Entropic Projection Alignment: Estimating, Explaining, and Improving Model Performance Under Distribution Shift

Salim I. Amoukou **Emanuele Albini** **Tom Bewley** **Saumitra Mishra** **Manuela Veloso**
J.P. Morgan AI Research

Abstract

We propose a unified framework for addressing three key challenges of distribution shift: ① estimating a model's performance on an unlabeled target domain, ② explaining the shift by identifying the features responsible, and ③ improving the target domain performance. Our method, Entropic Projection Alignment (**EPA**), aligns the source distribution to the target by matching carefully selected moments while simultaneously minimising the KL divergence from the source. This formulation yields a unique closed-form solution for importance weights, achieving robustness through implicit variance control. Drawing on domain adaptation theory, we establish that moment matching is sufficient for reliable estimation and adaptation, avoiding the need for full density ratio recovery. Extensive experiments, together with strong theoretical guarantees, demonstrate that **EPA** consistently outperforms state-of-the-art baselines while offering substantial computational efficiency.

1 INTRODUCTION

Machine learning models are often trained on data from a *source* domain and deployed in a *target* domain where the data distribution may differ. This discrepancy between training and deployment distributions can degrade model performance. This work considers a scenario where both features and labels are available in the source domain, but only unlabeled features are observed in the target domain. This mirrors applications such as medical diagnosis or credit scoring, where predictions involve future outcomes and immediate access to labels during deployment is not feasible.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

When the target distribution can differ arbitrarily from the source distribution, estimating target performance or adapting the model becomes an ill-posed problem. To address this, the literature often imposes invariance assumptions about the nature of the distribution shift (Tasche, 2023a). Common assumptions include changes in the marginal feature distribution $P_{\mathbf{X}}$ while the conditional label distribution $P_{Y|\mathbf{X}}$ remains unchanged, known as *covariate shift* (Shimodaira, 2000); shifts in the marginal label distribution P_Y while $P_{\mathbf{X}|Y}$ remains constant, referred to as *label shift* (Saerens et al., 2002; Lipton et al., 2018); or changes in the joint distribution $P_{\mathbf{X},Y}$ under the condition that there exists a subset of features S such that $P_{\mathbf{X}_{\bar{S}}|\mathbf{X}_S,Y}$ is invariant, known as *sparse joint shift* (Chen et al., 2022; Tasche, 2023a,b).

Instead of committing to these specific shift assumptions, we adopt the perspective of domain adaptation theory (Ben-David et al., 2006; Mansour et al., 2009), which assumes only that there exists a hypothesis that performs reasonably well across both domains. In addition, we focus on machine learning models that process semantically meaningful features, such as those in tabular data, rather than unstructured data like images. Within this context, we introduce a unified framework that: ① estimates a model's performance on the unlabeled target domain; ② explains performance changes by identifying the features responsible under sparse joint shifts; and ③ improves the model to enhance its target-domain performance.

Our method transforms the source distribution by solving a constrained optimization problem: it matches key statistics of the target domain while minimizing the KL divergence from the source. This formulation yields a unique closed-form solution for the importance weights, which guarantees efficient computation and reduced estimation variance. The central theoretical insight is that accurate performance estimation and adaptation don't require recovering the full distribution ratio, but matching carefully chosen moments suffices. Our framework is supported by strong theoretical foundations and extensive empirical valida-

tion: across several distribution shifts, EPA achieves $30-70\%$ better target error estimation, more precise feature attribution, and superior model adaptation over state-of-the-art methods, all while running an order of magnitude faster.

Related Works. The work most closely related to ours is SEES (Chen et al., 2022), which also generates importance weights, estimates model performance, and identifies features driving sparse joint shifts. However, our approach diverges fundamentally in its objective. SEES attempts to directly learn the density ratio $w_S(x,y) = \frac{dQ}{dP}(\boldsymbol{x}_S, y)$ by using a class of basis functions. In contrast, we circumvent the notoriously difficult task of density ratio estimation. Instead, we find the distribution closest to the source – thereby reducing weight variance – while satisfying selected statistics of the target domain, which ensures accuracy for downstream estimation and adaptation. This formulation is not only more robust but also yields a closed-form solution, making our method significantly faster. Furthermore, unlike SEES, our approach is not limited to classification tasks.

Our approach also shares conceptual similarities with kernel mean matching (KMM) (Gretton et al., 2008; Sugiyama et al., 2012), which aligns feature means in an RKHS. KMM, however, requires solving a quadratic program whose complexity scales with the number of samples n, making it inefficient for large datasets. Our entropic projection approach scales with the feature dimension $k \ll n$. Moreover, the KL minimisation in our objective is equivalent to entropy maximisation, which provides implicit regularisation and prevents the unstable, high-variance weights that KMM can sometimes produce.

Finally, to our knowledge, no prior work has integrated all three aspects, estimating, explaining, and improving, into a single framework; most focus on only some. Refer to H for further discussion of prior works.

2 ENTROPIC PROJECTION ALIGNMENT

Notations. Let $\mathcal{X}$ and $\mathcal{Y}$ denote the input and label spaces. We consider a predictive model $f : \mathcal{X} \to \mathcal{Y}$ from a hypothesis class $\mathcal{H}$ and bounded loss function $\mathcal{L} : \mathcal{Y}^2 \to [0,1]$. We are given a labeled dataset $\mathcal{D}^{\text{src}} = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n$, drawn i.i.d from a *source* distribution $P = P_X \times P_{Y|X}$, and an unlabeled dataset $\mathcal{D}^{\text{tar}} = \{\mathbf{X}_\mathbf{i}^\mathbf{t}\}_{i=1}^m$ drawn from a *target* distribution $Q = Q_X \times Q_{Y|X}$. We define the risk of a hypothesis $f \in \mathcal{H}$ under distribution P with respect to its true labeling function c_P as $\epsilon_P(f) := \mathbb{E}_{x\sim P_X}\left[\ell(f(x), c_P(x))\right]$ and analogously for Q. We use the hat notation $\hat{\cdot}$ to represent empirical distributions.

EPA casts alignment as an *entropic projection* (Csiszár, 1984) – a classical optimisation problem that seeks the distribution closest in KL divergence to a reference distribution while satisfying linear constraints. This framework was recently used by Bachoc et al. (2023) for sensitivity analysis, where it addresses counterfactual questions such as: how would a model's predictions change if the data distribution were projected to increase the mean or correlations of the features?

In contrast, we use entropic projection for a different goal: domain alignment. Our **EPA** method imposes constraints that align the source with the target distribution. Concretely, **EPA** seeks a distribution $P^\star$ that is as close as possible to P while matching moment conditions derived from the target domain $Q_{\boldsymbol{X}}$.

Definition 2.1 (**EPA** - empirical version)**.** Let $\Phi : \mathcal{X} \to \mathbb{R}^k$ a feature map and $\widehat{\mu}_Q := \mathbb{E}_{\hat{Q}_X}[\Phi(X)]$ be the empirical target moment. Assume $\widehat{\mu}_Q$ is in the relative interior of the convex hull of $\Phi_n = \{\Phi(X_i)\}_{i=1}^n$, and $\Phi_n^\intercal$ are linearly independent. **EPA** solves

$$\hat{P}^\star \in \arg\min_{\tilde{P}} \mathrm{KL}\Big(\tilde{P} \,\|\, \hat{P}\Big) \quad \text{s.t.} \quad \mathbb{E}_{\tilde{P}}[\Phi(X)] = \widehat{\mu}_Q.$$

If $\widehat{\mu}_Q$ lies outside $\mathrm{conv}\{\Phi(X_i)\}_{i=1}^n$, the equality-constrained problem is infeasible. In that case, we project $\widehat{\mu}_Q$ onto the source convex hull and run **EPA** with the projected moment; §W shows that this is equivalent to imposing a hard proximity constraint around the target moment.

In our setting, both *estimation* and *explanation* revolve around finding the alignment (via the matching function Φ) that most accurately estimates and explains the model's performance. Then, we leverage the distribution $P^\star$, drawn from the best estimate, to *improve* the model.

Proposition 2.2. *Under the assumptions of Definition 2.1, the optimal distribution $P^\star$ is a reweighting of the source data, $\hat{P}^\star = \sum_{i=1}^n \lambda_i \delta_{\{\mathbf{X}_i, Y_i\}}$, with unique weights given by*

$$\lambda_i = \frac{\exp\{\langle \xi,\, \Phi(X_i)\rangle\}}{\sum_{j=1}^n \exp\{\langle \xi,\, \Phi(X_j)\rangle\}}, \qquad i = 1, \ldots, n,$$

where $\xi \in \mathbb{R}^k$ is the unique minimizer of the strictly convex function:

$$\mathcal{J}(\xi) = \log\left(\frac{1}{n}\sum_{i=1}^n e^{\langle \Phi(\mathbf{X}_i),\, \xi\rangle}\right) - \left\langle \xi,\, \mathbb{E}_{\hat{Q}_\mathbf{X}}[\Phi(\mathbf{X})]\right\rangle.$$

As noted by Bachoc et al. (2023), this result follows as a corollary of the theorems in Csiszár (1984). In §J.1, we provide a simpler direct proof that avoids relying on those general results. Moreover, in §R we establish new finite-sample concentration bounds comparing the empirical **EPA** in §2.1 with its population analogue.

The **EPA** approach offers several advantages:

1. *Uniqueness and implicit regularisation.* The solution is unique and corresponds to the maximum-entropy set of weights satisfying the constraints, since $\mathrm{KL}(\tilde{P} \,\|\, \hat{P}) = \log n - H(\lambda)$. This encourages weights close to uniform, acting as a strong regulariser that penalises concentration and reduces the variance of importance-weighted estimators. A formal variance analysis is provided in §J.2.

2. *Scalability.* The optimisation involves only $k = \dim(\Phi)$ parameters, whereas KMM requires solving a quadratic program scaling with n, which is prohibitive for large datasets.

3. *Distribution-free formulation.* **EPA** requires no explicit assumptions on the underlying data distribution to approximate KL divergence, unlike prior approaches such as SEES.

3 ESTIMATING MODEL PERFORMANCE

The previous section introduced the **EPA** framework, which produces a reweighted source distribution $P^\star$ that matches the target moments of Φ under Q. The ultimate goal is to ensure that the risk on the reweighted source, $\epsilon_{P^\star}(f)$, is a reliable estimator for the target risk $\epsilon_Q(f)$. This raises the central question:

Under what conditions does matching the moments of Φ lead to a small estimation gap $|\epsilon_Q(f) - \epsilon_{P^\star}(f)|$?

To address this, we draw on the theory of domain adaptation, which provides formal tools for reasoning about performance under distribution shift.

Definition 3.1 (ℓ-Discrepancy)**.** For distributions $P_X^\star, Q_X$ on $\mathcal{X}$, a hypothesis class $\mathcal{H}$, and a bounded loss $\ell \in [0,1]$, the **ℓ-discrepancy** is $\mathrm{disc}_\ell(P_X^\star, Q_X) := \sup_{f,g\in\mathcal{H}} \left|\mathbb{E}_{Q_X}[\ell(f(X), g(X))] - \mathbb{E}_{P_X^\star}[\ell(f(X), g(X))]\right|$.

This measure captures the maximum disagreement between the two distributions over the worst-case pair of hypotheses from $\mathcal{H}$. A standard result (Mansour et al., 2009; Ben-David et al., 2006), shown below, bounds the estimation gap using this discrepancy.

Proposition 3.2. *For any hypothesis $h \in \mathcal{H}$, assuming that the loss ℓ satisfies the triangle inequality (e.g., 0-1 loss, absolute loss), the following inequality holds:*

$$\left|\epsilon_Q(h) - \epsilon_{P^\star}(h)\right| \le \mathrm{disc}_\ell(P_X^\star, Q_X) + \lambda_{P^\star,Q},$$

where $\lambda_{P^\star,Q} := \inf_{f\in\mathcal{H}}(\epsilon_{P^\star}(f) + \epsilon_Q(f))$ is the combined irreducible error of the best possible hypothesis on both distributions.

See derivations in K. This bound clarifies our objective: to make the estimation gap small, we must choose our feature map Φ such that the resulting $P^\star$ has small discrepancy $\mathrm{disc}_\ell(P_X^\star, Q_X)$. The following example shows that naively matching simple moments is not enough in general.

Example 3.3 (Mismatch between Mean and Thresholds)**.** Let $\mathcal{X} = \mathbb{R}$ and $\Phi(x) = x$, so that moment matching enforces $\mathbb{E}_{P_X^\star}[X] = \mathbb{E}_{Q_X}[X]$. Let $\mathcal{H}$ be the class of threshold classifiers, $\mathcal{H} = \{h_c(x) = \mathbf{1}\{x > c\} \mid c \in \mathbb{R}\}$. The associated discrepancy with binary 0-1 loss is $\mathrm{disc}_\ell(P_X, Q_X) = 2\sup_{I\subseteq\mathbb{R}\text{ interval}} \left|\mathbb{P}_{P_X^\star}(X \in I) - \mathbb{P}_{Q_X}(X \in I)\right|$. Consider the distributions $Q_X = \frac{1}{2}\delta_{-2} + \frac{1}{2}\delta_{+2}$ and $P_X^\star = \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_{+1}$. Both distributions have a mean of 0, satisfying the moment constraint. However, for the interval $I = (1.5, 2.5)$, we have $\mathbb{P}_{Q_X}(X \in I) = 0.5$ while $\mathbb{P}_{P_X^\star}(X \in I) = 0$. This implies $\sup_I |\mathbb{P}_{P_X^\star}(I) - \mathbb{P}_{Q_X}(I)| \ge 0.5$, and $\mathrm{disc}_\ell(P_X^\star, Q_X) = 1$, rendering the bound vacuous.

3.1 Alignment

The failure of the previous example demonstrates that the feature map Φ must be chosen in a way that is sensitive to the hypothesis class $\mathcal{H}$ and loss function ℓ. We formalise this intuition with the concept of alignment.

Definition 3.4 (Disagreement Function Space)**.** Given a hypothesis class $\mathcal{H}$ and a bounded loss $\ell : \mathcal{Y} \times \mathcal{Y} \to [0,1]$, the *disagreement function space* $\mathcal{G}_{\mathcal{H},\ell}$ is defined as $\mathcal{G}_{\mathcal{H},\ell} = \big\{g : \mathcal{X} \to [0,1] \,\big|\, \exists f, h \in \mathcal{H} \text{ s.t. } g(x) = \ell(f(x), h(x))\big\}$. For the binary 0-1 loss, this space consists of indicator functions of the form $\{\mathbf{1}\{f(x) \ne h(x)\} \mid f, h \in \mathcal{H}\}$.

Definition 3.5 (Alignment)**.** A feature map $\Phi = (\phi_1, \ldots, \phi_k) : \mathcal{X} \to \mathbb{R}^k$ is **aligned** with $(\mathcal{H}, \ell)$ if the span of the disagreement functions is contained within the affine span of the components of Φ: $\mathrm{span}(\mathcal{G}_{\mathcal{H},\ell}) \subseteq \mathrm{span}\big(\{1, \phi_1, \ldots, \phi_k\}\big)$. Equivalently, for every $g \in \mathcal{G}_{\mathcal{H},\ell}$, there exist $\beta_0 \in \mathbb{R}$ and $\beta \in \mathbb{R}^k$ such that $g(x) = \beta_0 + \beta^\top \Phi(x)$ for all $x \in \mathcal{X}$.

Intuitively, alignment means that Φ is expressive enough to exactly represent any disagreement pattern between hypotheses. When alignment holds, moment matching eliminates discrepancy entirely.

Theorem 3.6 (Alignment Implies Zero Discrepancy)**.** *Let $P_X^\star$ and Q_X such that $\mathbb{E}_{P_X^\star}[\Phi(X)] = \mathbb{E}_{Q_X}[\Phi(X)]$. If Φ is aligned with $(\mathcal{H}, \ell)$, then $\mathrm{disc}_\ell(P_X^\star, Q_X) = 0$. Consequently, for any loss covered by Proposition 3.2, $\left|\epsilon_Q(h) - \epsilon_P^\star(h)\right| \le \lambda_{P^\star,Q}$, where $\lambda_{P^\star,Q}$ is small if a model exists that performs well on both domains.*

The remaining term $\lambda_{P^\star,Q}$ captures the shared-good-hypothesis assumption. Under standard covariate or label shift, it is zero whenever the hypothesis class can represent the common labelling rule. If source and target are genuinely incompatible, then $\lambda_{P^\star,Q}$ is large, and neither EPA nor any other reweighting method can be expected to provide accurate target risk estimation or reliable adaptation.

To illustrate how this abstract condition manifests in practice, we now present several concrete examples.

Example 3.7 (Alignment for Linear Models)**.** Consider linear predictors $f_w(x) = w^\top x$ on $x \in \mathbb{R}^d$ with squared loss $\ell(\hat{y}, y) = (\hat{y} - y)^2$. A disagreement function is $g(x) = \ell(f_{w_1}(x), f_{w_2}(x)) = ((w_1 - w_2)^\top x)^2$, which is a quadratic form in x. The expectation of $g(x)$ depends on the second raw moments $\mathbb{E}[x_i x_j]$. If Φ includes all degree-2 monomials, e.g., $\Phi(x) = \big(x_1, \ldots, x_d, x_1^2, x_1x_2, \ldots, x_d^2\big)^\top$, then every $g \in \mathcal{G}_{\mathcal{H},\ell}$ is an affine function of $\Phi(x)$, alignment holds, and discrepancy vanishes under moment matching.

Remark 3.8. For linear models with a squared loss, Mansour et al. (2009) obtain the reweighted source distribution that minimises disc_ℓ via a semi-definite program. Under our assumption that $\hat{\mu}_Q$ lies in the convex hull of Φ_n, however, no such optimization is required.

Example 3.9 (Alignment for Covariate Shift)**.** Under covariate shift, suppose the distributional change is driven entirely by the prevalence of a subgroup $s(X) \in \{0, 1\}$. If the conditional risk is approximately constant within the subgroup and its complement, then choosing $\Phi(X) = s(X)$ provides effective control of the estimation gap.

See proofs and additional examples in L. Our focus is on tabular data, where tree-based models are state-of-the-art and thus the models we use in our experiments. Because these models can be represented as piecewise-constant functions, they also admit alignment, as shown below.

Theorem 3.10 (Alignment for Decision Trees)**.** *Let $\mathcal{H}$ be the class of decision trees on $\mathcal{X}$ and let ℓ be the 0-1 loss. Let $\mathcal{R}_\mathcal{H}$ be the set of all possible leaf regions generated by any tree in $\mathcal{H}$. Then the feature map $\Phi(x) = (\mathbf{1}\{x \in R\})_{R \in \mathcal{R}_\mathcal{H}}$ is aligned with $(\mathcal{H}, \ell)$.*

See proof in M. This result shows that, in principle, exact alignment can be achieved for decision trees: the indicator features $\mathbf{1}\{x \in R\}$ span the full disagreement space. In practice, however, such a construction is infeasible as $\mathcal{R}_\mathcal{H}$ can be extremely large, and most regions will be empty.

To make the approach tractable, we turn to approximations. Instead of explicitly representing all tree regions, we approximate the disagreement space using quantile binning of the feature space. Concretely, motivated by the examples and our empirical findings, we use the following general-purpose matching function:

$$\Phi(\boldsymbol{X}) = \big(\Psi_1(X^{(1)}), \ldots, \Psi_d(X^{(d)}), \Psi_f(f(\boldsymbol{X}))\big), \quad (1)$$

where each Ψ_j is a low-cardinality quantile binning. We use 10 bins in all our experiments. This approach has proven empirically effective for most common tabular models and datasets. See V for further discussion of alternative binning strategies for decision trees that may reduce approximation gap.

Remark 3.11. The key takeaway is that one does not need to recover the full density ratio – a notoriously difficult task. Instead, it is enough to match the moments of a Φ that captures model disagreement. The quality of performance estimation depends on how well Φ approximates the disagreement space (see O for approximate alignment analysis).

