varepsilon = 1.0$. We additionally fit a "False LIP" from the responses of Gemini 3 Flash. It uses identical hyperparameters but argmax at $g = 7$ $\mathrm{m/s^2}$ instead of $g = 9$ $\mathrm{m/s^2}$; this corresponds to LIP-G in the body and serves as the misspecified-prior ablation.

EM. For neural-network dynamics, the closed-form weighted-average M-step is meaningless across re-parameterizations, so we use the gradient-based generalized-EM (Algorithm 3 from Appendix D). At iterate $\theta^{(t)}$, the M-step takes 100 minibatch gradient steps on the weighted negative log-likelihood

$$\mathcal{L}(\theta) = -\log p(\mathcal{D}_0 \mid \theta) - \sum_k w_k^{(t)} \log p(\mathcal{D}_k \mid \theta),$$

with batch size 1024, AdamW (weight decay 10^{-4}), and a constant learning rate 10^{-5} that matches the pool's terminal LR. The outer loop runs at most 100 iterations with $\|\Delta w\|_\infty \le 10^{-3}$ patience-5 convergence. The tempering schedule uses $\nu = 0.1$ and $\beta_k^{(t)} = (1 - e^{-\nu t})/\sqrt{d_{\mathrm{eff}} N_k / N_0}$, where d_{eff} is the trainable parameter count of θ_{pool}. The null model is a single global density $\log p(\mathcal{D}_k \mid \theta_{\mathrm{pool}})$, computed once before the EM loop and held fixed; this replaces the empirical-Bayes leave-one-out null, which is unstable in the over-parameterized regime.

Weighted IQL. After EM, we train Implicit Q-Learning [Kostrikov et al., 2022] on the mixture sampling distribution $p(\text{target}) \propto N_0$, $p(\text{source } k) \propto w_k^{(\infty)} N_k$. We use a tanh-squashed Gaussian policy and twin Q-networks each with hidden width 256 and LayerNorm, 200,000 training steps, batch 256, Adam at 3×10^{-4} with cosine decay to zero, target soft-update $\tau = 0.005$, discount $\gamma = 0.99$, expectile 0.7, AWR temperature $\beta = 3.0$, and AWR weight clamp 100. The reported policy returns are the mean over 200 evaluation episodes at $g = 8.87$ $\mathrm{m/s^2}$.

F.3 Hardware and Training Time.

All experiments run on a single workstation with an NVIDIA RTX 5090 GPU, an AMD Ryzen 9 9950X3D CPU, and 64 GB of RAM. The Gaussian and C-MAPSS pipelines complete in seconds on CPU. Pool dynamics training takes $\sim$14 hours, *each EM completes in* $1 \sim 3$ *minutes*, and the IQL over 6×5 cells completes in roughly 5 hours.

F.4 Assessment of LLM-Generated Priors via Reasoning Analysis

As we observed in the experiments, the quality of the LIP generated by the LLM has a significant impact on the EM algorithm's convergence and final performance. In practice, we suggest the best way to audit the quality of the LIP is to directly inspect the LLM's reasoning and check its logic. We provide a detailed analysis of the LLM's reasoning in the hopper experiment, where we observed

three significant failure modes in the Gemini 3 Flash model. We also analyzed the reasoning patterns of Claude Code Opus 4.7 to understand why it produced a more accurate LIP.

F.4.1 Using Local Agent for Relevance Assessment

For Claude Code Opus 4.7, we used **Local Agent Assessment**, where a local agent, designed to mimic human reasoning, evaluates the relevance of each source based on the contextual information and provides a more structured analysis. One huge advantage of this approach is that the dataset does not need to be directly uploaded to the LLM, which can be a significant bottleneck in terms of both time and cost. In the hopper experiment, the 10 sources adds up to 1 GB data, which we could only sample a few trajectories to upload to Gemini. Instead of uploading the entire dataset, the local agent can analyze the dataset and extract relevant features or summaries that are then fed into the LLM for relevance assessment.

We now summarize the key steps of the Local Agent approach. The agent (Claude Code Opus 4.7) is given filesystem access and a Python execution tool, but no internet access and no exposure to the source-to-gravity mapping.

Step 1: Read context. The agent reads three input files: a task specification, a technical report describing the hopper telemetry (z, $\dot{z}$, joint angles/velocities, actions, terminal flag at $\Delta t = 0.008\,\mathrm{s}$), and a target description identifying the goal as Venus-like gravity ($\sim 8.87\ \mathrm{m/s^2}$).

Step 2: Inspect the source data. The agent lists the data directory and inspects the column format of the 10 anonymized CSV files (`source_01.csv`, ..., `source_10.csv`).