Remark 3.12. Although our results are stated for population distributions ($P^\star$ and Q), the derivations are purely algebraic and therefore apply to any pair of distributions, including the empirical ones $\hat{P}^\star$ and $\hat{Q}$. This is a crucial feature of the framework: given an aligned feature map Φ, performance on a finite target sample $\hat{Q}$ can be estimated simply by matching empirical moments between $\hat{P}^\star$ and $\hat{Q}$. The resulting gap between the empirical risk $\epsilon_{\hat{P}^\star}$ and the population risk ϵ_Q is governed only by the sampling deviation between $\hat{\mu}_Q$ and μ_Q (See Q.3 for derivations). Thus, unlike standard analyses that require generalisation bounds on distributional discrepancy (e.g., Corollary 7 in Mansour et al. (2009)), our setting reduces entirely to moment matching.

4 EXPLAINING MODEL PERFORMANCE

Understanding differences between source and target domains is crucial for providing guidelines to monitor similar changes or to anticipate further shifts. This is often approached by learning a constrained map T that transforms the source data to the target data $T(\hat{P}_{\boldsymbol{X}}) = \hat{Q}_{\boldsymbol{X}}$., e.g., via optimal transport (Koebler et al., 2023).

A challenge with this problem is identifiability: without assumptions of invariance between the source and target data or knowledge about the causal structure, there are infinitely many plausible mappings T, making it impossible to identify which one generated the given target data. Therefore, we need to specify certain invariance assumptions to make this problem identifiable. Unlike the estimating and improving components, the explaining aspect of our methodology focuses specifically on explaining sparse shifts, which assumes there exists a subset $S \subset \{1, \ldots, d\}$ such that $P(\mathbf{X}_{\bar{S}}|\mathbf{X}_S, Y) = Q(\mathbf{X}_{\bar{S}}|\mathbf{X}_S, Y)$. Notably, the sparse joint shift generalizes both classical label shift and sparse covariate shift. See Tasche (2023b,a) for a comprehensive analysis of these shifts. In this context, explaining the causes involves identifying the subset of features responsible for generating the shift (Chen et al., 2022).

Proposition 4.1 (Identification of Sparse Shifts)**.** *Let a sparse shift from a source P to a target Q be generated by a unique minimal set S^*. For any candidate subset S, define a reweighted source distribution*

P'_S via the density ratio restricted to the variables in S (and Y for joint shift): $w^*_S(\boldsymbol{x}_S) = q(\boldsymbol{x}_S)/p(\boldsymbol{x}_S)$ or $w^*_S(\boldsymbol{x}_S, y) = q(\boldsymbol{x}_S, y)/p(\boldsymbol{x}_S, y)$. The generating set S^* is the unique set of minimal cardinality such that: $S^* = \text{argmin}_{S\subseteq\{1,\dots,d\}}\{|S| \mid \text{KL}(P'_S \,\|\, Q) = 0\}$.

See derivation in P. Our approach is to operationalise this idea using **EPA**. We iteratively test candidate feature subsets. For each subset, we use **EPA** to perform a histogram matching, reweighting the source data to align the distribution of those specific features with the target's. We then select the subset that yields the best overall approximation of the target data, as measured by an estimated KL distance. Specifically, among all subsets within $p\%$ (default $p = 5$) of the minimum, we then select the subset with the smallest size. We can also combine feature matchings with the target variable to capture shifts involving both inputs and labels. Refer to §5 and §3 for a step-by-step description of histogram matching and details on **EPA**'s explanations.

5 IMPROVING MODEL PERFORMANCE

To enhance the model's performance under the target distribution Q, we propose to leverage the reweighted data generated by the matching Φ used for target error estimation, i.e., matching the binned version of all features and model prediction.

The use of reweighting to improve model performance has been extensively studied. Shimodaira (2000) demonstrated that, for parametric models, reweighting can be beneficial when the model is misspecified (e.g., using a linear model when the true relationship is quadratic), but offers no advantage when the model is well-specified. However, recent works such as Byrd and Lipton (2019) and Zhai et al. (2022) report negative results, indicating that reweighting does not outperform classical training with uniform weights for high-capacity models. In contrast, Gogolashvili et al. (2023) argue that these negative findings are due to evaluations conducted on well-specified cases, and they show that reweighting can be effective even for high-capacity models if the model is misspecified.

Additionally, it is well-known that directly using the exact ratio of source and target densities as weights may not be optimal in practice. Shimodaira (2000) discusses this issue in detail, highlighting that while such weights are asymptotically unbiased, they can suffer from high variance in practice. For example, these weights may disproportionately emphasize a small subset of the data, potentially leading to poor model. This phenomenon reflects the classic bias-variance trade-off: accepting a small bias can lead to more stable and effective learning. Our reweighting technique addresses this trade-off by ensuring that the reweighted distribution remains close to the original distribution, thereby preventing the weights from concentrating on few points while still matching key statistics of the target data useful for estimation and adaptation.

When learning with reweighted data, most prior work focuses on training a completely new model (Byrd and Lipton, 2019; Zhai et al., 2022; Gogolashvili et al., 2023). In contrast, we propose adding a corrective term to the existing model f. We find this approach more effective than retraining a new model from scratch under reweighting, as it builds upon the valuable information already captured by f. Specifically, we employ a boosting procedure to minimize the error under the reweighted distribution $P^\star$.

Starting with $f_0 = f$, we iteratively update the model as $f_t = f_{t-1} + \alpha_t h_t$, where at each iteration t, $h_t \in \mathcal{H}$ is a base learner chosen from a class of functions $\mathcal{H}$ to minimize the weighted loss $h_t = \arg\min_{h\in\mathcal{H}} \sum_{i=1}^n \lambda_i \mathcal{L}\big(f_{t-1}(\mathbf{X}_i) + h(\mathbf{X}_i), Y_i\big)$, and α_t is a learning rate determined by: $\alpha_t = \arg\min_\alpha \sum_{i=1}^n \lambda_i \mathcal{L}\big(f_{t-1}(\mathbf{X}_i) + \alpha h_t(\mathbf{X}_i), Y_i\big)$. After M iterations, we obtain the improved model $\tilde{f} = f_M$, which is expected to perform better under $P^\star$ and, consequently, in the target domain Q. This procedure can be efficiently implemented using gradient boosting libraries like XGBoost (Chen and Guestrin, 2016), which natively support instance weights.

The following theorem provides a formal guarantee: if the performance gain from boosting on the reweighted source data exceeds a certain threshold, the adapted model is provably better than the baseline on the target distribution.

Theorem 5.1. *Let $\tilde{f}$ be the boosted model and define the empirical performance gain as $\Delta_{\text{emp}} = \epsilon_{\hat{P}^\star}(f) - \epsilon_{\hat{P}^\star}(\tilde{f})$. If*

$$\Delta_{\text{emp}} > 2(\text{disc}_{\mathcal{L}}(\hat{P}^\star_X, \hat{Q}_X) + \lambda_{\hat{P}^\star,\hat{Q}}),$$

then $\epsilon_{\hat{Q}}(\tilde{f}) \le \epsilon_{\hat{Q}}(f)$. Furthermore, if Δ_{emp} also exceeds a constant $B_{m,k,\delta}$ (see details in Q), then with probability at least $1-\delta$, we have

$$\epsilon_Q(\tilde{f}) < \epsilon_Q(f).$$

Our matching Φ is designed to make the discrepancy term $\text{disc}_\ell(P^\star_X, Q_X)$ small, while boosting decreases the reweighted error $\epsilon_{\hat{P}^\star}(\tilde{f})$. Theorem 5.1 then ensures these two effects combine into a provable improvement guarantee over the baseline model f. See proof in Q.

6 EXPERIMENTAL SETUP

In this section, we detail the experimental setup used to determine which method – relying only on source

data and target domain features – yields the best weights across three core objectives: ① estimating the target domain performance, ② identifying the features responsible for performance changes in sparse shift scenarios, and ③ adapting the model to the target data.

We use three shift settings: *sparse covariate*, *sparse joint* and *natural shifts*. For each case, we have source and target datasets. We split the source data into a training set, used to train the initial model f, and a calibration set, which is made accessible to each method for weight estimation and model adaptation. Similarly, we split the target data into an evaluation and calibration sets.

Sparse Covariate Shifts: We use the Adult (Dua and Graff, 2017), HELOC (FICO, 2018), and NHANES I (CDC, 2022) datasets. To create different source and target domains, we split the data based on the values of one feature at a time. For each continuous feature, we split the data at its median value. For each categorical feature, we exclude data belonging to one category. In either case, splitting results in two subsets, A and B. We then randomly select 10% of B and add it to A to form our *source data*, while the remaining 90% of B constitutes our *target data*.

For each dataset, the number of distribution shifts studied equals the number of features times the number of splits: two for continuous features (since we split at the median) and one for each category of discrete features. This setup results in **292 different distribution shifts**. A similar methodology was used in prior work (Chen et al., 2022) on a smaller scale with selected features, whereas we evaluate all possible feature combinations, aligning with recent tabular shift benchmarks (Gardner et al., 2023).

Sparse Joint Shifts: In this setting, both features and labels shift. We apply the same subset-splitting strategy from the covariate-shift setup but additionally modify the label distribution by randomly undersampling the source label mean to 0.2 and oversampling the target label mean to 0.5.

Natural Shifts: We use the Folktables dataset (Ding et al., 2021), which comprises U.S. Census data (50 states plus Puerto Rico) from multiple years (2014-2018), capturing both geographic and temporal shifts. The predictive task is to determine whether income exceeds $50,000.

As with the previous datasets, we define subsets A and B as different state-year combinations, thereby leveraging the inherent distribution shifts across time and location.

For further details on the datasets, such as size and feature count, see §G.2.

Given our focus on tabular data, we utilize XGBoost Chen and Guestrin (2016), a well-established and high-performing model for this type of data. We perform extensive fine-tuning to ensure optimal performance at every step involving training. Details of the tuning process can be found in §F.

In I, we give a detailed algorithmic description of all components of **EPA**. We use the following baselines for comparison:

- **SEES-C** and **SEES-D**: The two versions of the method introduced by Chen et al. (2022). SEES-C uses continuous features, and SEES-D uses a discretized version of the features to estimate the sample weights. We use the original implementation available on Github.
- **KMM** (Cortes et al., 2008; Gretton et al., 2008) using a linear kernel to match the mean of all features.
- **Domain Classifier Weighting (COV-CLF)**: A commonly used method (Sugiyama et al., 2012) for covariate shift that trains a classifier to distinguish between source and target domains. The predicted probabilities $\hat{p}(\mathbf{X})$ are used to derive sample weights as $\hat{p}(\mathbf{X})/(1-\hat{p}(\mathbf{X}))$. We use XGBoost as classifier.

7 EXPERIMENTS: ESTIMATING

We evaluate each method's ability to accurately estimate model performance on target domains and detect harmful distribution shifts using the following metrics:

1. *Estimation Inaccuracy*: Mean absolute error between estimated and true target errors (lower is better; ↓). True errors are defined as the absolute difference between predicted probabilities and true labels. The results are organized into increasing bins (B1-B4), which reflect the difference between the model error in the target data and the source data. A positive difference indicates a *harmful* shift; otherwise, the shift is considered *benign*.
2. *Detection Capability*: Assesses the method's ability to detect when a harmful shift occurs—that is, when the target error exceeds the source error. Power (proportion of harmful shifts correctly detected; higher is better ↑) and False Positive rate (FP; lower is better ↓).

Takeaway ①: As shown in Table 1, **EPA** significantly outperforms the baselines in both estimation inaccuracy and detection capability across all shift scenarios, achieving 30–70% better performance, particularly in the more harmful bins (B3, B4). It is followed by **SEES-C**, **SEES-D**, and then **COV-CLF**.

Table 1: Performance across different shift scenarios. **Tasks:** Estimation, Explanation, Improvement.

		Sparse Covariate						Sparse Joint						Natural					
Task	Method	B1	B2	B3	B4	Power/Prop	FP	B1	B2	B3	B4	Power/Prop	FP	B1	B2	B3	B4	Power/Prop	FP
Estimation↓	**EPA**	**0.007**	**0.010**	**0.014**	**0.009**	0.765	0.056	**0.031**	**0.034**	**0.034**	**0.064**	**1.000**	0.000	**0.047**	**0.049**	**0.055**	**0.045**	**0.978**	0.000
	KMM	0.014	0.020	0.158	0.111	0.009	**0.000**	0.172	0.130	0.227	0.310	0.442	0.000	–	–	–	–	–	–
	SEES-D	0.016	0.022	0.053	0.044	**0.791**	0.178	0.084	0.048	0.141	0.235	0.976	0.000	0.119	0.116	0.163	0.160	0.717	0.000
	SEES-C	0.014	0.016	0.012	0.017	0.696	0.100	0.057	0.029	0.127	0.222	0.976	0.000	0.134	0.135	0.186	0.323	0.607	0.000
	COV-CLF	0.015	0.015	0.028	0.026	0.757	0.171	0.128	0.098	0.203	0.249	0.935	0.000	0.107	0.098	0.167	0.281	0.835	0.000
Explan.	**EPA**	**0.70/0.75**	**0.71/0.82**	**1.00/1.00**	**0.86/0.98**	–	–	**0.65/0.58**	**0.86**/0.73	**1.00/1.00**	0.63/**0.58**	–	–	–	–	–	–	–	–
	SEES-D	0.28/0.35	0.29/0.35	0.50/0.59	0.86/0.83	–	–	0.25/0.20	0.23/0.09	0.56/0.50	0.25/0.18	–	–	–	–	–	–	–	–
	SEES-C	0.47/0.69	0.60/0.81	1.00/1.00	0.62/**0.98**	–	–	0.61/0.40	0.83/**0.82**	**1.00/1.00**	**0.87**/0.19	–	–	–	–	–	–	–	–
Improvement↑	**EPA**	13.86	11.89	10.71	24.48	0.979	–	17.65	17.74	8.90	6.21	0.983	–	13.67	13.90	20.07	42.27	0.981	–
	KMM	13.52	**13.09**	9.77	23.56	**0.997**	–	6.50	17.21	8.20	4.25	0.973	–	–	–	–	–	–	–
	SEES-D	**14.05**	12.03	9.15	23.52	0.976	–	16.59	17.12	8.41	4.40	0.973	–	9.68	12.46	19.27	41.23	0.879	–
	SEES-C	13.72	12.22	9.28	24.51	0.969	–	16.20	17.07	7.82	4.10	0.969	–	9.91	11.59	20.76	42.08	0.780	–
	COV-CLF	13.80	12.09	8.45	24.66	0.979	–	16.50	16.56	8.32	4.20	0.966	–	12.85	13.09	10.28	36.85	0.978	–
	SRC-UNI	14.01	12.05	9.27	24.17	0.969	–	16.55	16.92	8.45	4.18	0.973	–	12.76	12.90	10.22	37.40	0.984	–
	TAR-UNI	13.37	11.02	**11.93**	**26.57**	0.979	–	**37.27**	**49.14**	**42.07**	**49.65**	**1.000**	–	**24.87**	**27.43**	**34.68**	**50.62**	**1.000**	–

Bin ranges: Sparse Covariate: B1=(-0.149,-0.075], B2=(-0.075,0], B3=(0,0.106], B4=(0.106,0.211]; Sparse Joint: B1=(0.082,0.146], B2=(0.146,0.209], B3=(0.209,0.272], B4=(0.272,0.336]; Natural: B1=(0.042,0.141], B2=(0.141,0.24], B3=(0.24,0.338], B4=(0.338,0.437]. **Notes:** Estimation shows inaccuracy (lower is better); Explanation shows Acc/Corr pairs; Improvement shows average improvement (higher is better). Power/Prop column shows Power for Estimation and I-prop for Improvement. FP column shows false positives for Estimation only. Bold indicates best performance and underline second best performance within each task-shift. – indicates unavailable data. Note that **SRC-UNI** and **TAR-UNI** are not baselines but serve as control group

KMM is the poorest-performing method. In the natural shift setting, which reflects real-world data variability, **EPA** maintains strong accuracy and the highest power, while other methods degrade more noticeably. Due to scalability issues with larger datasets, **KMM** is omitted under natural shifts.

8 EXPERIMENTS: EXPLAINING

Next, we evaluate each method's explanation performance, focusing on *sparse covariate* and *sparse joint* shifts. In contrast to the natural shift scenario, these cases have well-defined ground truth explanations, as they correspond to the splitting features used to generate the shifts, allowing straightforward evaluation. Among the methods considered, only our approach (**EPA**) and the two baselines (**SEES-C** and **SEES-D**) provide feature-level explanations.

We measure the average detection rate of the features responsible for the shifts (*Acc*). To account for potential biases arising from correlated features, we also compute the correlation (*Corr*), which quantifies the relationship between the selected features and the ground truth features across the dataset (see details in §G).

Takeaway ②. As shown in Table 1, **EPA** outperforms the baselines across most bins and shift types, with up to 60% higher accuracy under sparse joint shifts. **SEES-C** generally outperforms **SEES-D**.

9 EXPERIMENTS: IMPROVING

We now turn to adaptation effectiveness, which measures how effectively the reweighted source data from each method can be used to adapt the model and improve its performance on the target domain. We report two relative improvement metrics to isolate the contribution of the reweighting scheme from the strenght of the starting model:

- *Average Improvement*: The relative percentage difference in performance between the adapted model and the original model on the target data (higher is better; ↑). Results are averaged across bins representing different performance shift intensities.
- *Improvement Proportion (I-prop)*: The proportion of shifts in which the adapted model outperforms the original model ranges from 0 to 1 (higher is better; ↑).

For model adaptation comparison, we include two additional baselines as **control methods**:

- **Source Calibration with Uniform Weights** (**SRC-UNI**): Uses the source calibration set with uniform weights as a naive approach without reweighting.
- **Target Calibration with Uniform Weights** (**TAR-UNI**): Uses the target calibration set with uniform weights as an idealized or strong baseline as it leverages ground truth labels from the target domain. Note that the target calibration set is not used for evaluation but comes from the same distribution as the one used for evaluation. It provides a reference for what can be achived with the same calibration budget when target labels are available.

Table 1 summarizes the results across three shift types: Sparse Covariate, Sparse Joint, and Natural shifts. A horizontal line separates the control baselines (**SRC-UNI**, **TAR-UNI**) from the actual baselines.

9.1 Sparse Covariate Shifts

Benign Shifts (Negative Bins) When distribution shifts benefit performance, most methods perform similarly, with differences in average improvement typically under 0.5% for bin B1.

Harmful Shifts (Positive Bins) When shifts become harmful, **TAR-UNI** (which leverages labeled target data) achieves the highest improvements. Among methods using only unlabeled target data, **EPA** stands out in the moderate bin $B_3 = (0, 0.106]$ with improvements close to **TAR-UNI**. In the highest bin $B_4 = (0.106, 0.211]$, all methods show comparable results, with **COV-CLF** having a slight edge in average improvement.

Interestingly, **SRC-UNI** (the naive approach) also performs well, supporting the findings from (Gogolashvili et al., 2023) that uniform weights may often suffice when the model is well specified. Next, we examine performance when the model is misspecified, where naive training provides less benefit, allowing a clearer comparison of each method's reweighting effectiveness.

Remark 9.1 (Model Misspecification). To isolate the benefits of reweighting, we conducted a study with a misspecified linear model, a scenario where the model class cannot easily approximate the true data-generating process (Shimodaira, 2000). The results show: naive fine-tuning with uniform weights (**SRC-UNI**) failed, yielding negligible or even negative gains on harmful shifts. In contrast, all importance reweighting methods delivered substantial improvements, demonstrating their ability to compensate for the model-data mismatch. Among them, **EPA provided the most consistent and significant gains** (see Table C), underscoring the critical role of effective reweighting for robust adaptation, especially when a model is misspecified.

9.2 Sparse Joint and Natural Shifts

Under *sparse joint shifts*, both the features and the labels change, creating more pronounced and inherently harmful distribution shifts even for a strong model like XGBoost. As expected, **TAR-UNI** achieves the highest performance gains by leveraging labeled target data. Among methods that only use unlabeled target data, **EPA** achieves the largest overall improvements, followed closely by **SEES-D**, **SEES-C**, and **COV-CLF**.

For *natural shifts*, **TAR-UNI** again demonstrates superior performance, especially under severe shifts. **EPA** maintains its position as the strongest method among those not using target labels across all bins, while **SEES-D**, **SEES-C**, perform similarly, with **COV-CLF** showing slightly better results.

Takeaway ③: When using boosting approaches, even simple uniform weighting on a new independent source dataset can yield modest improvements as the algorithm naturally adapts. However, as distributional shifts become more pronounced relative to the model class, our results demonstrate that reweighting becomes particularly beneficial, aligning with findings from Gogolashvili et al. (2023). Among the methods using only unlabelled target data, **EPA** performs best overall across diverse types of shifts, especially when the shift becomes harmful. Another observation from these experiments is that even suboptimal weights for estimating model performance can still lead to substantial model improvement.

10 ABLATION STUDIES

Appendices A-E present additional ablation experiments. These experiments demonstrate that the *correction* or *boosting* strategies (outlined in Section 5) consistently outperform *learning from scratch* under the studied shifts. We analyze which method generates reweighted data that aligns most closely with the target data. We conduct simulations demonstrating **EPA**'s stability advantage over **KMM** in importance weight estimation. Further experiments are included on linear models and regression problems.

11 COMPLEXITY ANALYSIS AND RUNNING TIME

We compare the computational costs of computing the sample weights for our method (**EPA**) against the direct baselines: **SEES-C** and **SEES-D**. In terms of complexity, **EPA** requires estimating a vector proportional to the number of features by minimizing the convex loss in (2.2). In contrast, the baselines aim to learn basis functions to approximate the density ratio $w_S(x, y) = \frac{dQ}{dP}(x_S, y)$. This approach involves a number of parameters dependent on the size of the basis class function, which can often exceed the dimensionality of the feature space.

Table 2: Runtime per method (in seconds), reporting the *mean*, standard deviation (*Std*), minimum (*Min*), and maximum (*Max*) across sparse-shift experiments on a MacBook Pro M2.

Metrics	*Mean*	*Std*	*Min*	*Max*
EPA	4.6	2.3	1.4	10.1
KMM	238	320	1.0	927
SEES-D	17.3	10.6	1.8	37.6
SEES-C	448	452	30	2026
COV-CLF	**2.4**	**1.2**	**0.5**	**4.4**

Table 2 provides the observed runtimes for each method during sparse-shift experiments. These results show that **EPA** is significantly faster than its closest baselines, particularly the best baseline **SEES-C**, which requires an order of magnitude more time. Notably, **COV-CLF**, which simply trains a domain classifier to distinguish between source and target labels

and converts the predictions into sample weights, is the quickest, requiring only 2 seconds at most.

12 CONCLUSION

We introduced **EPA**, an entropic projection-based framework to address three core challenges of distribution shift: ① estimating target domain performance, ② identifying the features driving the shift, and ③ adapting models to improve target domain performance, all using only unlabeled target data and the source data. Empirical evaluations across various shifts demonstrate that **EPA** outperforms state-of-the-art baselines while providing improved computational efficiency.

A key element of our framework is the choice of a feature map Φ that is well aligned with the model disagreement class under study. Future work could extend this framework beyond the tabular setting to unstructured domains, such as images, and to richer model classes, such as deep neural networks, where alignment may be achieved through learned representations. This direction is especially relevant for cases where explicit mappings are difficult or intractable, such as tree-based models. Another promising avenue is to employ hypothesis-specific discrepancy bounds rather than worst-case bounds over the entire function class, which could lead to tighter guarantees and improved adaptability.

DISCLAIMER

This paper was prepared for informational purposes by the Artificial Intelligence Research group of JP-Morgan Chase & Co., and its affiliates ("JP Morgan") and is not a product of the Research Department of JP Morgan. JP Morgan makes no representation, and warranty whatsoever, and disclaims all liability, for the completeness, accuracy or reliability of the information contained herein. This document is not intended as investment research or investment advice, or a recommendation, offer or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction, and shall not constitute a solicitation under any jurisdiction or to any person, if such solicitation under such jurisdiction or to such person would be unlawful.

© 2026 JP Morgan Chase & Co. All rights reserved.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Bachoc, F., Gamboa, F., Halford, M., Loubes, J.-M., and Risser, L. (2023). Explaining machine learning models using entropic variable projection. *Information and Inference: A Journal of the IMA*, 12(3):1686–1715.

Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. (2006). Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19.

Byrd, J. and Lipton, Z. (2019). What is the effect of importance weighting in deep learning? In *International conference on machine learning*, pages 872–881. PMLR.

CDC (1999-2022). National health and nutrition examination survey.

Chen, L., Zaharia, M., and Zou, J. Y. (2022). Estimating and explaining model performance when both covariates and labels shift. *Advances in Neural Information Processing Systems*, 35:11467–11479.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Cortes, C., Mohri, M., Riley, M., and Rostamizadeh, A. (2008). Sample selection bias correction theory. In *International conference on algorithmic learning theory*, pages 38–53. Springer.

Csiszár, I. (1984). Sanov property, generalized i-projection and a conditional limit theorem. *The Annals of Probability*, pages 768–793.

Ding, F., Hardt, M., Miller, J., and Schmidt, L. (2021). Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34.

Dua, D. and Graff, C. (2017). UCI machine learning repository.

FICO (2018). Fico. explainable machine learning challenge.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., and Lempitsky, V. (2016). Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35.

Gardner, J., Popovic, Z., and Schmidt, L. (2023). Benchmarking distribution shift in tabular data with tableshift. *Advances in Neural Information Processing Systems*, 36:53385–53432.

Gogolashvili, D., Zecchin, M., Kanagawa, M., Kountouris, M., and Filippone, M. (2023). When is im-

portance weighting correction needed for covariate shift adaptation? *arXiv preprint arXiv:2303.04020*.

Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., and Schölkopf, B. (2008). Covariate shift by kernel mean matching.

Koebler, A., Decker, T., Lebacher, M., Thon, I., Tresp, V., and Buettner, F. (2023). Towards explanatory model monitoring. In *XAI in Action: Past, Present, and Future Applications*.

Lipton, Z., Wang, Y.-X., and Smola, A. (2018). Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, pages 3122–3130. PMLR.

Liu, Y., Dong, J., Jiang, Z., Aloui, A., Li, K., Klein, H., Tarokh, V., and Carlson, D. (2023). Domain adaptation via rebalanced sub-domain alignment. *arXiv preprint arXiv:2302.02009*.

Long, M., Cao, Y., Wang, J., and Jordan, M. (2015). Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR.

Mansour, Y., Mohri, M., and Rostamizadeh, A. (2009). Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*.

Saerens, M., Latinne, P., and Decaestecker, C. (2002). Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural computation*, 14(1):21–41.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. (2019). Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244.

Sugiyama, M., Suzuki, T., and Kanamori, T. (2012). *Density ratio estimation in machine learning*. Cambridge University Press.

Sun, B. and Saenko, K. (2016). Deep coral: Correlation alignment for deep domain adaptation. In *Computer vision–ECCV 2016 workshops: Amsterdam, the Netherlands, October 8-10 and 15-16, 2016, proceedings, part III 14*, pages 443–450. Springer.

Tasche, D. (2023a). Invariance assumptions for class distribution estimation. *arXiv preprint arXiv:2311.17225*.

Tasche, D. (2023b). Sparse joint shift in multinomial classification. *arXiv preprint arXiv:2303.16971*.

Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. (2017). Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176.

Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., and Darrell, T. (2014). Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*.

Zhai, R., Dan, C., Kolter, Z., and Ravikumar, P. (2022). Understanding why generalized reweighting does not improve over erm. *arXiv preprint arXiv:2201.12293*.

CHECKLIST

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes. Proofs and detailed statements of the theorems, including their assumptions, are provided in the Appendix. In Section I, we give a detailed algorithmic description of all components of **EPA**.]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Code will be released after acceptance]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Code to reproduce all experiments will be released after acceptance]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes. Additional details can be found in Appendix, and code to reproduce all experiments will be released after acceptance]

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes. We report mean, standard deviation, min, and max for runtime statistics related to weight computation (Table 2). However, due to the extensive hyperparameter tuning detailed in Appendix F and the large number of shift scenarios evaluated (292 for sparse covariate shifts alone), computing comprehensive error bars (e.g., via multiple runs with full tuning) for the primary performance metrics in Tables 1 was computationally prohibitive: each full evaluation (with hyperparameter tuning) took approximately one week per method. However, we ensured robustness by evaluating a wide range of settings and reporting aggregated results across multiples shifts and several datasets.]

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes. The weight computation experiments were conducted on a MacBook Pro M2, as noted in the caption of Table 2. This figure provides mean, std, min, and max runtimes. Other experiments were run on AWS c5.4xlarge (16 vCPU, 32 GB RAM).]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Yes]

   (b) The license information of the assets, if applicable. [Not Applicable]

   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

   (d) Information about consent from data providers/curators. [Not Applicable]

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials

APPENDIX: TABLE OF CONTENTS

A. **Stability: EPA vs KMM**

- **Highlight: EPA** produces sample weights with lower variance and closer to the uniform distribution than **KMM**.

B. **Learning from Scratch vs. Correcting the Existing Model**

- Comparison of training new models from scratch vs. using correction terms.
- **Highlight:** Correction approach outperforms training from scratch.

C. **Experiments: Estimation and Improvement using a Linear Base**

- Evaluation of methods under sparse covariate shifts, joint shifts, and natural shifts using a Linear model.
- **Highlight: EPA** excels in estimation accurary; reweighting improves performance significantly over the traditional uniform weighting of the calibration set (**SRC-UNI**), in contrast to the case with tree base for the XGBoost model.

D. **Experiments on Regression Model**

- Evaluation on the California Housing Pricing dataset (Dua and Graff, 2017) using a linear base.
- **Highlight:** While the baselines **SEES-D**, and **SEES-C** are not available for regression, **EPA** significantly outperforms **COV-CLF** in estimation and improvement.

E. **Target Data Estimation**

- Comparison of methods in reweighting source data to approximate target data.
- **Highlight: EPA** consistently achieves the lowest mean and histogram difference norms.

F. **Hyperparameter Optimization for XGBoost Model**

- **Highlight:** Details of hyperparameter tuning strategies for XGBoost. We use the same strategy for all parts of our methodology requiring learning a model.

G. **Additional experiments details**

H **Additional related works**

I. **Algorithmic Description of EPA**

J-V **corresponds to all the theoretical analyses and proofs.**

A STABILITY: EPA VS KMM

In this section, we present a simulation to demonstrate the advantages of the **EPA** framework over **KMM**, complementing our earlier experiments on real-world datasets. As discussed, the regularization in **EPA** reduces variance in importance weight estimation. To validate this, we generate datasets multiple times under controlled conditions.

We define two distributions, $P \sim \mathcal{N}(0_p, 5)$ and $Q \sim \mathcal{N}(2 \times I_p, 1)$, and sample 1000 points from each. The target output is modeled as $Y = f(X) = \sum_{i=1}^{p} X_i$, with $p = 100$.

Both **KMM** and **EPA** are used to estimate importance weights, and we fit a linear model using the reweighted samples. Note that the optimal model remains f, even after the distribution shift. Performance is evaluated by the Relative Absolute Difference with respect to the optimal model by

$$\text{Relative Absolute Difference} = \frac{\text{MSE(linear_reweighted_method)} - \text{MSE(true_model)}}{\text{MSE(true_model)}}.$$

The experiment is repeated 500 times, and the results, illustrated in figure 1, highlight that **EPA** achieves zero variance, consistently recovering the true model, while **KMM** exhibits high variance. Moreover, in figure 2 we show the distribution of the KL distance between the sample weights of each method and the uniform weights. It shows that, as expected, the **EPA** weights are closer to the uniform distribution than those estimated by **KMM**.

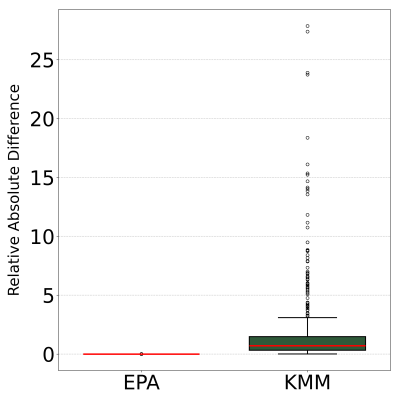

Figure 1: Comparison of each reweighted method's MSE with the true model's MSE. **EPA**'s variance is negligible, while **KMM**'s variance is large.

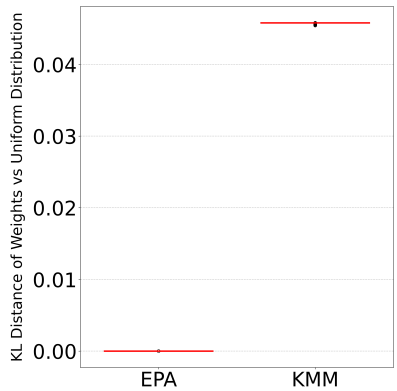

Figure 2: Distribution of the KL distance between sample weights and uniform weights. **EPA** is closer to uniform, as expected.

B LEARNING FROM SCRATCH VS. CORRECTING THE EXISTING MODEL

We evaluate the effectiveness of training a new model from scratch compared to our proposed method of adding corrective terms to an existing model (as described in Section 5). Table 3 presents the results. When comparing these findings with Table 1 (where correction was applied), we observe that training from scratch is significantly less effective across all experiments.

In the "learning from scratch" setting, most methods fail to improve performance and even degrade performance in most cases. The best-performing approach in this case is calibration with uniform weights (**SRC-UNI**), which produces results similar to the original model. This outcome is expected, as the calibration data follows the same distribution as the data used to train the original model.

This pattern holds across all other experiments. These findings underscore the limitations of training a new model on reweighted data in sparse shifts. In contrast, our correction approach effectively utilizes reweighted data to achieve better adaptation and performance improvements.

Table 3: Model improvement metrics for Sparse Covariate Shift when Learning a new model instead of correction on the reweighted data.

Metrics	*Average Improvement* (↑)				*I-prop* (↑)
Bins	(-0.091, -0.045]	(-0.045, 0]	(0, 0.153)	(0.153, 0.305)	All-Bins
EPA	-64.623	-62.970	-54.696	-51.008	-63.225
SEES-D	-58.976	-64.096	-61.102	-28.211	-60.621
SEES-C	-62.311	-59.761	-54.342	-32.398	-60.086
COV-CLF	-67.654	-61.408	-55.903	-82.166	-64.797
TAR-UNI	-4.520	-3.448	3.130	12.302	-3.366
SRC-UNI	-0.755	0.088	-0.593	1.825	-0.29

C EXPERIMENTS: ESTIMATION AND IMPROVEMENT USING A LINEAR BASE

This section evaluates the performance of the methods in terms of estimation and improvement under sparse covariate shifts, sparse joint shifts, and natural shifts using a linear base model. This setup highlights scenarios where the model class is underperforming, and reweighting is known to help (Gogolashvili et al., 2023) in mitigating the performance degradation.

C.1 Sparse Covariate Shift: Linear Model

Key Observations:

- **Estimation Inaccuracy (Table 4):**
  - **EPA** achieves the best estimation accuracy across all bins, with significantly lower errors compared to other methods. For detection power, **EPA** significantly outperforms baselines, achieving 0.848 compared to approximately 0.14 for others.
  - Baselines **SEES-C**, **SEES-D**, and **COV-CLF** struggle, particularly in cases of large performance shifts.
- **Model Improvement (Table 11):**
  - While most methods perform similarly in lower shift bins, **EPA** demonstrates the highest average improvement in bins with the largest shifts.
  - Uniform calibration (**SRC-UNI**) fails to deliver meaningful improvements, often resulting in negligible or negative performance gains.

Table 4: Estimation metrics for sparse covariate shifts using linear models.

Metrics	*Estimation Inaccuracy* (↓)				*Power* (↑)	*FP* (↓)
Bins	(-0.091, -0.045]	(-0.045, 0]	(0, 0.153)	(0.153, 0.305)]	All-Bins	All-Bins
EPA	0.004	0.005	0.004	0.015	0.848	0
KMM	0.012	0.024	0.067	0.230	0.	0.
SEES-D	0.048	0.068	0.025	0.189	0.141	0
SEES-C	0.054	0.066	0.023	0.242	0.141	0
COV-CLF	0.012	0.024	0.063	0.217	0.130	0

Table 5: Model improvement metrics for sparse covariate shift using linear models.

Metrics	*Average Improvement* (↑)				*I-prop* (↑)
Bins	(-0.091, -0.045]	(-0.045, 0]	(0, 0.153)	(0.153, 0.305)	All-Bins
EPA	56.662	32.590	**40.587**	**25.072**	**1.00**
KMM	56.645	32.586	39.592	23.067	0.97
SEES-D	56.664	**35.914**	39.171	21.187	0.712
SEES-C	56.679	34.569	39.353	21.138	0.801
COV-CLF	**56.690**	35.908	39.175	21.124	0.712
TAR-UNI	33.243	1.044	6.061	-8.091	0.805
SRC-UNI	4.380	1.764	2.241	-0.502	0.928

C.2 Sparse Joint Shift: Linear Model

Key Observations:

- **Estimation Inaccuracy (Table 6):**
  - **EPA** achieves near-perfect accuracy with negligible errors across all bins.
  - Baselines show large errors and low detection power (near 0).
- **Model Improvement (Table 7):**
  - All methods demonstrate similar improvement levels, but **EPA** lags slightly in improvement proportion (*I-prop*) compared to the baselines.
  - **SRC-UNI** continues to underperform, offering limited or negligible improvements.

Table 6: Estimation metrics for sparse joint shifts using linear models.

Metrics	*Estimation Inaccuracy* (↓)				*Power* (↑)	*FP* (↓)
Bins	(0.159, 0.167)	(0.167, 0.174)	(0.174, 0.182)	(0.182, 0.189)	All-Bins	All-Bins
EPA	0.000	0.000	0.000	0.001	1.00	0
SEES-D	0.235	0.240	0.225	0.241	0.120	0
SEES-C	0.222	0.242	0.216	0.210	0.027	0
COV-CLF	0.175	0.183	0.169	0.160	0.045	0

C.3 Natural Shift: Linear model

Key Observations:

- **Estimation Inaccuracy (Table 10):**
  - **EPA** demonstrates consistently strong performance, achieving low estimation errors across all bins.
  - Baselines such as **SEES-C**, **SEES-D**, and **COV-CLF** exhibit significantly higher errors.
- **Model Improvement (Table 9):**
  - All methods show comparable levels of improvement.

Table 7: Model improvement metrics for sparse joint shift using linear models.

Metrics	*Average Improvement* (↑)				*I-prop* (↑)
Bins	(0.159, 0.167)	(0.167, 0.174)	(0.174, 0.182)	(0.182, 0.189)	All-Bins
EPA	42.263	39.622	45.551	39.964	0.712
SEES-D	42.273	39.636	45.562	39.977	0.712
SEES-C	40.088	35.798	43.399	39.962	0.795
COV-CLF	36.673	30.403	39.310	39.928	1.00
SRC-UNI	1.775	1.256	1.698	2.358	1.00
TAR-UNI	4.709	4.328	3.586	3.182	0.945

– **TAR-UNI** and **SRC-UNI** fail to provide meaningful enhancements in lower bins. **TAR-UNI** gives the highest improvement on the most harmful bin.

Table 8: Estimation metrics for natural shifts using linear models.

Metrics	*Estimation Inaccuracy* (↓)				*Power* (↑)	*FP* (↓)
Bins	(-0.082, -0.041)	(-0.041, -0.001)	(-0.001, 0.16)	(0.16, 0.319)	All-Bins	All-Bins
EPA	0.018	0.034	0.005	0.009	0.963	0
SEES-D	0.157	0.278	0.017	0.044	0.	0
SEES-C	0.146	0.274	0.019	0.017	0.	0
COV-CLF	0.087	0.211	0.015	0.081	0.	0

Table 9: Model improvement metrics for natural shifts using linear models.