Step 3: Estimate gravity per source. The agent iteratively writes and runs six gravity-estimation scripts and compare their performance. The methods are:

1. average $\ddot{z}$ at hop apex (where $\dot{z} \approx 0$ and z is high);
2. histogram mode of $\ddot{z}$ over high-z samples;
3. short-window quadratic fit to $z(t)$ in a ± 5-step window around each apex;
4. joint-motion-minimal filter, restricting to samples whose joint velocities are small *and* z is high;
5. approximate centre-of-mass acceleration from link kinematics;
6. long-flight ballistic fit over runs of ≥ 20 consecutive high-z timesteps with RMSE $< 5\,\mathrm{cm}$.

Methods 1–3 disagreed on `source_04` (e.g., apex 5.02 vs. mode $8.19\ \mathrm{m/s^2}$). Method 4 was inconclusive due to perpetual joint motion in `source_10`. Method 5 was also inconclusive due to an inaccurate kinematic-chain approximation. Method 6 yielded tight per-source standard deviations (0.16–0.84) and resolved the residual ambiguities, resulting in `source_10` at $9.10\ \mathrm{m/s^2}$ (Venus-closest) and `source_08` at $9.50\ \mathrm{m/s^2}$ (Earth-closer).

F.4.2 Failure Modes of Gemini 3 Flash on hopper

This subsection catalogs three recurring reasoning errors observed in Gemini 3 Flash subgroup queries that produced the "False LIP" (LIP-G) used in the hopper ablation. Although our framework is provably robust to such errors as data accumulates, a false LIP can cause the slower cold-start recovery in Table 2. We document them here as the errors are not random noise but specific reasoning patterns that recur across queries.

Failure mode 1: Anchoring bias on Earth gravity This is the dominant systematic error. Gemini 3 Flash picks one source from the subgroup, assumes it represents Earth ($g = 9.81\ \mathrm{m/s^2}$), and then identifies the source approximately $1\ \mathrm{m/s^2}$ below it as Venus. Such selection rule clearly stands on no physical or statistical principle.

When the anchor is wrong, the entire chain of inference is wrong. In Q11 the subgroup contained `source_10` (true $g = 9$), `source_04` (true $g = 8$), and `source_08` (true $g = 10$):

> *`source_04` has the lowest gravity, with `source_10` being $1.0\ \mathrm{m/s^2}$ higher and `source_08` being $2.0\ \mathrm{m/s^2}$ higher. Given that Venus gravity ($8.87\ \mathrm{m/s^2}$) is approximately $1.0\ \mathrm{m/s^2}$ less*

> *than Earth gravity (*$9.81\ \mathrm{m/s^2}$*),* `source_04` *is the most dynamically consistent candidate for a Venus-like environment, assuming* `source_10` *represents the Earth baseline.*

Even worse, in Q48, the subgroup was $[g{=}1, g{=}4, g{=}5, g{=}10]$, in which `source_08` ($g = 10$) is the only proximate match for Venus. Gemini 3 Flash nonetheless picked `source_05` ($g = 4$):

> *Source 05 consistently shows an acceleration* $1.0\ \mathrm{m/s^2}$ *greater (less negative) than Source 07. Assuming Source 07 represents Earth gravity (*$9.81\ \mathrm{m/s^2}$*), Source 05's gravity is approximately* $8.81\ \mathrm{m/s^2}$.

Failure mode 2: Single-row noisy selection A correct estimate of gravity from the replay buffer requires aggregating across many free-fall transitions. Gemini 3 Flash sometimes commits to a single row whose acceleration matches the target, ignoring the rest of the trajectory. In Q27, the subgroup was $[g{=}8, g{=}5, g{=}2]$ and Gemini 3 Flash chose $g = 2$:

> *In* `source_09`*, analyzing Episode* 19 *from step* 7 *to* 8*, the vertical velocity (*$\dot{z}$*) changes from* $-0.1471\ \mathrm{m/s}$ *to* $-0.2181\ \mathrm{m/s}$ *over a* 0.008 *s interval, yielding an acceleration of* $-8.875\ \mathrm{m/s^2}$*, which is dynamically consistent with the target environment.*

Failure mode 3: Arithmetic errors mid-reasoning. Even when the measured statistics are correct, Gemini 3 Flash sometimes makes arithmetic errors in its own chain of thought and does not notice. Example from Q06:

> `source_01` *has an* a_z *of* $-7.425\ \mathrm{m/s^2}$*,* `source_02` *has* $-8.425\ \mathrm{m/s^2}$*, and* `source_10` *has* $-10.425\ \mathrm{m/s^2}$*. These differences of exactly* 1.0 *and* $3.0\ \mathrm{m/s^2}$ *indicate that* `source_02` *uses Earth gravity (*$9.81\ \mathrm{m/s^2}$*)...*