Metrics	*Average Improvement* (↑)				*I-prop* (↑)
Bins	(-0.082, -0.041)	(-0.041, -0.001)	(-0.001, 0.16)	(0.16, 0.319)	All-Bins
TAR-UNI	-7.162	-14.898	17.471	52.100	0.242
EPA	25.824	18.047	27.189	38.826	1.00
SEES-D	25.821	18.034	27.178	38.588	1.00
SEES-C	25.819	18.043	27.161	38.720	1.00
COV-CLF	25.814	18.025	27.161	38.575	1.00
SRC-UNI	1.317	0.866	0.568	-4.069	0.934

Takeaway: In terms of estimation, **EPA** is by far the most accurate method, significantly outperforming all baselines. Regarding improvement, all methods perform similarly overall, highlighting that reweighting can still drive performance gains even when error estimation is suboptimal—showing the learning algorithm's ability to benefit from reweighted data itself. Additionally, when the model is not high-performing, classical training with uniform weighting (**SRC-UNI**) fails to improve and can even degrade performance. Interestingly, revealing labels (**TAR-UNI**) can also underperform in this context, emphasizing the challenges of adaptation when the model class and data distribution are misaligned.

D EXPERIMENTS ON REGRESSION MODEL

In this section, we extend the evaluation of our method to regression problems using the California Housing Pricing dataset Dua and Graff (2017) following the previous section by using a Linear base. Among the baselines, only **COV-CLF** is applicable, as **SEES-C** and **SEES-D** are not designed for regression settings.

Key Observations:

- **Estimation Inaccuracy (Table 10):**
  - **EPA** achieves the lowest estimation errors across all bins, significantly outperforming the **COV-CLF** baseline.
  - The detection power of **EPA** is high (0.889), while **COV-CLF** fails to detect shifts at all (*Power* (↑) = 0).
- **Model Improvement (Table 11):**
  - **EPA** demonstrates consistent improvements across all bins, slightly outperforming the **COV-CLF** baseline.

Table 10: Estimation metrics for natural shifts using linear models.

Metrics	*Estimation Inaccuracy* (↓)				*Power* (↑)	*FP* (↓)
Bins	(-0.025, -0.018]	(-0.018, 0]	(0, 0.054]	(0.054, 0.092]	All-Bins	All-Bins
EPA	0.015	0.019	0.012	0.017	0.889	0
COV-CLF	0.023	0.013	0.074	0.028	0.	0

Table 11: Model improvement metrics for sparse covariate shift on California House Price using linear models.

Metrics	*Average Improvement* (↑)				*I-prop* (↑)
Bins	(-0.025, -0.018]	(-0.018, 0]	(0, 0.054]	(0.054, 0.092]	All-Bins
TAR-UNI	5.993	4.343	30.014	1.980	0.625
EPA	35.803	38.933	51.553	40.051	1.0
COV-CLF	35.701	38.693	51.498	40.004	1.0
SRC-UNI	0.680	1.256	1.730	0.666	0.938

Takeaway: Once again, **EPA** significantly outperforms the baseline in terms of error estimation. In terms of model improvement, **EPA** shows a slight advantage. This example also demonstrates that it is possible to have weights that are not necessarily accurate for estimation but can still contribute to improving model performance.

E TARGET DATA ESTIMATION

In this section, we evaluate how well each method reweights the source data to approximate the target data. Two metrics are used for evaluation:

- *Mean Matching for All Features*: Measures the discrepancy in feature means between the reweighted source data and the target data.
- *Histogram Matching for All Features*: Measures the discrepancy in feature distributions (histogram binning) computed marginally for each features independently and averaged. The number of bins is set at 10.

E.1 Sparse Covariate Shift

Table 12 summarizes the results for sparse covariate shifts. **EPA** achieves the lowest values for both metrics, indicating superior alignment with the target data. **SEES-C** performs moderately well, followed by **COV-CLF** and **SEES-D**, which show larger discrepancies.

Table 12: Comparison of mean and histogram difference norms for sparse covariate shift.

Method	*Mean Matching for All Features*	*Histogram Matching for All Features*
EPA	0.03	0.07
SEES-D	0.08	0.18
SEES-C	0.04	0.08
COV-CLF	0.06	0.11

E.2 Sparse Joint Shift

Table 13 presents the results for sparse joint shifts. Similar to the covariate shift setting, **EPA** demonstrates the best performance, achieving the smallest mean approximation. Regarding histogram differences **SEES-C** is slightly better, while **COV-CLF** and **SEES-D** perform worse, with higher values for both metrics.

Table 13: Comparison of mean and histogram difference norms for sparse joint shifts.

Method	*Mean Matching for All Features*	*Histogram Matching for All Features*
EPA	0.04	0.12
SEES-D	0.10	0.20
SEES-C	0.06	0.11
COV-CLF	0.12	0.17

Takeaway: These findings reinforce the effectiveness of **EPA** in approximating target distributions, making it a reliable choice for reweighting source data under distribution shifts.

F HYPERPARAMETER OPTIMIZATION FOR XGBOOST MODEL

To fine-tune the XGBoost model, a hyperparameter optimization process was conducted using 1000 iterations. The optimization strategy varied depending on the booster type (`gbtree` vs `gblinear`). The following details the process used to suggest and train the optimal hyperparameters. The hyperparameters were dynamically suggested during training based on the booster type:

- For `gbtree` boosters, the following hyperparameters were optimized:
  - `eta` (learning rate): Log-uniformly sampled in the range [0.01, 0.3].
  - `max_depth`: Integer value between 3 and 10.
  - `subsample`: Float value between 0.6 and 1.0.
  - `colsample_bytree`: Float value between 0.6 and 1.0.
  - `lambda` (L2 regularization): Log-uniformly sampled in the range [1e-3, 10.0].
  - `alpha` (L1 regularization): Log-uniformly sampled in the range [1e-3, 10.0].
  - `min_child_weight`: Integer value between 1 and 10.
- For the `gblinear` booster, the following hyperparameters were optimized:
  - `lambda` (L2 regularization): Log-uniformly sampled in the range [1e-3, 10.0].
  - `alpha` (L1 regularization): Log-uniformly sampled in the range [1e-3, 10.0].

The trained model was optimized for performance on the validation dataset, leveraging early stopping and dynamically selected hyperparameters to achieve the best results using the package Optuna Akiba et al. (2019).

G Additional experiments details

G.1 Correlation metric for the explainability part

Given two features, A and B, we determine their types and apply an appropriate correlation metric:

- **Both Continuous:** We use the *Pearson correlation coefficient*, which measures the linear relationship between two continuous variables:
$$\rho(A, B) = \frac{\text{cov}(A, B)}{\sigma_A \sigma_B},$$
where $\text{cov}(A, B)$ is the covariance of A and B, and σ_A and σ_B are their respective standard deviations.
- **Both Categorical:** We use *Cramér's V*, which is derived from the chi-squared (χ^2) statistic computed from the contingency table:
$$V = \sqrt{\frac{\chi^2/n}{\min(k_A - 1, k_B - 1)}},$$
where χ^2 is the chi-squared statistic, n is the total number of observations, and k_A and k_B are the number of categories in A and B, respectively.
- **One Continuous, One Binary:** We use the *Point-Biserial correlation coefficient*, which measures the correlation between a continuous variable and a binary categorical variable:
$$r_{pb} = \frac{\overline{X}_1 - \overline{X}_0}{s} \sqrt{\frac{n_1 n_0}{n(n-1)}},$$
where $\overline{X}_1$ and $\overline{X}_0$ are the means of the continuous variable for the two binary groups, s is the standard deviation of the continuous variable, and n_1, n_0, and n are the sizes of the respective groups and the total sample.

Finally, we take the absolute value of each metric to ensure that all correlation values fall within the range $[0, 1]$. We compute the values using all the dataset.

G.2 Dataset Details

Our experiments leverage a diverse array of publicly available, real-world datasets. These datasets vary in terms of sample size, feature dimensionality, and application domain, facilitating a comprehensive evaluation of our proposed framework under different conditions. Specific details for each dataset are provided below. For all datasets, categorical features were one-hot encoded, and the resulting feature dimensions are reported.

- **Adult:** This dataset, commonly used for income prediction (binary classification of whether income exceeds \$50K/yr), is sourced from the UCI Machine Learning Repository Dua and Graff (2017). It originally contains **14 features** (a mix of continuous and categorical) for **48,842 samples**. After one-hot encoding of categorical variables, the feature space expands to **109 dimensions**.
- **HELOC (Home Equity Line of Credit):** Provided by FICO (2018), this dataset is used for credit risk assessment. It comprises **9,871 samples** and **24 features**. The task is to predict whether a homeowner will default on their HELOC.
- **NHANES I (National Health and Nutrition Examination Survey):** This dataset consists of health-related data from the first National Health and Nutrition Examination Survey, made available by the CDC (2022). In our experiments, we use a version with **8,593 samples** and **18 features**, focusing on a mortality prediction task.
- **California Housing Prices:** Also from the UCI Machine Learning Repository Dua and Graff (2017), this dataset contains **20,640 samples** and **9 features**. It is a standard benchmark for regression tasks, where the objective is to predict the median house value in California districts.

Folktables (ACS-based Tasks)

The Folktables suite Ding et al. (2021) provides benchmark tasks derived from the U.S. Census Bureau's American Community Survey (ACS). These datasets allow for the study of model performance under various demographic shifts. Table 14 summarizes the characteristics of several tasks available within this suite.

Task Name	No. of Features (Original)	No. of Datapoints
ACSIncome	10	1,664,500
ACSPublicCoverage	19	1,138,289
ACSMobility	21	620,937
ACSEmployment	17	3,236,107
ACSTravelTime	16	1,466,648

Table 14: Overview of selected Folktables tasks based on ACS data. Feature counts are prior to one-hot encoding.

For our experiments involving natural distribution shifts, we primarily utilize the **ACSIncome** task from Folktables. While the original ACSIncome task includes 10 features, after one-hot encoding of its categorical variables, the dimensionality increases to approximately **284 features**. The full ACSIncome dataset across all available years and states contains over 1.6 million records. In our experimental setup, we generate source and target domains by selecting data for specific U.S. states and years. This typically results in experimental datasets ranging from **30,000 to 100,000 samples**, though some state-year combinations may yield larger datasets.

H ADDITIONAL RELATED WORKS DISCUSSION

One important related body of literature is representation learning. Methods within this field have progressed in managing unstructured data such as images by using mean-matching techniques to align learned embedding features for adapting to distributional changes, as highlighted in works like (Long et al., 2015; Tzeng et al., 2014, 2017; Long et al., 2015; Ganin et al., 2016; Sun and Saenko, 2016) or sub-domain alignment methods (Liu et al., 2023). However, the EPA approach differs from these works. These approaches primarily address domain adaptation alone. In contrast, EPA is specifically designed to tackle three interrelated problems simultaneously: (1) performance evaluation, (2) explanation of shifts, and (3) model adaptation. Thus, these earlier methods do not fully compare across all three aspects. In addition, direct empirical comparisons with these methods are challenging, as many are tailored to specific model architectures, such as deep neural networks with unique loss functions and feature alignment strategies—whereas EPA is not restricted to a particular model type. Rather than replacing existing models, EPA integrates a corrective component using estimated importance weights, setting it apart in its goals and operation. Putting differently, EPA is a post-hoc method. However, extending EPA to fields like computer vision or textual data would require an additional phase for identifying meaningful feature representations where mean-matching could be effectively applied. Developing this extension is complex and separate from EPA's core function and thus is not covered in this paper. As stated in our conclusion, we have left this extension for future exploration.

Given EPA's core mechanism of deriving importance weights, its contributions are particularly relevant to the literature on reweighting techniques for distribution shift. Within this domain, SEES (Chen et al., 2022) is our main competitor and best among the reweighting techniques as demonstrated in their paper. Our work therefore emphasizes comparison with such state-of-the-art reweighting approaches to clearly benchmark EPA's advancements in robust weight estimation, along several additional dimensions such as performance prediction, explanation of shifts, and model adaptation.

Lastly, EPA can be contrasted with distributionally robust optimization (DRO) strategies (Sagawa et al., 2019). DRO methods typically aim to optimize model performance under worst-case distributional shifts, often by reweighting data from predefined groups. While effective for enhancing worst-case generalization, DRO approaches require these group definitions. Specifying such groups comprehensively can be challenging, may not capture all relevant types of shifts, or the necessary groupings might not be apparent in all domains (for instance, while image datasets sometimes allow for grouping by attributes like gender or hair color, this approach has its limitations and is not universally applicable). Furthermore, DRO techniques typically do not offer the direct target performance estimation or the explanatory insights inherent to the EPA framework.

I ALGORITHMIC DESCRIPTION OF EPA

This section provides a detailed algorithmic description of the proposed Entropic Projection Alignment (**EPA**) framework. The following algorithms outline the core components and applications of **EPA**:

- Algorithm 1 describes the core **EPA** procedure for computing importance weights and the reweighted source distribution.
- Algorithm 2 details how **EPA** is used to estimate model performance on an unlabeled target domain.
- Algorithm 3 outlines the process for identifying feature subsets responsible for distribution shifts, utilizing Algorithm 5 for discrepancy measurement.
- Algorithm 4 shows the methodology for adapting an existing model to improve its performance in the target domain using **EPA**-derived weights.
- Algorithm 5 provides the method for computing the histogram-based symmetric KL divergence used in the shift explanation process.
- Algorithm 6 provide the method for computing the target estimation with label prevalence.

Algorithm 1 Core **EPA** (Entropic Projection Alignment)

Input:
Source dataset: $D_{\text{src}} = \{(X_i, Y_i)\}_{i=1}^n$
Target dataset: $D_{\text{tar}} = \{X_i^t\}_{i=1}^m$
Feature mapping function: $\Phi(X)$
Output:
Importance weights: $\lambda = (\lambda_1, \ldots, \lambda_n)$
Reweighted source distribution: P^*
Procedure:

1: **Step 1:** Define the convex loss function:

$$\text{Loss}(\xi) = \log\left(\frac{1}{n}\sum_{i=1}^{n}\exp(\langle\Phi(X_i), \xi\rangle)\right) - \langle\xi, \mu_t\rangle$$

where $\mu_t = \frac{1}{m}\sum_{j=1}^{m}\Phi(X_j^t)$ is the mean of the feature mapping over the target data.

2: **Step 2:** Optimize to find the optimal parameter vector ξ^*:

$$\xi^* = \arg\min_{\xi} \text{Loss}(\xi)$$

using convex optimization (e.g., L-BFGS optimizer).

3: **Step 3:** Compute importance weights and define the reweighted source distribution:

$$\lambda_i = \frac{\exp(\langle\xi^*, \Phi(X_i)\rangle)}{\sum_{j=1}^{n}\exp(\langle\xi^*, \Phi(X_j)\rangle)}, \quad \text{for } i = 1, \ldots, n$$

$$P^* = \sum_{i=1}^{n}\lambda_i\delta_{(X_i, Y_i)}$$

4: **Return:** Importance weights λ and reweighted source distribution P^*.

Remark I.1. Assuming d features and the scalar score $f(X)$ are discretized into B bins, then EPA optimizes $k = (d+1)(B-1)$ parameters. Evaluating the objective in (2.2) and its gradient costs $O(nk) = O(n\,d\,B)$, so I L-BFGS iterations yield total complexity $O(I\,n\,d\,B)$, linear in the number of source samples.

Algorithm 2 Target Performance Estimation using **EPA**

Input:
Source dataset: $D_{\text{src}} = \{(X_i, Y_i)\}_{i=1}^n$
Target dataset: $D_{\text{tar}} = \{X_i^t\}_{i=1}^m$
Predictive model: $f : \mathcal{X} \to \mathcal{Y}$
Loss function: $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$

Output:
Estimated target error: $\hat{E}_{\text{tar}}(f)$

Procedure:

1: **Step 1:** Choose quantile bins for each feature and for the scalar score $f(X)$. Let $\psi_j(X^{(j)})$ and $\psi_f(f(X))$ denote the corresponding bin-indicator blocks with one redundant bin removed, and define:

$$\Phi_{\text{est}}(X) = [\psi_1(X^{(1)}), \ldots, \psi_d(X^{(d)}), \psi_f(f(X))]$$

2: **Step 2:** Apply Algorithm 1 with Φ_{est} to obtain importance weights:

$$\lambda = (\lambda_1, \ldots, \lambda_n) \leftarrow \text{Core-}\mathbf{EPA}(D_{\text{src}}, D_{\text{tar}}, \Phi_{\text{est}})$$

3: **Step 3:** Calculate the estimated target error using importance weights:

$$\hat{E}_{\text{tar}}(f) = \sum_{i=1}^{n} \lambda_i \cdot \mathcal{L}(f(X_i), Y_i)$$

4: **Return:** Estimated target error $\hat{E}_{\text{tar}}(f)$.

Algorithm 3 Shift Explanation using **EPA**

Input:
Source dataset: $D_{\text{src}} = \{(X_i, Y_i)\}_{i=1}^n$
Target dataset (unlabeled features): $D_{\text{tar}} = \{X_i^t\}_{i=1}^m$
Candidate feature subsets: $\mathcal{S} = \{S_1, S_2, \ldots, S_k\}$, where $S_j \subseteq \{1, \ldots, d\}$

Output:
Feature subset S^* that best explains the distribution shift.

Procedure:

1: **Step 1:** For each candidate subset $S_j \in \mathcal{S}$:

1.1 Define feature mapping using only features in subset S_j:

$$\Phi_{S_j}(X) = [\psi_s(X^{(s)})]_{s \in S_j}$$

1.2 Apply Algorithm 1 with Φ_{S_j} to obtain the reweighted source distribution $P^*_{S_j}$(features only) and corresponding weights λ_{S_j}:

$$(P^*_{S_j}(\text{features only}), \lambda_{S_j}) \leftarrow \text{Core-}\mathbf{EPA}(D_{\text{src}}, D_{\text{tar}}, \Phi_{S_j})$$

1.3 Compute discrepancy $D(S_j)$ between the reweighted source feature distribution (using $P^*_{S_j}$ or weights λ_{S_j} on D_{src} features) and the target feature distribution Q_X^{tar} (empirical distribution of D_{tar}). Typically using Algorithm 5:

$$D(S_j) = \text{SymmetricKL}(\{(X_i, (\lambda_{S_j})_i)\}_{i=1}^n, D_{\text{tar}} \text{ for features in } S_j)$$

2: **Step 2:** Select the feature subset that minimizes the discrepancy:

$$S^* = \arg\min_{S_j \in \mathcal{S}} D(S_j)$$

3: **Return:** Optimal feature subset S^*.

Algorithm 4 Model Adaptation using **EPA** (Boosting Approach)

Input:
Source dataset: $D_{\text{src}} = \{(X_i, Y_i)\}_{i=1}^n$
Target dataset: $D_{\text{tar}} = \{X_i^t\}_{i=1}^m$
Initial model: $f : \mathcal{X} \to \mathcal{Y}$
Loss function: $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$
Base learner class: $\mathcal{H}$
Number of boosting iterations: M

Output:
Adapted model: $\tilde{f} : \mathcal{X} \to \mathcal{Y}$

Procedure:

1: **Step 1:** Use the same histogram-based matching map as in estimation:

$$\Phi_{\text{adapt}}(X) = [\psi_1(X^{(1)}), \ldots, \psi_d(X^{(d)}), \psi_f(f(X))]$$

2: **Step 2:** Apply Algorithm 1 with Φ_{adapt} to obtain importance weights:

$$\lambda = (\lambda_1, \ldots, \lambda_n) \leftarrow \text{Core-}\mathbf{EPA}(D_{\text{src}}, D_{\text{tar}}, \Phi_{\text{adapt}})$$

3: **Step 3:** Initialize the adapted model:

$$f_0 = f$$

4: **Step 4:** For each iteration $t = 1, 2, \ldots, M$:

4.1 Fit base learner $h_t \in \mathcal{H}$ by minimizing the weighted loss:

$$h_t = \arg\min_{h \in \mathcal{H}} \sum_{i=1}^n \lambda_i \cdot \mathcal{L}(f_{t-1}(X_i) + h(X_i), Y_i)$$

4.2 Determine the learning rate (or step size) α_t:

$$\alpha_t = \arg\min_{\alpha \in \mathbb{R}} \sum_{i=1}^n \lambda_i \cdot \mathcal{L}(f_{t-1}(X_i) + \alpha \cdot h_t(X_i), Y_i)$$

4.3 Update the model:

$$f_t = f_{t-1} + \alpha_t \cdot h_t$$

5: **Step 5:** Set the final adapted model:

$$\tilde{f} = f_M$$

6: **Return:** Adapted model $\tilde{f}$.

Note: In practice, gradient boosting libraries like XGBoost can implement this procedure by accepting instance weights λ_i.

Algorithm 5 Histogram-based Symmetric KL Divergence

Input:

Reweighted source data (features and weights): $D_s = \{(X_i, \lambda_i)\}_{i=1}^n$, where $X_i \in \mathbb{R}^d$

Target data (features): $D_t = \{X_j^t\}_{j=1}^m$, where $X_j^t \in \mathbb{R}^d$

Number of bins for histogram: B (default: 10)

Smoothing parameter: ε (default: 10^{-6})

Output:

Average Symmetric KL divergence over all features: D_{KL}^{avg}

Procedure:

Initialize: Total divergence $D_{KL}^{\text{total}} \leftarrow 0$

For each feature dimension $k \in \{1, \ldots, d\}$:

Step 1: Extract k-th feature values from source and target:

Source values: $x_s^{(k)} = [X_1^{(k)}, X_2^{(k)}, \ldots, X_n^{(k)}]$ (with corresponding weights λ_i)

Target values: $x_t^{(k)} = [X_1^{t,(k)}, X_2^{t,(k)}, \ldots, X_m^{t,(k)}]$

Step 2: Compute empirical probability distributions $\mathbf{p}^{(k)}$ and $\mathbf{q}^{(k)}$ using histograms:

Determine common bin edges for the k-th feature based on the combined range of $x_s^{(k)}$ and $x_t^{(k)}$.

$\mathbf{p}^{(k)} = [p_1^{(k)}, \ldots, p_B^{(k)}]$: weighted histogram of $x_s^{(k)}$ using weights λ_i.

$\mathbf{q}^{(k)} = [q_1^{(k)}, \ldots, q_B^{(k)}]$: unweighted histogram of $x_t^{(k)}$.

Step 3: Apply smoothing and normalize distributions:

$p_b^{(k)} \leftarrow p_b^{(k)} + \varepsilon$ and $q_b^{(k)} \leftarrow q_b^{(k)} + \varepsilon$ for all bins $b = 1, \ldots, B$.

Normalize $\mathbf{p}^{(k)}$ such that $\sum_b p_b^{(k)} = 1$.

Normalize $\mathbf{q}^{(k)}$ such that $\sum_b q_b^{(k)} = 1$.

Step 4: Calculate symmetric KL divergence for the k-th feature:

$KL(\mathbf{p}^{(k)} \parallel \mathbf{q}^{(k)}) = \sum_{b=1}^B p_b^{(k)} \log\left(\frac{p_b^{(k)}}{q_b^{(k)}}\right)$

$KL(\mathbf{q}^{(k)} \parallel \mathbf{p}^{(k)}) = \sum_{b=1}^B q_b^{(k)} \log\left(\frac{q_b^{(k)}}{p_b^{(k)}}\right)$

$D_{KL}^{(k)} = KL(\mathbf{p}^{(k)} \parallel \mathbf{q}^{(k)}) + KL(\mathbf{q}^{(k)} \parallel \mathbf{p}^{(k)})$

Step 5: Accumulate total divergence:

$D_{KL}^{\text{total}} \leftarrow D_{KL}^{\text{total}} + D_{KL}^{(k)}$

$D_{KL}^{\text{avg}} = D_{KL}^{\text{total}}/d$

Return: Average Symmetric KL divergence D_{KL}^{avg}.

Algorithm 6 Target Performance Estimation with Label Prevalence Optimization (EPA-LPO)

Input:

Source dataset: $D_{\text{src}} = \{(X_i, Y_i)\}_{i=1}^{n}$ (where $Y_i \in \{0, 1\}$)

Target features: $D_{\text{tar_X}} = \{X_j^t\}_{j=1}^{m}$ ▷ Unlabeled target data

Predictive model: $f : \mathcal{X} \to \mathbb{R}$ (e.g., a scorer)

Loss function: $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$

Candidate target prevalences for $Y = 1$: $\mathcal{Q}_1 = \{q_1^{(1)}, q_1^{(2)}, \ldots, q_1^{(K)}\}$ ▷ Grid like $[0.05, 0.1, \ldots, 0.95]$

Discrepancy metric: Discrepancy$(\cdot, \cdot)$ ▷ Typically Symmetric KL Divergence

Output:

Estimated target error: $\hat{E}_{\text{tar}}(f)$

Estimated target label prevalence for $Y = 1$: q_1^*

Procedure:

Define base feature mapping (features and model predictions): $\Phi_{\text{base}}(X) = [X^{(1)}, \ldots, X^{(d)}, f(X)]$

Compute target moments for $\Phi_{\text{base}}(X)$ from $D_{\text{tar_X}}$: $\mu_{\text{base_target}} = \frac{1}{m} \sum_{j=1}^{m} \Phi_{\text{base}}(X_j^t)$

Initialize $D_{\min} \leftarrow \infty$, $q_1^* \leftarrow$ None, $\lambda^* \leftarrow$ None

for each candidate prevalence $q_1 \in \mathcal{Q}_1$ **do** ▷ Try different label prevalence values

  Define augmented feature mapping including the label Y: $\Phi_{\text{aug}}(X, Y) = [\Phi_{\text{base}}(X), Y]$

  Construct the target moment vector for Φ_{aug}: $\mu_{\text{aug_target}}(q_1) = [\mu_{\text{base_target}}, q_1]$ ▷ Hypothesize target moments with current q_1

  Solve for $\xi^*(q_1)$ to get weights $\lambda_i(q_1)$: ▷ Entropic projection step

    $\text{Loss}(\xi; q_1) = \log\left(\frac{1}{n} \sum_{i=1}^{n} \exp(\langle \Phi_{\text{aug}}(X_i, Y_i), \xi \rangle)\right) - \langle \xi, \mu_{\text{aug_target}}(q_1) \rangle$

    $\xi^*(q_1) = \arg\min_{\xi} \text{Loss}(\xi; q_1)$ ▷ Convex optimization

    $\lambda_i(q_1) = \frac{\exp(\langle \xi^*(q_1), \Phi_{\text{aug}}(X_i, Y_i) \rangle)}{\sum_{k=1}^{n} \exp(\langle \xi^*(q_1), \Phi_{\text{aug}}(X_k, Y_k) \rangle)}$, for $i = 1, \ldots, n$

  Calculate discrepancy $D(q_1)$ between reweighted source and target distributions:

    $D(q_1) = \text{Discrepancy}(\{(X_i, \lambda_i(q_1))\}_{i=1}^{n}, D_{\text{tar_X}})$ ▷ Measure alignment quality

  if $D(q_1) < D_{\min}$ **then**

    $D_{\min} \leftarrow D(q_1)$

    $q_1^* \leftarrow q_1$ ▷ Store best prevalence estimate

    $\lambda^* \leftarrow \lambda(q_1)$

  end if

end for

Calculate the final estimated target error using optimal weights: $\hat{E}_{\text{tar}}(f) = \sum_{i=1}^{n} \lambda_i^* \cdot \mathcal{L}(f(X_i), Y_i)$ ▷ Weighted average of errors

Return: Estimated target error $\hat{E}_{\text{tar}}(f)$ and estimated prevalence q_1^*.

J THEORETICAL ANALYSIS OF EPA

J.1 Proof of Proposition 2.2

To prove Proposition 2.2, we begin with the fundamental theorem governing the minimization of KL divergence relative to a prior measure under linear moment constraints. This result corresponds to Theorem A.1 in Bachoc et al. (2023), stated there without proof as a consequence of the theorems of Csiszár (1984).

Theorem J.1 (Minimum KL Divergence Solution)**.** *Let $(E, \mathcal{B}(E))$ be a measurable space and Q a probability measure on E. Consider $t \in \mathbb{R}^k$ and a measurable function $\Phi : E \to \mathbb{R}^k$. We assume that, for $v \in \mathbb{R}^k, b \in \mathbb{R}$, $Q(\{x \in E; \langle v, \Phi(x)\rangle = b\}) = 1$ if and only if $v = 0$ and $b = 0$. Let $\mathbb{P}_{\Phi,t}$ be the set of all probability measures P on $(E, \mathcal{B}(E))$ such that $\int_E \Phi(x)dP(x) = t$. Assume that $\mathbb{P}_{\Phi,t}$ contains a probability measure that is mutually absolutely continuous with respect to Q.*

For a vector $\xi \in \mathbb{R}^k$, let $Z(\xi) := \int_E e^{\langle \xi, \Phi(x)\rangle} dQ(x)$. We assume that the set on which Z is finite is open. Define now $\xi(t)$ as the unique minimizer of the strictly convex function $H(\xi) - \langle \xi, t\rangle$, where $H(\xi) := \log Z(\xi)$.

Then, the solution to the optimization problem

$$Q_t := \arg \inf_{P \in \mathbb{P}_{\Phi,t}} KL(P, Q)$$

exists and is unique. Furthermore, it can be computed as

$$Q_t = \frac{\exp(\langle \xi(t), \Phi\rangle)}{Z(\xi(t))} Q.$$

This means its Radon-Nikodym derivative is $\frac{dQ_t}{dQ}(x) = \frac{\exp(\langle \xi(t), \Phi(x)\rangle)}{Z(\xi(t))}$.

Proof. Assume P is absolutely continuous with respect to Q, we can frame the problem as minimising the functional $I[p] = \int_E p \log p \, dQ$ subject to the constraints $\int_E p \, dQ = 1$ and $\int_E \Phi p \, dQ = t$. We form the Lagrangian functional with multipliers λ and $\xi \in \mathbb{R}^k$:

$$L[p] = \int_E p \log p \, dQ - \lambda \left(\int_E p \, dQ - 1 \right) - \left\langle \xi, \int_E \Phi p \, dQ - t \right\rangle.$$

Combining terms under a single integral:

$$L[p] = \int_E \left[p(x) \log p(x) - \lambda p(x) - \langle \xi, \Phi(x)\rangle p(x) \right] dQ(x) + \lambda + \langle \xi, t\rangle.$$

To find the stationary point, we take the functional derivative with respect to $p(x)$ and set it to zero. This is equivalent to finding the point-wise minimum of the integrand:

$$\frac{\partial}{\partial p} \left[p \log p - \lambda p - \langle \xi, \Phi\rangle p \right] = \log p + 1 - \lambda - \langle \xi, \Phi\rangle = 0.$$

Solving for $p(x)$ reveals its exponential form:

$$p(x) = \exp(\lambda - 1 + \langle \xi, \Phi(x)\rangle).$$

The multipliers are determined by the constraints. The normalization constraint $\int p(x)dQ(x) = 1$ implies $e^{\lambda - 1} = 1/Z(\xi)$, where $Z(\xi)$ is the partition function. This gives the density form:

$$p_\xi(x) = \frac{e^{\langle \xi, \Phi(x)\rangle}}{Z(\xi)}.$$

Next, consider $\int_E \Phi(x) p_\xi(x) \, dQ(x) = t$.

$$\int_E \Phi(x) \frac{\exp(\langle \xi, \Phi(x)\rangle)}{Z(\xi)} \, dQ(x) = t.$$

We can relate the integral to the gradient of $Z(\xi)$. Since the set where $Z(\xi)$ is finite is open, we can differentiate under the integral sign:

$$\nabla_\xi Z(\xi) = \nabla_\xi \int_E e^{\langle \xi, \Phi(x) \rangle} dQ(x) = \int_E \Phi(x) e^{\langle \xi, \Phi(x) \rangle} dQ(x).$$

The moment constraint becomes:

$$\frac{1}{Z(\xi)} \nabla_\xi Z(\xi) = t.$$

Now, consider $H(\xi) = \log Z(\xi)$. Its gradient is $\nabla_\xi H(\xi) = \frac{\nabla_\xi Z(\xi)}{Z(\xi)}$. Thus, the Lagrange multiplier vector ξ, which we now denote $\xi(t)$, must satisfy:

$$\nabla_\xi H(\xi) = t.$$

This is precisely the first-order condition for minimizing $H(\xi) - \langle \xi, t \rangle$. By the theorem's assumption, this function is strictly convex and thus has a unique minimizer $\xi(t)$. □

Proof of Proposition 2.2. The **EPA** is a direct, discrete instance of Theorem J.1. The key is to replace the general measures and integrals with their empirical counterparts.

The solution from Theorem J.1 translates directly. The optimal distribution P^* must be absolutely continuous with respect to $\hat{P}$, meaning it is also a discrete distribution on the same n samples, defined by a vector of weights $\lambda^* = (\lambda_1^*, \ldots, \lambda_n^*)$. The Radon-Nikodym derivative $\frac{dP^*}{d\hat{P}}$ at a sample point X_i is $\frac{\lambda_i^*}{1/n} = n\lambda_i^*$. Applying the theorem's solution form:

$$n\lambda_i^* = \frac{\exp(\langle \xi^*, \Phi(X_i) \rangle)}{Z(\xi^*)},$$

where the partition function is now a sum: $Z(\xi^*) = \int e^{\langle \xi^*, \Phi(x) \rangle} d\hat{P}(x) = \frac{1}{n} \sum_{j=1}^n e^{\langle \xi^*, \Phi(X_j) \rangle}$. Substituting $Z(\xi^*)$ gives the explicit formula for the optimal weights:

$$\lambda_i^* = \frac{1}{nZ(\xi^*)} \exp(\langle \xi^*, \Phi(X_i) \rangle) = \frac{\exp(\langle \xi^*, \Phi(X_i) \rangle)}{\sum_{j=1}^n \exp(\langle \xi^*, \Phi(X_j) \rangle)}.$$

The vector ξ^* is the unique parameter that satisfies the moment constraint: $\sum_{i=1}^n \lambda_i^* \Phi(X_i) = \hat{\mu}_Q$. □

J.2 EPA's weights minimises variance upper-bound by construction

Interestingly, **EPA**'s definition in 2.1 can be rewritten as the following optimisation problem:

Definition J.2 (EPA - empirical weight version)**.** EPA solves the I-projection of $\hat{P}$ onto the affine moment set:

$$\lambda^* \in \arg\min_{\lambda \in \Delta_n} \mathrm{KL}(\lambda \| u) \quad \text{s.t.} \quad \sum_{i=1}^n \lambda_i \, \Phi(X_i) = \widehat{\mu}_Q. \tag{2}$$

Equivalently, (2) maximizes Shannon entropy on the feasible set, since $\mathrm{KL}(\lambda \| u) = \log n - H(\lambda)$.

Let $S = (X_{1:n}, \tilde{X}_{1:m})$ denote the (unlabeled) features from source and target. EPA produces weights

$$\lambda(S) \in \Delta_n, \qquad \sum_{i=1}^n \lambda_i \, \Phi(X_i) = \widehat{\mu}_Q, \qquad \lambda_i \propto \exp\{\langle \xi, \Phi(X_i) \rangle\},$$

which depend only on S (not on labels). For any $f \in \mathcal{H}$, define the centered label-noise term

$$\Delta_f \;:=\; \sum_{i=1}^n \lambda_i \Big(\ell(f(X_i), Y_i) - \mathbb{E}[\ell(f(X_i), Y_i) \mid X_i] \Big).$$

Write u for the uniform weights $u_i = 1/n$, and set $N_{\text{eff}}(\lambda) := 1/\sum_{i=1}^n \lambda_i^2$.

Theorem J.3 (Variance upper bound via KL; EPA minimises the bound)**.** *Assume $\ell \in [0,1]$. Conditionally on S, for any mean-matching $w \in \Delta_n$ satisfying $\sum_i \lambda_i \Phi(X_i) = \widehat{\mu}_Q$,*

$$\mathrm{Var}(\Delta_f \mid S; \lambda) \;\le\; \tfrac{1}{4}\sum_{i=1}^{n} \lambda_i^2 \;\le\; \tfrac{1}{4}\left(\frac{1}{n} + 2\,\mathrm{KL}(\lambda\|u)\right) \;\le\; \tfrac{1}{4}\left(\frac{1}{n} + 2\,\mathrm{KL}(w\|u)\right). \tag{3}$$

Consequently, EPA's λ minimises the KL-based upper bound *in (3) over the feasible set. In particular, EPA favors high effective sample size $N_{\mathrm{eff}}(\lambda)$, although it does not necessarily maximise $N_{\mathrm{eff}}(\lambda)$ exactly.*

Proof sketch. Given S, the λ_i are fixed and label-independent. The summands in Δ_f are independent, mean-zero, and bounded by λ_i, so $\mathrm{Var}(\Delta_f \mid S) = \sum_i \lambda_i^2 \,\mathrm{Var}(\cdot \mid X_i) \le \frac{1}{4}\sum_i \lambda_i^2$. Next, relate dispersion to KL via $\sum_i \lambda_i^2 = \frac{1}{n} + \|\lambda - u\|_2^2 \le \frac{1}{n} + \|\lambda - u\|_1^2 \le \frac{1}{n} + 2\,\mathrm{KL}(\lambda\|u)$ (Pinsker). EPA minimises $\mathrm{KL}(\cdot\|u)$ subject to the same moment constraint, giving the final inequality for any feasible w. □

Proof. **Step 1: centring, independence and a basic variance bound.** For each i, write

$$\ell_i \;:=\; \ell\big(f(X_i), Y_i\big) \in [0,1], \qquad m_i \;:=\; \mathbb{E}[\ell_i \mid X_i] \in [0,1], \qquad Z_i \;:=\; \ell_i - m_i.$$

By construction, $\mathbb{E}[Z_i \mid X_i] = 0$. Since $S = (X_{1:n}, \tilde{X}_{1:m})$ contains $X_{1:n}$ but not the labels, conditioning on S fixes λ and $X_{1:n}$, and the random variation is only through the labels $Y_{1:n}$. Under the i.i.d. assumption, the Y_i (hence the ℓ_i and Z_i) are conditionally independent given S. Moreover, $\ell_i \in [0,1]$ implies

$$Z_i = \ell_i - m_i \in [-m_i,\, 1 - m_i] \subseteq [-1, 1].$$

Therefore, each summand $\lambda_i Z_i$ is conditionally mean-zero and lies in the interval $[-\lambda_i, \lambda_i]$.

Let $\Delta_f = \sum_{i=1}^n \lambda_i Z_i$. Conditional independence and zero means yield

$$\mathrm{Var}(\Delta_f \mid S; \lambda) = \sum_{i=1}^{n} \lambda_i^2 \,\mathrm{Var}(Z_i \mid S) = \sum_{i=1}^{n} \lambda_i^2 \,\mathrm{Var}(\ell_i \mid X_i),$$

where we used that m_i is X_i-measurable and hence constant given S.

Step 2: bounding each $\mathrm{Var}(\ell_i \mid X_i)$ by $1/4$. Since $\ell_i \in [0,1]$, Popoviciu's inequality on variances gives $\mathrm{Var}(\ell_i \mid X_i) \le (1-0)^2/4 = 1/4$.[1] Hence

$$\mathrm{Var}(\Delta_f \mid S; \lambda) \;\le\; \frac{1}{4}\sum_{i=1}^{n} \lambda_i^2,$$

which is the first inequality in (3).

Step 3: relating $\sum_i \lambda_i^2$ to $\mathrm{KL}(\lambda\|u)$. Let u be the uniform weights, $u_i = 1/n$. Compute

$$\sum_{i=1}^{n} \lambda_i^2 = \sum_{i=1}^{n} \Big(u_i + (\lambda_i - u_i)\Big)^2 = \sum_{i=1}^{n} u_i^2 \;+\; 2\sum_{i=1}^{n} u_i(\lambda_i - u_i) \;+\; \sum_{i=1}^{n} (\lambda_i - u_i)^2.$$

Since $\sum_i u_i(\lambda_i - u_i) = \frac{1}{n}\sum_i (\lambda_i - u_i) = 0$, we get

$$\sum_{i=1}^{n} \lambda_i^2 \;=\; \frac{1}{n} \;+\; \|\lambda - u\|_2^2. \tag{4}$$

Next, $\|\lambda - u\|_2 \le \|\lambda - u\|_1$. Pinsker's inequality states

$$\|\lambda - u\|_1^2 \;\le\; 2\,\mathrm{KL}(\lambda\|u),$$

[1] A quick proof: for any random variable $W \in [a,b]$, $\mathrm{Var}(W) = \min_{t\in\mathbb{R}} \mathbb{E}[(W-t)^2] \le \mathbb{E}[(W - \frac{a+b}{2})^2] \le (b-a)^2/4$.

so from (4) we obtain

$$\sum_{i=1}^{n} \lambda_i^2 \;\le\; \frac{1}{n} + 2\,\mathrm{KL}(\lambda\|u),$$

which is the second inequality in (3).

Step 4: effective sample size. Since $N_{\text{eff}}(\lambda) = 1/\sum_i \lambda_i^2$, the first two inequalities show

$$\mathrm{Var}(\Delta_f \mid S;\lambda) \;\le\; \frac{1}{4\,N_{\text{eff}}(\lambda)} \;\le\; \frac{1}{4}\left(\frac{1}{n} + 2\,\mathrm{KL}(\lambda\|u)\right).$$

Minimising $\mathrm{KL}(\cdot\|u)$ therefore minimises this KL-based *upper bound* on variance and favors more dispersed weights. This does not imply exact maximisation of N_{eff}, but it gives a principled surrogate objective controlling concentration of the weights. □

K PROOF OF PROPOSITION 3.2

Proposition K.1 (Discrepancy Gap Bound)**.** *For any hypothesis $h \in \mathcal{H}$, the following inequality holds:*

$$\big|\epsilon_Q(h) - \epsilon_P(h)\big| \le \mathrm{disc}_\ell(P,Q) + \lambda^\ell_{P,Q},$$

where the risks are defined with respect to (potentially different) labeling functions c_P and c_Q.

Proof. Assume ℓ satisfies the triangle inequality and its reverse form, i.e., $\ell(a,c) \le \ell(a,b) + \ell(b,c)$ and $\ell(a,c) \ge \ell(a,b) - \ell(b,c)$ (pointwise). For any distribution D over $\mathcal{X}$, define the pairwise divergence

$$d_{D_X}(g,g') \;:=\; \mathbb{E}_{x\sim D_X}\big[\ell\big(g(x), g'(x)\big)\big],$$

and define the ℓ-discrepancy

$$\mathrm{disc}_\ell(P,Q) \;:=\; \sup_{g,g'\in\mathcal{H}} \big|d_{Q_X}(g,g') - d_{P_X}(g,g')\big|.$$

Let the "joint" hypothesis be

$$h_{\text{joint}} \in \operatorname*{argmin}_{g\in\mathcal{H}} \big(\epsilon_P(g) + \epsilon_Q(g)\big),$$

so the joint error is $\lambda^\ell_{P,Q} = \epsilon_P(h_{\text{joint}}) + \epsilon_Q(h_{\text{joint}})$.

Part 1: Bounding $\epsilon_Q(h) - \epsilon_P(h)$. By the triangle inequality,

$$\begin{aligned}
\epsilon_Q(h) &= \mathbb{E}_{x\sim Q_X}\big[\ell(h(x), c_Q(x))\big] \\
&\le \mathbb{E}_{x\sim Q_X}\big[\ell(h(x), h_{\text{joint}}(x)) + \ell(h_{\text{joint}}(x), c_Q(x))\big] \\
&= d_{Q_X}(h, h_{\text{joint}}) + \epsilon_Q(h_{\text{joint}}).
\end{aligned}$$

By the reverse triangle inequality,

$$\begin{aligned}
\epsilon_P(h) &= \mathbb{E}_{x\sim P_X}\big[\ell(h(x), c_P(x))\big] \\
&\ge \mathbb{E}_{x\sim P_X}\big[\ell(h(x), h_{\text{joint}}(x)) - \ell(h_{\text{joint}}(x), c_P(x))\big] \\
&= d_{P_X}(h, h_{\text{joint}}) - \epsilon_P(h_{\text{joint}}),
\end{aligned}$$

so $-\epsilon_P(h) \le -d_{P_X}(h, h_{\text{joint}}) + \epsilon_P(h_{\text{joint}})$. Combining,

$$\begin{aligned}
\epsilon_Q(h) - \epsilon_P(h) &\le \big(d_{Q_X}(h, h_{\text{joint}}) - d_{P_X}(h, h_{\text{joint}})\big) + \big(\epsilon_P(h_{\text{joint}}) + \epsilon_Q(h_{\text{joint}})\big) \\
&\le \mathrm{disc}_\ell(P,Q) + \lambda^\ell_{P,Q}.
\end{aligned}$$

Part 2: Bounding $\epsilon_P(h) - \epsilon_Q(h)$. Swapping the roles of P and Q in Part 1 gives

$$\epsilon_P(h) - \epsilon_Q(h) \le \mathrm{disc}_\ell(P,Q) + \lambda^\ell_{P,Q}.$$

Conclusion. Combining the two one-sided bounds yields

$$\left|\epsilon_Q(h) - \epsilon_P(h)\right| \;\leq\; \mathrm{disc}_\ell(P, Q) \;+\; \lambda^\ell_{P,Q}. \qquad \square$$

L PROOFS OF THE RESULTS ON ALIGNMENT

L.1 Proof of Proposition 3.6

Theorem L.1 (Alignment Implies Zero Discrepancy)**.** *Let $P^\star_X$ and Q_X such that $\mathbb{E}_{P^\star_X}[\Phi(X)] = \mathbb{E}_{Q_X}[\Phi(X)]$. If Φ is aligned with $(\mathcal{H}, \ell)$, then $\mathrm{disc}_\ell(P^\star_X, Q_X) = 0$. Consequently, $\left|\epsilon_Q(h) - \epsilon_{P^\star}(h)\right| \leq \lambda_{P,Q}$, where $\lambda_{P^\star,Q}$ is small if a model exists that performs well on both domains.*

Proof of Theorem 3.6. The discrepancy is defined as $\sup_{g\in\mathcal{G}_{\mathcal{H},\ell}} |\mathbb{E}_{Q_X}[g(X)] - \mathbb{E}_{P^\star_X}[g(X)]|$. Let g be an arbitrary function in $\mathcal{G}_{\mathcal{H},\ell}$. By the perfect alignment condition, there exist $\beta_0 \in \mathbb{R}$ and $\beta_g \in \mathbb{R}^k$ such that $g(x) = \beta_0 + \beta_g^\top \Phi(x)$. Consider the difference in expectations:

$$\begin{aligned}
\mathbb{E}_{Q_X}[g(X)] - \mathbb{E}_{P^\star_X}[g(X)] &= \mathbb{E}_{Q_X}[\beta_0 + \beta_g^\top \Phi(X)] - \mathbb{E}_{P^\star_X}[\beta_0 + \beta_g^\top \Phi(X)] \\
&= (\beta_0 + \beta_g^\top \mathbb{E}_{Q_X}[\Phi(X)]) - (\beta_0 + \beta_g^\top \mathbb{E}_{P^\star_X}[\Phi(X)]) \\
&= \beta_g^\top \left(\mathbb{E}_{Q_X}[\Phi(X)] - \mathbb{E}_{P^\star_X}[\Phi(X)]\right).
\end{aligned}$$

By the moment-matching assumption, the term in the parenthesis is the zero vector. Thus, the entire expression is zero. Since this holds for any $g \in \mathcal{G}_{\mathcal{H},\ell}$, the supremum over the absolute values is also zero. Therefore, $\mathrm{disc}_\ell(P^\star_X, Q_X) = 0$.

Under the conditions of Theorem 3.6, the domain adaptation bound for a general bounded loss simplifies to:

$$|\epsilon_Q(f) - \epsilon_{P^\star}(f)| \leq \lambda^\ell_{P^\star,Q},$$

where $\lambda^\ell_{P^\star,Q} = \inf_{h\in\mathcal{H}}(\epsilon_{P^\star}(h) + \epsilon_Q(h))$. The distribution shift component of the error bound has been eliminated by the reweighting procedure. $\square$

L.2 Alignment Under Covariate Shift

Proposition L.2. *Let the covariate shift be defined solely by a change in the prevalence of a subgroup $s(X)$. If the conditional risk $r_f(X)$ is approximately constant within the subgroup and its complement, then choosing $\Phi(X) = s(X)$ provides effective control over the error estimation gap.*

Proof. Let us assume for clarity that the conditional risk $r_f(X)$ is piece-wise constant:

$$r_f(X) = \begin{cases} c_1 & \text{if } s(X) = 1 \\ c_0 & \text{if } s(X) = 0 \end{cases}$$

This is a reasonable model if the subgroup is homogeneous with respect to the model's performance. We can rewrite the conditional risk as a linear function of $s(X)$:

$$r_f(X) = c_0(1 - s(X)) + c_1 s(X) = c_0 + (c_1 - c_0)s(X).$$

This shows that the conditional risk function $r_f(X)$ lies in $\mathrm{span}(\{1, s(X)\})$. This is a perfect alignment between the risk function and the feature $s(X)$.

Now, we choose the feature map for EPA to be $\Phi(X) = s(X)$. The moment-matching constraint forces:

$$\mathbb{E}_{P^\star_X}[\Phi(X)] = \mathbb{E}_{Q_X}[\Phi(X)] \implies \mathbb{E}_{P^\star_X}[s(X)] = \pi_Q.$$

The reweighting procedure adjusts the source data to match the target subgroup prevalence exactly. Let's analyze the residual error estimation gap after reweighting:

$$\begin{aligned}
\epsilon_Q(f) - \epsilon_{P^\star}(f) &= \mathbb{E}_{Q_X}[r_f(X)] - \mathbb{E}_{P^\star_X}[r_f(X)] \\
&= \mathbb{E}_{Q_X}[c_0 + (c_1 - c_0)s(X)] - \mathbb{E}_{P^\star_X}[c_0 + (c_1 - c_0)s(X)] \\
&= [c_0 + (c_1 - c_0)\mathbb{E}_{Q_X}[s(X)]] - [c_0 + (c_1 - c_0)\mathbb{E}_{P^\star_X}[s(X)]] \\
&= (c_1 - c_0)(\pi_Q - \pi_Q) = 0.
\end{aligned}$$

The error estimation gap is completely eliminated. Even if the conditional risk is not perfectly piece-wise constant, if it is well-approximated by a linear function of $s(X)$, choosing $\Phi(X) = s(X)$ will significantly reduce the transfer gap. □

L.3 Alignment of prediction histogram

Setup. Let $p_f(X) \in [0,1]$ and partition $[0,1]$ into bins $B_1, \ldots, B_K$. Define model-derived features $\psi_j(x) := \mathbb{1}\{p_f(x) \in B_j\}$ and set $\Phi(x) = (\psi_1(x), \ldots, \psi_K(x))$. Let $\alpha_{D,j} := \Pr_D(p_f(X) \in B_j)$ so that $\sum_j \alpha_{D,j} = 1$. Suppose the conditional risk is well-approximated as piecewise-constant on the bins: $r_j \approx \mathbb{E}[\ell(f(X), Y) \mid p_f(X) \in B_j]$. Then

$$\epsilon_D(f) \approx \sum_{j=1}^{K} \alpha_{D,j}\, r_j.$$

EPA with prediction-histogram features. EPA enforces $\mathbb{E}_{P^\star}[\psi_j] = \mathbb{E}_Q[\psi_j]$ for all j, i.e. it matches the entire prediction histogram: $\alpha_{P^\star,j} = \alpha_{Q,j}$. Hence $\epsilon_{P^\star}(f) \approx \epsilon_Q(f)$, and *exactly* equals it if the piecewise-constant assumption holds.

M PROOF OF THEOREM 3.10 - ALIGNMENT FOR DECISION TREES

M.1 The Hypothesis Class of Decision Trees

Let the input space be $\mathcal{X} \subseteq \mathbb{R}^d$. A decision tree classifier partitions the input space into a set of disjoint regions and assigns a class label to each region.

Definition M.1 (Decision Tree)**.** A decision tree $f : \mathcal{X} \to \mathcal{Y}$ induces a partition of the input space $\mathcal{X}$ into a set of disjoint hyper-rectangular regions $\mathcal{P}_f = \{R_1, R_2, \ldots, R_m\}$, where $\mathcal{X} = \bigcup_{i=1}^m R_i$. Each region R_i is called a **leaf region**. The function f is piecewise constant, taking a single value $y_i \in \mathcal{Y}$ for all $x \in R_i$.

Let $\mathcal{H}$ be the class of all decision trees on $\mathcal{X}$ up to a maximum depth D. For our analysis, we consider the binary classification setting where $\mathcal{Y} = \{0,1\}$ and the loss function is the 0–1 loss, $\ell(y_1, y_2) = \mathbf{1}\{y_1 \neq y_2\}$.

N CHARACTERIZING THE DISAGREEMENT SPACE

To construct an aligned feature map, we must first understand the structure of the disagreement functions $g(x) = \mathbf{1}\{f(x) \neq h(x)\}$ where $f, h \in \mathcal{H}$.

Proposition N.1 (Structure of Disagreement Functions)**.** *Let $f, h \in \mathcal{H}$ be two decision trees, inducing partitions $\mathcal{P}_f = \{R_{f,i}\}$ and $\mathcal{P}_h = \{R_{h,j}\}$ respectively. The disagreement function $g(x) = \mathbf{1}\{f(x) \neq h(x)\}$ is a piecewise constant function. The region where $g(x) = 1$ is a union of disjoint hyper-rectangles.*

Proof. Consider the common refinement of the two partitions, defined as $\mathcal{P}_{f,h} = \{R_{f,i} \cap R_{h,j} \mid R_{f,i} \in \mathcal{P}_f, R_{h,j} \in \mathcal{P}_h\}$. This is also a partition of $\mathcal{X}$ into disjoint hyper-rectangular regions.

Let $R' \in \mathcal{P}_{f,h}$ be an arbitrary region from this refined partition. For any $x \in R'$, both $f(x)$ and $h(x)$ are constant. Specifically, if $R' = R_{f,i} \cap R_{h,j}$, then $f(x)$ takes a constant value y_i and $h(x)$ takes a constant value y_j for all $x \in R'$.

Therefore, the disagreement function $g(x) = \mathbf{1}\{f(x) \neq h(x)\}$ is also constant on R'. Its value is either 0 (if $y_i = y_j$) or 1 (if $y_i \neq y_j$).

The set on which the two trees disagree, $A = \{x \in \mathcal{X} \mid f(x) \neq h(x)\}$, is the union of all regions $R' \in \mathcal{P}_{f,h}$ where the predictions of f and h differ.

$$A = \bigcup_{R' \in \mathcal{P}_{f,h} \text{ s.t. } f(x) \neq h(x) \text{ for } x \in R'} R'.$$

Since each R' is a hyper-rectangle, the disagreement set A is a union of disjoint hyper-rectangles, and its indicator function $g(x) = \mathbf{1}\{x \in A\}$ is the disagreement function. □

N.1 Constructing the Aligned Feature Map

The previous proposition shows that any disagreement function is an indicator function of a set composed of elementary hyper-rectangles. This suggests that the basis for our feature map should be the indicator functions of these elementary regions.

Definition N.2 (Set of All Leaf Regions)**.** Let $\mathcal{R}_{\mathcal{H}}$ be the set of all possible leaf regions that can be generated by any decision tree $f \in \mathcal{H}$.

$$\mathcal{R}_{\mathcal{H}} = \{R \mid \exists f \in \mathcal{H}, R \text{ is a leaf region of } f\}.$$

Definition N.3 (Aligned Feature Map for Decision Trees)**.** Let the feature map $\Phi : \mathcal{X} \to \mathbb{R}^{|\mathcal{R}_{\mathcal{H}}|}$ be defined as the vector of indicator functions for every region in $\mathcal{R}_{\mathcal{H}}$:

$$\Phi(x) = (\mathbf{1}\{x \in R\})_{R \in \mathcal{R}_{\mathcal{H}}}.$$

N.2 Main Result and Proof of Alignment

We now prove that this construction of Φ provides perfect alignment.

Theorem N.4 (Perfect Alignment for Decision Trees)**.** *Let $\mathcal{H}$ be the class of decision trees and ℓ be the 0–1 loss. The feature map $\Phi(x) = (\mathbf{1}\{x \in R\})_{R \in \mathcal{R}_{\mathcal{H}}}$ is perfectly aligned with $(\mathcal{H}, \ell)$.*

Proof. To prove perfect alignment, we must show that any function $g \in \mathcal{G}_{\mathcal{H},\ell}$ can be written as a linear combination of the components of $\Phi(x)$ (and a constant, though it is not needed here).

Let g be an arbitrary function in $\mathcal{G}_{\mathcal{H},\ell}$. By definition, there exist two decision trees $f, h \in \mathcal{H}$ such that $g(x) = \mathbf{1}\{f(x) \neq h(x)\}$. Let $A = \{x \mid f(x) \neq h(x)\}$ be the disagreement set. Then $g(x) = \mathbf{1}\{x \in A\}$.

From Proposition 1, the set A can be written as a finite union of disjoint hyper-rectangular regions from the refined partition $\mathcal{P}_{f,h}$. Let these regions be $\{R'_1, R'_2, \ldots, R'_m\}$. So, $A = \bigcup_{k=1}^{m} R'_k$.

Each region R'_k is formed by an intersection of a leaf region from f and a leaf region from h. Since leaf regions are defined by conjunctions of axis-aligned half-spaces, their intersection is also a region defined by such a conjunction. This means that each R'_k is a valid leaf region for some decision tree in $\mathcal{H}$ (assuming $\mathcal{H}$ is sufficiently rich, e.g., closed under refinement). Therefore, each R'_k is an element of the set of all possible leaf regions, $\mathcal{R}_{\mathcal{H}}$.

Since the regions R'_k are disjoint, the indicator function of their union is the sum of their individual indicator functions:

$$g(x) = \mathbf{1}\{x \in A\} = \mathbf{1}\left\{x \in \bigcup_{k=1}^{m} R'_k\right\} = \sum_{k=1}^{m} \mathbf{1}\{x \in R'_k\}.$$

Each term $\mathbf{1}\{x \in R'_k\}$ in the sum is, by definition, a component of the feature vector $\Phi(x)$, since $R'_k \in \mathcal{R}_{\mathcal{H}}$. The expression above is therefore a linear combination of the components of $\Phi(x)$, where the coefficients are all 1.

Since we have shown that an arbitrary $g \in \mathcal{G}_{\mathcal{H},\ell}$ is in the linear span of the components of Φ, we conclude that $\text{span}(\mathcal{G}_{\mathcal{H},\ell}) \subseteq \text{span}(\{\phi_R\}_{R \in \mathcal{R}_{\mathcal{H}}})$. This satisfies the condition for perfect alignment. □

Corollary N.5. *Let $P^\star_X$ and Q_X be two marginal distributions on $\mathcal{X}$. If the moment-matching condition $\mathbb{E}_{P^\star_X}[\Phi(X)] = \mathbb{E}_{Q_X}[\Phi(X)]$ holds for the feature map Φ defined above, then the discrepancy is zero:*

$$\text{disc}_\ell(P^\star_X, Q_X) = 0.$$

Proof. The moment matching condition $\mathbb{E}_{P^\star_X}[\Phi(X)] = \mathbb{E}_{Q_X}[\Phi(X)]$ implies that for every region $R \in \mathcal{R}_{\mathcal{H}}$, we have $\mathbb{P}_{P^\star_X}(X \in R) = \mathbb{P}_{Q_X}(X \in R)$. For any $g \in \mathcal{G}_{\mathcal{H},\ell}$, we have $g(x) = \sum_{k=1}^{m} \mathbf{1}\{x \in R'_k\}$ for some $R'_k \in \mathcal{R}_{\mathcal{H}}$. Then, by linearity of expectation:

$$\begin{aligned}
\mathbb{E}_{Q_X}[g(X)] - \mathbb{E}_{P^\star_X}[g(X)] &= \mathbb{E}_{Q_X}\left[\sum_{k=1}^{m} \mathbf{1}\{X \in R'_k\}\right] - \mathbb{E}_{P^\star_X}\left[\sum_{k=1}^{m} \mathbf{1}\{X \in R'_k\}\right] \\
&= \sum_{k=1}^{m} \left(\mathbb{P}_{Q_X}(X \in R'_k) - \mathbb{P}_{P^\star_X}(X \in R'_k)\right) \\
&= \sum_{k=1}^{m} (0) = 0.
\end{aligned}$$

Since the difference in expectations is zero for all $g \in \mathcal{G}_{\mathcal{H},\ell}$, the supremum over them is also zero. □

O APPROXIMATE ALIGNMENT AND THE RESIDUAL

Let $\mathcal{A}_\Phi = \{x \mapsto \beta_0 + \beta^\top \Phi(x) : \beta_0 \in \mathbb{R},\ \beta \in \mathbb{R}^k\}$ be the affine hull of Φ. For $g \in \mathcal{G}_{\mathcal{H},\ell}$, define the *residual* relative to (β_0, β): $r_g(x) = g(x) - \beta_0 - \beta^\top \Phi(x)$.

Theorem O.1 (Approximate alignment bounds)**.** *If $\mathbb{E}_{P_X^\star}[\Phi] = \mathbb{E}_{Q_X}[\Phi]$, then*

$$\mathrm{disc}_\ell(P_X^\star, Q_X) \ \le\ \sup_{g \in \mathcal{G}_{\mathcal{H},\ell}} \ \inf_{\beta_0,\beta} \ \Big(\mathbb{E}_{Q_X}|r_g| + \mathbb{E}_{P_X^\star}|r_g|\Big).$$

In particular, if for each g there exist (β_0, β) with $\|r_g\|_\infty \le \eta$, then $\mathrm{disc}_\ell \le 2\eta$.

Proof. Fix any $g \in \mathcal{G}_{\mathcal{H},\ell}$ and any pair (β_0, β). Since

$$g(x) = \beta_0 + \beta^\top \Phi(x) + r_g(x),$$

we have

$$\mathbb{E}_{Q_X}[g(X)] - \mathbb{E}_{P_X^\star}[g(X)] = \beta^\top \big(\mathbb{E}_{Q_X}[\Phi] - \mathbb{E}_{P_X^\star}[\Phi]\big) + \mathbb{E}_{Q_X}[r_g] - \mathbb{E}_{P_X^\star}[r_g].$$

The exact moment-matching assumption cancels the affine term, so

$$\big|\mathbb{E}_{Q_X}[g(X)] - \mathbb{E}_{P_X^\star}[g(X)]\big| \le \mathbb{E}_{Q_X}|r_g| + \mathbb{E}_{P_X^\star}|r_g|.$$

Taking the infimum over (β_0, β) and then the supremum over g proves the first claim. If $\|r_g\|_\infty \le \eta$ uniformly, then each expectation is at most η, yielding $\mathrm{disc}_\ell \le 2\eta$. □

When moments are matched only approximately, let $\Delta = \mathbb{E}_{Q_X}[\Phi] - \mathbb{E}_{P_X^\star}[\Phi]$ and fix any norm pair $(\|\cdot\|, \|\cdot\|_*)$ on $\mathbb{R}^k$.

Theorem O.2 (Approximate moments & approximate alignment)**.** *For any norm pair $(\|\cdot\|, \|\cdot\|_*)$,*

$$\mathrm{disc}_\ell(P_X^\star, Q_X) \ \le\ \sup_{g \in \mathcal{G}_{\mathcal{H},\ell}} \ \inf_{\beta_0,\beta} \ \Big\{ \ \|\beta\|\,\|\Delta\|_* \ +\ \mathbb{E}_{Q_X}|r_g| + \mathbb{E}_{P_X^\star}|r_g| \ \Big\}.$$

If there exist witnesses with $\|\beta\| \le B$ and $\|r_g\|_\infty \le \eta$ uniformly over g, then $\mathrm{disc}_\ell \le B\|\Delta\|_* + 2\eta$.

O.1 How to Control the Residual: Examples

Example O.3 (Threshold classifiers with 0–1 loss (binning gives $O(\delta)$ control).)**.** Let $\mathcal{X} = [0,1]$, $\mathcal{H} = \{h_t(x) = \mathbf{1}\{x > t\} : t \in [0,1]\}$, and $\ell = \mathbf{1}\{\cdot \ne \cdot\}$. For $f = h_{t_1}$, $h = h_{t_2}$ (w.l.o.g. $t_1 < t_2$), the disagreement is $g(x) = \mathbf{1}\{x \in (t_1, t_2]\}$. Choose a grid $0 = \tau_0 < \tau_1 < \cdots < \tau_k = 1$ with mesh $\delta = \max_j(\tau_j - \tau_{j-1})$, and take histogram features

$$\Phi(x) = \big(\psi_1(x), \ldots, \psi_k(x)\big), \qquad \psi_j(x) = \mathbf{1}\{x \in (\tau_{j-1}, \tau_j]\}.$$

Let j_L be the index with $\tau_{j_L} \le t_1 < \tau_{j_L+1}$ and j_R with $\tau_{j_R} \le t_2 < \tau_{j_R+1}$. Approximate g by $\tilde{g}(x) = \sum_{j=j_L+1}^{j_R} \psi_j(x) = \beta^\top \Phi(x)$ ($\beta_0 = 0$, $\beta_j = \mathbf{1}\{j \in \{j_L+1, \ldots, j_R\}\}$). Then $r_g(x) = g(x) - \tilde{g}(x)$ can be nonzero only in the two boundary bins $(\tau_{j_L}, \tau_{j_L+1}]$ and $(\tau_{j_R}, \tau_{j_R+1}]$. If Q_X and $P_X^\star$ admit densities bounded by M on $[0,1]$,

$$\mathbb{E}_{Q_X}|r_g| + \mathbb{E}_{P_X^\star}|r_g| \ \le\ 2M\delta \ +\ 2M\delta \ =\ 4M\delta.$$

Hence Theorem O.1 yields $\mathrm{disc}_\ell(P_X^\star, Q_X) \le 4M\delta$.

Takeaway: Coarse bin features turn interval-type disagreements into affine functions of Φ, with residuals controlled by grid width.

P PROOF OF PROPOSITION 4.1

Proof. Write $\bar{S} := \{1, \ldots, d\} \setminus S$ and assume overlap so the ratios below are well-defined. For $S \subseteq \{1, \ldots, d\}$ define P'_S by

$$\frac{dP'_S}{dP}(x, y) = \frac{q(x_S, y)}{p(x_S, y)}.$$

Then

$$p'_S(x, y) = \frac{q(x_S, y)}{p(x_S, y)}\, p(x, y) = q(x_S, y)\, p(x_{\bar{S}} \mid x_S, y), \quad q(x, y) = q(x_S, y)\, q(x_{\bar{S}} \mid x_S, y).$$

Hence

$$\mathrm{KL}(P'_S \| Q) = 0 \iff P'_S = Q \text{ (a.s.)} \iff p(x_{\bar{S}} \mid x_S, y) = q(x_{\bar{S}} \mid x_S, y) \text{ (a.s.)}. \tag{5}$$

By the sparse joint shift assumption for the unique minimal generating set S^*, $p(x_{\bar{S}^*} \mid x_{S^*}, y) = q(x_{\bar{S}^*} \mid x_{S^*}, y)$, so $\mathrm{KL}(P'_{S^*} \| Q) = 0$.

If $S \supseteq S^*$, then from $p(x_{\bar{S}}, x_{S \setminus S^*} \mid x_{S^*}, y) = q(x_{\bar{S}}, x_{S \setminus S^*} \mid x_{S^*}, y)$, divide by the common marginal $p(x_{S \setminus S^*} \mid x_{S^*}, y) = q(\cdot)$ to get $p(x_{\bar{S}} \mid x_S, y) = q(x_{\bar{S}} \mid x_S, y)$; by (5), $\mathrm{KL}(P'_S \| Q) = 0$.

Conversely, if $\mathrm{KL}(P'_S \| Q) = 0$, then (5) shows S is a generating set. By the defining minimality assumption, any generating S must contain S^*; thus among all S with $\mathrm{KL}(P'_S \| Q) = 0$, the unique minimal-cardinality one is S^*.

Therefore,

$$S^* = \arg \min_{S \subseteq \{1, \ldots, d\}} \big\{ |S| \ : \ \mathrm{KL}(P'_S \| Q) = 0 \big\}.$$

The covariate-shift case is identical with y dropped everywhere. □

Q PROOF OF THEOREM 5.1

Recall the weight construction:

Definition Q.1 (EPA Weights)**.** Given a feature map $\Phi : \mathcal{X} \to \mathbb{R}^k$ and target moments $\hat{\mu}_Q$, the EPA weights are:

$$\lambda_i = \frac{\exp(\langle \beta^\star, \Phi(X_i) \rangle)}{\sum_{j=1}^n \exp(\langle \xi^\star, \Phi(X_j) \rangle)}, \quad i = 1, \ldots, n$$

where $\xi^\star \in \mathbb{R}^k$ minimizes the strictly convex objective:

$$\mathcal{J}(\xi) = \log \left(\frac{1}{n} \sum_{i=1}^n \exp(\langle \xi, \Phi(X_i) \rangle) \right) - \langle \xi, \hat{\mu}_Q \rangle$$

These weights define a reweighted source distribution:

$$\hat{P}^\star = \sum_{i=1}^n \lambda_i \delta_{(X_i, Y_i)}$$

The first part of the theorem is proved in the following lemma.

Lemma Q.2 (Transfer of Improvement)**.** *Assume ℓ satisfies the conditions of Proposition 3.2. Then, for any $f, \tilde{f} \in \mathcal{F}$:*

$$\Big| [\epsilon_Q(\tilde{f}) - \epsilon_Q(f)] - [\epsilon_{P^\star}(\tilde{f}) - \epsilon_{P^\star}(f)] \Big| \leq 2[\mathrm{disc}_\ell(P^\star_X, Q_X) + \lambda_{P^\star, Q}]$$

In particular, if $\epsilon_{P^\star}(\tilde{f}) \leq \epsilon_{P^\star}(f) - \varepsilon$ with:

$$\varepsilon > 2[\mathrm{disc}_\ell(P^\star_X, Q_X) + \lambda_{P^\star, Q}]$$

then $\epsilon_Q(\tilde{f}) < \epsilon_Q(f)$.

Proof. It consists of applying proposition 3.2 to both f and $\tilde{f}$, then using the triangle inequality.

$$\begin{aligned} \big(\epsilon_Q(\tilde{f}) - \epsilon_Q(f)\big) - \big(\epsilon_{P^\star}(\tilde{f}) - \epsilon_{P^\star}(f)\big) &= \big(\epsilon_Q(\tilde{f}) - \epsilon_{P^\star}(\tilde{f})\big) - \big(\epsilon_Q(f) - \epsilon_{P^\star}(f)\big) \\ &\leq |\epsilon_Q(\tilde{f}) - \epsilon_{P^\star}(\tilde{f})| + |\epsilon_Q(f) - \epsilon_{P^\star}(f)| \\ &\leq 2\big(\mathrm{disc}_\ell(P^\star_X, Q_X) + \lambda_{P^\star, Q}\big), \end{aligned}$$

and the reverse inequality is analogous. □

Now, we control the population discrepancy by combining exact empirical moment matching with the approximate-alignment bound from §O.

Lemma Q.3 (Discrepancy Control under Empirical Moment Matching)**.** *Let $\hat{P}^*$ be the reweighted source distribution produced by EPA 2.1 using n source samples and m target samples to compute the empirical target moment $\hat{\mu}_Q = \frac{1}{m}\sum_{j=1}^{m}\Phi(X_j^t)$. Let Q be the true target distribution with population moment $\mu_Q = \mathbb{E}_{Q_X}[\Phi(X)]$.*

Assume the practical feature map Φ satisfies the uniform witness condition of Theorem O.2: for every $g \in \mathcal{G}_{\mathcal{H},\ell}$ there exists a representation $g(x) = \beta_{0,g} + \beta_g^\top \Phi(x) + r_g(x)$ such that $\|\beta_g\|_* \le B_\Phi$ and $\|r_g\|_\infty \le \eta_\Phi$. Then, with $\Delta = \mu_Q - \hat{\mu}_Q$,

$$\mathrm{disc}_\ell(\hat{P}_X^*, Q_X) \le B_\Phi \|\Delta\| + 2\eta_\Phi. \tag{6}$$

Furthermore, if the features are bounded ($\|\Phi(x)\|_2 \le R$), then with probability at least $1-\delta$,

$$\|\Delta\|_2 \le C_{m,k,\delta} := R\sqrt{\frac{2k\log(2/\delta)}{m}}, \tag{7}$$

and therefore

$$\mathrm{disc}_\ell(\hat{P}_X^*, Q_X) \le B_\Phi C_{m,k,\delta} + 2\eta_\Phi. \tag{8}$$

Proof. By construction of EPA, $\mathbb{E}_{\hat{P}_X^\star}[\Phi] = \hat{\mu}_Q$, hence

$$\Delta = \mathbb{E}_{Q_X}[\Phi] - \mathbb{E}_{\hat{P}_X^\star}[\Phi].$$

Applying Theorem O.2 with the stated uniform witnesses yields $\mathrm{disc}_\ell(\hat{P}_X^*, Q_X) \le B_\Phi\|\Delta\| + 2\eta_\Phi$. The concentration bound on $\|\Delta\|_2$ follows from a vector Hoeffding inequality for the bounded vectors $\Phi(X_j^t)$. □

Remark Q.4. For exact alignment, one can set $\eta_\Phi = 0$ and recover the sharper discrepancy bound proportional only to the target-moment sampling error. The practical histogram map used in EPA is better modelled through the approximate-alignment constants (B_Φ, η_Φ).

Theorem Q.5. *Assume ℓ satisfies the conditions of Proposition 3.2. Let $\tilde{f}$ be the boosted model and define the empirical performance drop as $\Delta_{\mathrm{emp}} = \epsilon_{\hat{P}^\star}(f) - \epsilon_{\hat{P}^\star}(\tilde{f})$. If*

$$\Delta_{\mathrm{emp}} > 2\big(\mathrm{disc}_{\mathcal{L}}(\hat{P}_X^\star, \hat{Q}_X) + \lambda_{\hat{P}^\star,\hat{Q}}\big),$$

then $\epsilon_{\hat{Q}}(\tilde{f}) \le \epsilon_{\hat{Q}}(f)$. Furthermore, under the assumptions of Lemma Q.3, if

$$\Delta_{\mathrm{emp}} > B_{m,k,\delta}, \quad \textit{with} \quad B_{m,k,\delta} = 2\big(B_\Phi C_{m,k,\delta} + 2\eta_\Phi + \lambda_{\hat{P}^\star,Q}\big)$$

then with probability at least $1-\delta$, we have $\epsilon_Q(\tilde{f}) < \epsilon_Q(f)$.

Proof. The empirical statement follows directly from Lemma Q.2 applied to the pair $(\hat{P}^\star, \hat{Q})$. For the population statement, apply Lemma Q.2 to $(\hat{P}^\star, Q)$ and substitute the discrepancy bound from Lemma Q.3. On the high-probability event $\mathrm{disc}_\ell(\hat{P}_X^\star, Q_X) \le B_\Phi C_{m,k,\delta} + 2\eta_\Phi$, the transfer lemma yields

$$\epsilon_Q(\tilde{f}) - \epsilon_Q(f) \le -\Delta_{\mathrm{emp}} + 2\big(B_\Phi C_{m,k,\delta} + 2\eta_\Phi + \lambda_{\hat{P}^\star,Q}\big),$$

which is negative under the displayed condition. □

R PROOF ANALYSING THE EMPIRICAL VERSION OF EPA WITH ITS POPULATION VERSION.

Note that this section complements Bachoc et al. (2023), which provides an asymptotic analysis of the entropic variable projection, by instead offering a non-asymptotic analysis of the solution.

R.1 Problem Formulation

We recall the population and empirical problems.

- **Population Problem:** Given a source distribution P_X and target moment $\mu_Q = \mathbb{E}_{Q_X}[\Phi(X)]$, find $P^* = \operatorname{argmin}_{P \ll P_X} \mathrm{KL}(P \| P_X)$ subject to $\mathbb{E}_P[\Phi(X)] = \mu_Q$. The solution is given by the density $\frac{dP^*}{dP_X}(x) = \exp(\langle \xi^*, \Phi(x)\rangle - H(\xi^*))$, where ξ^* is the true dual parameter and $H(\xi)$ is the log-partition function.
- **Empirical Problem:** Given source samples $\{X_i\}_{i=1}^n \sim P_X$ and target samples $\{Z_j\}_{j=1}^m \sim Q_X$, we compute the empirical measure $\hat{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ and estimate the target moment as $\hat{\mu}_{Q,m} = \frac{1}{m}\sum_{j=1}^m \Phi(Z_j)$. We then find $\hat{P}^*_{n,w} = \operatorname{argmin}_{\tilde{P} \ll \hat{P}_n} \mathrm{KL}(\tilde{P} \| \hat{P}_n)$ subject to $\mathbb{E}_{\tilde{P}}[\Phi(X)] = \hat{\mu}_{Q,m}$. The solution is a discrete distribution on $\{X_i\}_{i=1}^n$ with weights
$$\lambda_i^* = \frac{\exp\{\langle \hat{\xi}^*_{n,m}, \Phi(X_i)\rangle\}}{\sum_{\ell=1}^n \exp\{\langle \hat{\xi}^*_{n,m}, \Phi(X_\ell)\rangle\}},$$
where $\hat{\xi}^*_{n,m}$ is the empirical dual parameter.

For simplicity, we will denote $\hat{\xi}^*_{n,m}$ as $\hat{\xi}^*$.

Remark R.1 (The Object of Analysis). A crucial distinction must be made. The direct solution to the empirical problem is the discrete weighted measure $\hat{P}^*_{n,w}$. However, comparing a continuous distribution like P^* to a discrete one via KL divergence or TV distance is uninformative (indeed often infinite). The theoretical bounds in Theorem R.6 apply not to the discrete measure $\hat{P}^*_{n,w}$, but to the **continuous distribution from the same exponential family parameterized by the estimated parameter** $\hat{\xi}^*$. We denote this distribution by $P_{\hat{\xi}^*}$. Its density is:
$$\frac{dP_{\hat{\xi}^*}}{dP_X}(x) = \exp(\langle \hat{\xi}^*, \Phi(x)\rangle - H(\hat{\xi}^*))$$
Since P^* and $P_{\hat{\xi}^*}$ are both continuous and defined with respect to P_X, their KL divergence and TV distance are well-defined. In Section U, we will show how to connect the bounds on $P_{\hat{\xi}^*}$ back to the practical discrete solution $\hat{P}^*_{n,w}$. For notational consistency, we will use $\hat{P}^*_n$ to refer to this continuous distribution $P_{\hat{\xi}^*}$ in Sections 1–3.

R.2 Assumptions

Assumption R.2 (Bounded Features)**.** The feature map $\Phi : E \to \mathbb{R}^k$ is uniformly bounded, i.e., there exists a constant $R > 0$ such that $\|\Phi(x)\|_2 \le R$ for all $x \in E$.

Assumption R.3 (Strong Convexity and Smoothness)**.** The population log-partition function $H(\xi) = \log \mathbb{E}_{P_X}[e^{\langle \xi, \Phi(X)\rangle}]$ is σ-strongly convex and L-smooth in a neighborhood of ξ^*. Equivalently, its Hessian is bounded: $\sigma I \preceq \nabla^2 H(\xi) \preceq LI$ for some $L \ge \sigma > 0$.

Assumption R.4 (Feasibility and Compact Parameter Set)**.** There exists a nonempty compact convex set $\Xi \subset \mathbb{R}^k$ such that: (i) $\xi^* \in \Xi$ and H is σ-strongly convex and L-smooth on Ξ; (ii) the empirical dual solution $\hat{\xi}^*$ exists, is unique, and lies in Ξ by assuming that $\hat{\mu}_{Q,m}$ lies in the relative interior of $\operatorname{conv}\{\Phi(X_i)\}_{i=1}^n$).

Assumption R.5 (Uniform gradient concentration)**.** There exists a deterministic function $\alpha_n : (0,1) \to (0,\infty)$ such that, for every $\delta \in (0,1)$,
$$\mathbb{P}\left(\sup_{\xi \in \Xi} \|\nabla H(\xi) - \nabla H_{\hat{P}_n}(\xi)\|_2 \le \alpha_n(\delta)\right) \ge 1 - \delta,$$
where
$$H_{\hat{P}_n}(\xi) := \log\left(\frac{1}{n}\sum_{i=1}^n e^{\langle \xi, \Phi(X_i)\rangle}\right).$$

R.3 Main Theorem

The following theorem provides finite-sample bounds for the estimated continuous model.

Theorem R.6 (Finite-Sample Bounds for Empirical I-Projection)**.** *Let Assumptions R.2, R.3, and R.4 hold. Let $\hat{P}_n^*$ be the continuous distribution $P_{\hat{\xi}^*}$ defined in Remark R.1. Then for any confidence level $\delta \in (0,1)$, with probability at least $1-2\delta$, the following bounds hold for the empirical solution based on n source and m target samples:*

1. ***Dual Parameter Error:***
$$\|\hat{\xi}^* - \xi^*\|_2 \leq \frac{1}{\sigma}\left(\alpha_n(\delta) + \beta_m(\delta)\right).$$

2. ***KL Divergence Error:***
$$\mathrm{KL}(P^* \| \hat{P}_n^*) \leq \frac{L}{2}\|\hat{\xi}^* - \xi^*\|_2^2.$$

3. ***Total Variation Error:***
$$d_{TV}(P^*, \hat{P}_n^*) \leq \sqrt{\frac{L}{4}\|\hat{\xi}^* - \xi^*\|_2^2}.$$

where
$$\beta_m(\delta) := R\sqrt{\frac{2k\log(2k/\delta)}{m}}.$$

The remainder of this document is dedicated to proving this theorem and connecting it to the discrete empirical solution.

S PROOF OF THE BOUND ON THE DUAL PARAMETER

The proof relies on establishing a basic inequality that links the output error $\|\hat{\xi}^* - \xi^*\|$ to input errors, which are then bounded using concentration inequalities.

S.1 Step 1: A Basic Inequality

Let $H(\xi) = \log \mathbb{E}_{P_X}[e^{\langle \xi, \Phi(X)\rangle}]$ and $H_{\hat{P}_n}(\xi) = \log\left(\frac{1}{n}\sum_{i=1}^n e^{\langle \xi, \Phi(X_i)\rangle}\right)$. Set $G = \nabla H$ and $\hat{G}_n = \nabla H_{\hat{P}_n}$. The optimality conditions are $G(\xi^*) = \mu_Q$ and $\hat{G}_n(\hat{\xi}^*) = \hat{\mu}_{Q,m}$.

By strong convexity of H on Ξ (Assumption R.4), the gradient map G is σ-strongly monotone: $\langle G(y) - G(x), y - x\rangle \geq \sigma\|y-x\|_2^2$. Hence, by Cauchy–Schwarz, $\|G(y) - G(x)\|_2 \geq \sigma\|y-x\|_2$. Thus

$$\sigma\|\hat{\xi}^* - \xi^*\|_2 \leq \|G(\hat{\xi}^*) - G(\xi^*)\|_2.$$

Substituting $G(\xi^*) = \mu_Q$ and using the empirical optimality condition,

$$\begin{aligned}
\sigma\|\hat{\xi}^* - \xi^*\|_2 &\leq \|G(\hat{\xi}^*) - \mu_Q\|_2 \\
&= \|G(\hat{\xi}^*) - \hat{G}_n(\hat{\xi}^*) + \hat{\mu}_{Q,m} - \mu_Q\|_2 \\
&\leq \|G(\hat{\xi}^*) - \hat{G}_n(\hat{\xi}^*)\|_2 + \|\hat{\mu}_{Q,m} - \mu_Q\|_2.
\end{aligned}$$

This is our basic inequality.

S.2 Step 2: Bounding the Error Terms via Concentration

Lemma S.1 (Bound on Target Moment Error)**.** *Under Assumption R.2, for any $\delta \in (0,1)$, with probability at least $1-\delta$:*

$$\|\hat{\mu}_{Q,m} - \mu_Q\|_2 \leq R\sqrt{\frac{2k\log(2k/\delta)}{m}} = \beta_m(\delta).$$

Proof. Apply a vector Hoeffding inequality to the i.i.d. bounded vectors $\{\Phi(Z_j)\}_{j=1}^m$. □

S.3 Step 3: Combining the Bounds

By a union bound applied to Lemmas S.1 and Assumptions R.5, with probability at least $1 - 2\delta$,

$$\sigma\|\hat{\xi}^* - \xi^*\|_2 \leq \alpha_n(\delta) + \beta_m(\delta),$$

which yields Item (1) of Theorem R.6 after dividing by σ.

T PROOF OF THE BOUNDS ON THE DISTRIBUTION ERROR

We now translate the bound on the dual parameter error into bounds on statistical distances between the continuous distributions P^* and $\hat{P}_n^* = P_{\hat{\xi}^*}$.

T.1 Step 1: Bounding the KL Divergence

The KL divergence between two distributions from an exponential family is equal to the Bregman divergence of their natural parameters with respect to the log-partition function.

Proposition T.1. *For the optimal distributions P^* and $\hat{P}_n^*$, we have:*

$$\mathrm{KL}(P^*\|\hat{P}_n^*) = H(\hat{\xi}^*) - H(\xi^*) - \langle \nabla H(\xi^*), \hat{\xi}^* - \xi^* \rangle =: D_H(\hat{\xi}^*\|\xi^*).$$

Proof. This is the standard identity linking exponential-family KL to the Bregman divergence of H. □

Proposition T.2 (Smoothness upper bound for D_H)**.** *If H is L-smooth on a convex set containing ξ^* and $\hat{\xi}^*$, then*

$$D_H(\hat{\xi}^*\|\xi^*) \leq \frac{L}{2}\,\|\hat{\xi}^* - \xi^*\|_2^2.$$

Proof. A standard property of L-smooth functions: $H(y) \leq H(x) + \langle \nabla H(x), y - x\rangle + \frac{L}{2}\|y - x\|^2$ with $x = \xi^*$, $y = \hat{\xi}^*$. □

Combining Propositions T.1 and T.2 proves Item (2) of Theorem R.6.

T.2 Step 2: Bounding the Total Variation Distance

Lemma T.3 (Pinsker's Inequality)**.** *For any two probability measures P and Q, $d_{TV}(P, Q)^2 \leq \frac{1}{2}\,\mathrm{KL}(P\|Q)$.*

Proof. Classical result; see, e.g., standard information-theory references. □

Applying Lemma T.3 together with Item (2) of Theorem R.6, we obtain Item (3):

$$d_{\mathrm{TV}}(P^*, \hat{P}_n^*)^2 \leq \tfrac{1}{2}\mathrm{KL}(P^*\|\hat{P}_n^*) \leq \tfrac{L}{4}\,\|\hat{\xi}^* - \xi^*\|_2^2.$$

U Connection to the Discrete Empirical Solution

We now relate the continuous tilted model $P_{\hat{\xi}^*}$ to the practical discrete optimizer $\hat{P}_{n,w}^*$. Since these measures need not be mutually absolutely continuous, Wasserstein distance is the appropriate metric.

Assumption U.1 (Metric regularity for the discrete comparison)**.** The space (E, d) is a metric space with finite diameter

$$\mathrm{diam}(E) \leq D < \infty.$$

Moreover, Φ is L_Φ-Lipschitz:

$$\|\Phi(x) - \Phi(y)\|_2 \leq L_\Phi d(x, y) \qquad \text{for all } x, y \in E.$$

Proposition U.2 (Stability of exponential reweighting in W_1)**.** *Assume Assumptions R.2-R.4, and U.1. Let*

$$B_\Xi := \sup_{\xi\in\Xi}\|\xi\|_2, \qquad m_\Xi := e^{-B_\Xi R}, \qquad M_\Xi := e^{B_\Xi R}, \qquad L_w := B_\Xi L_\Phi e^{B_\Xi R}.$$

Then

$$W_1(P_{\hat{\xi}^*}, \hat{P}_{n,w}^*) \leq C_{\Xi,\Phi,D}\, W_1(P_X, \hat{P}_n),$$

where

$$C_{\Xi,\Phi,D} := \frac{M_\Xi + DL_w}{m_\Xi} + \frac{DM_\Xi L_w}{m_\Xi^2}.$$

Proof. Set

$$w(x) := e^{\langle \hat{\xi}^*, \Phi(x) \rangle}.$$

Since $\hat{\xi}^* \in \Xi$ and $\|\Phi(x)\|_2 \le R$,

$$m_\Xi \le w(x) \le M_\Xi \qquad \text{for all } x \in E.$$

Also, by the Lipschitz property of Φ,

$$|w(x) - w(y)| \le B_\Xi L_\Phi e^{B_\Xi R}\, d(x,y) = L_w d(x,y).$$

Define the reweighting map

$$T_w(\mu)(dx) := \frac{w(x)}{\mu(w)}\, \mu(dx)$$

for any probability measure μ on E. Then

$$P_{\hat{\xi}^*} = T_w(P_X), \qquad \hat{P}^*_{n,w} = T_w(\hat{P}_n).$$

Fix $x_0 \in E$. By the Kantorovich–Rubinstein dual representation on a bounded metric space,

$$W_1(T_w(\mu), T_w(\nu)) = \sup_{\substack{\mathrm{Lip}(f) \le 1 \\ f(x_0)=0}} \left| \frac{\mu(fw)}{\mu(w)} - \frac{\nu(fw)}{\nu(w)} \right|.$$

For every such f, the diameter bound implies $|f| \le D$, hence

$$\mathrm{Lip}(fw) \le M_\Xi + DL_w, \qquad |\nu(fw)| \le DM_\Xi, \qquad \mu(w), \nu(w) \ge m_\Xi.$$

Therefore,

$$\begin{aligned} \left| \frac{\mu(fw)}{\mu(w)} - \frac{\nu(fw)}{\nu(w)} \right| &\le \frac{|\mu(fw) - \nu(fw)|}{\mu(w)} + |\nu(fw)| \frac{|\mu(w) - \nu(w)|}{\mu(w)\nu(w)} \\ &\le \frac{M_\Xi + DL_w}{m_\Xi}\, W_1(\mu,\nu) + \frac{DM_\Xi L_w}{m_\Xi^2}\, W_1(\mu,\nu), \end{aligned}$$

where we used Kantorovich–Rubinstein duality both for fw and for w. Taking the supremum over all admissible f yields the result. □

Corollary U.3 (Discrete error decomposition)**.** *Assume the conditions of Theorem R.6 and Proposition U.2. Then, on the same event as in Theorem R.6,*

$$W_1(P^*, \hat{P}^*_{n,w}) \le \frac{D\sqrt{L}}{2\sigma} \big(\alpha_n(\delta) + \beta_m(\delta)\big) + C_{\Xi,\Phi,D}\, W_1(P_X, \hat{P}_n).$$

Proof. By the triangle inequality,

$$W_1(P^*, \hat{P}^*_{n,w}) \le W_1(P^*, P_{\hat{\xi}^*}) + W_1(P_{\hat{\xi}^*}, \hat{P}^*_{n,w}).$$

Since $\mathrm{diam}(E) \le D$,

$$W_1(P^*, P_{\hat{\xi}^*}) \le D\, d_{\mathrm{TV}}(P^*, P_{\hat{\xi}^*}).$$

Applying Item (3) of Theorem R.6 yields

$$W_1(P^*, P_{\hat{\xi}^*}) \le \frac{D\sqrt{L}}{2\sigma} \big(\alpha_n(\delta) + \beta_m(\delta)\big).$$

The second term is controlled by Proposition U.2. Combining the two bounds proves the claim. □

V DISCUSSION OF THE APPROXIMATION GAP OF AXIS-ALIGNED BINNING

Definition V.1 (Axis-Aligned Histogram Map)**.** Let the feature space be $\mathcal{X} \subseteq \mathbb{R}^d$. For each feature dimension $j \in \{1, \ldots, d\}$, let $\mathcal{B}_j = \{B_{j,1}, \ldots, B_{j,m_j}\}$ be a partition of its range (e.g., from quantile binning). The **axis-aligned histogram map** is:

$$\Phi_{hist}(x) = \left(\mathbb{1}\{x^{(1)} \in B_{1,1}\}, \ldots, \mathbb{1}\{x^{(1)} \in B_{1,m_1}\}, \ldots, \mathbb{1}\{x^{(d)} \in B_{d,m_d}\}\right)$$

Matching moments with Φ_{hist} enforces that the marginal probability mass in each bin is the same for $P_X^\star$ and Q_X, but it does not constrain the joint distribution of features. This lack of constraint on feature interactions is the source of the gap.

To close the gap, we suggest exploring feature maps that can capture more complex relations, which may be worth investigating in future work.

V.1 Method 1: Feature Crosses

We can explicitly model interactions by binning the joint distribution of feature pairs (or higher-order tuples).

V.2 Method 2: Tree-Based Binning

A more adaptive approach is to use a decision tree to define the bins, as its structure naturally captures interactions relevant to the data.

Definition V.2 (Tree-Based Map)**.** Let $\mathcal{T}$ be a decision tree trained on a task related to the distribution shift (e.g., to classify whether a point x comes from P_X or Q_X). Let its leaf regions be $\mathcal{L} = \{L_1, \ldots, L_m\}$. The **tree-based feature map** is:

$$\Phi_{tree}(x) = (\mathbb{1}\{x \in L_1\}, \ldots, \mathbb{1}\{x \in L_m\})$$

By construction, the regions $\{L_i\}$ are sensitive to the parts of the feature space where P_X and Q_X differ. This makes Φ_{tree} a far better approximation of the disagreement space $\mathcal{G}_{\mathcal{H},\ell}$ than Φ_{hist}.

Another solution is to use a tighter bound of K that is specific to a given hypothesis h for the decision tree in question.

The original discrepancy term can be replaced by one that measures the divergence only between the specific hypothesis h and the ideal joint hypothesis h_{joint}.

Proposition V.3 (Hypothesis-Specific Discrepancy Gap Bound)**.** *For any given hypothesis $h \in \mathcal{H}$, the following inequality holds:*

$$\left|\epsilon_Q(h) - \epsilon_P(h)\right| \leq \left|d_{Q_X}(h, h_{joint}) - d_{P_X}(h, h_{joint})\right| + \lambda^{\ell}_{P,Q}$$

This bound is tighter than the original, as the hypothesis-specific term is necessarily less than or equal to the supremum taken over all pairs in $\mathcal{H}$:

$$\left|d_{Q_X}(h, h_{\text{joint}}) - d_{P_X}(h, h_{\text{joint}})\right| \leq \sup_{g,g' \in \mathcal{H}} \left|d_{Q_X}(g, g') - d_{P_X}(g, g')\right|$$

Using the tighter, hypothesis-specific bound for your decision tree provides a more precise and practical target for your feature mapping.

Instead of needing a feature map that works for all possible trees, you only need one that works for your specific tree (according to its leaves). This is a less demanding and more realistic goal.

Proof. Assume the loss function ℓ satisfies the triangle inequality, $\ell(a, c) \leq \ell(a, b) + \ell(b, c)$, and its reverse form, $\ell(a, c) \geq \ell(a, b) - \ell(c, b)$.

Step 1: Bound the difference from one side. We first seek an upper bound for the term $\epsilon_Q(h) - \epsilon_P(h)$.

- Using the triangle inequality, we can bound $\epsilon_Q(h)$:

$$\begin{aligned}\epsilon_Q(h) &= \mathbb{E}_{x\sim Q_X}[\ell(h(x), c_Q(x))] \\ &\leq \mathbb{E}_{x\sim Q_X}[\ell(h(x), h_{\text{joint}}(x)) + \ell(h_{\text{joint}}(x), c_Q(x))] \\ &= d_{Q_X}(h, h_{\text{joint}}) + \epsilon_Q(h_{\text{joint}})\end{aligned}$$

- Using the reverse triangle inequality, we find a lower bound for $\epsilon_P(h)$:

$$\begin{aligned}\epsilon_P(h) &= \mathbb{E}_{x\sim P_X}[\ell(h(x), c_P(x))] \\ &\geq \mathbb{E}_{x\sim P_X}[\ell(h(x), h_{\text{joint}}(x)) - \ell(c_P(x), h_{\text{joint}}(x))] \\ &= d_{P_X}(h, h_{\text{joint}}) - \epsilon_P(h_{\text{joint}})\end{aligned}$$

Combining these two results yields a one-sided inequality:

$$\begin{aligned}\epsilon_Q(h) - \epsilon_P(h) &\leq \big(d_{Q_X}(h, h_{\text{joint}}) + \epsilon_Q(h_{\text{joint}})\big) - \big(d_{P_X}(h, h_{\text{joint}}) - \epsilon_P(h_{\text{joint}})\big) \\ &= \big(d_{Q_X}(h, h_{\text{joint}}) - d_{P_X}(h, h_{\text{joint}})\big) + \big(\epsilon_P(h_{\text{joint}}) + \epsilon_Q(h_{\text{joint}})\big) \\ &= \big(d_{Q_X}(h, h_{\text{joint}}) - d_{P_X}(h, h_{\text{joint}})\big) + \lambda^{\ell}_{P,Q}\end{aligned}$$

Step 2: Combine with the symmetric argument. By swapping the roles of P and Q in the argument above, we obtain the bound for $\epsilon_P(h) - \epsilon_Q(h)$:

$$\epsilon_P(h) - \epsilon_Q(h) \leq \big(d_{P_X}(h, h_{\text{joint}}) - d_{Q_X}(h, h_{\text{joint}})\big) + \lambda^{\ell}_{P,Q}$$

Together, these two one-sided inequalities imply the final result with the absolute value:

$$\big|\epsilon_Q(h) - \epsilon_P(h)\big| \leq \big|d_{Q_X}(h, h_{\text{joint}}) - d_{P_X}(h, h_{\text{joint}})\big| + \lambda^{\ell}_{P,Q}$$

□

W EPA WHEN THE TARGET MEAN LIES OUTSIDE THE CONVEX HULL

Let $\Phi_n := \mathrm{conv}\{\Phi(X_1), \ldots, \Phi(X_n)\} \subset \mathbb{R}^k$ be the convex hull of source features and let $t \in \mathbb{R}^k$ denote the (target) mean we wish to match. If $t \notin \Phi_n$, the equality-constrained EPA problem is infeasible. A principled remedy is to *project* t onto Φ_n (with respect to a chosen norm) and then impose *equality* to this projection.

Step 1: metric projection onto Φ_n. Fix a norm $\|\cdot\|$ on $\mathbb{R}^k$. Define the projection

$$s^\star \in \operatorname*{argmin}_{s \in \Phi_n} \|s - t\| \qquad \text{and its distance} \qquad \delta_n(t) := \|s^\star - t\|.$$

Equivalently, in the weights $\lambda \in \Delta_n$,

$$s^\star \in \operatorname*{argmin}_{\lambda \in \Delta_n} \Big\| \sum_{i=1}^n \lambda_i \Phi(X_i) - t \Big\|. \tag{9}$$

For $\|\cdot\|_2$ this is a small QP; for ℓ_1/ℓ_∞ it is an LP.

Step 2: EPA at the projected target. Run EPA with the equality constraint set to $s^\star$:

$$\lambda^\star \in \operatorname*{argmin}_{\lambda \in \Delta_n} \mathrm{KL}(\lambda \| u) \quad \text{s.t.} \quad \sum_{i=1}^n \lambda_i \, \Phi(X_i) = s^\star, \qquad u_i = \tfrac{1}{n}. \tag{10}$$

By Theorem J.1, the unique solution has the exponential-tilt form

$$\lambda_i^\star = \frac{\exp\{\langle \xi^\star, \Phi(X_i)\rangle\}}{\sum_{j=1}^n \exp\{\langle \xi^\star, \Phi(X_j)\rangle\}}, \qquad \sum_{i=1}^n \lambda_i^\star \, \Phi(X_i) = s^\star,$$

where $\xi^\star$ is the (unique) minimizer of $H(\xi) - \langle \xi, s^\star \rangle$ with $H(\xi) = \log\big(\frac{1}{n}\sum_j e^{\langle \xi, \Phi(X_j)\rangle}\big)$.

Equivalence to hard proximity EPA. The projection approach is equivalent to solving EPA with a hard proximity tolerance equal to the projection distance:

$$\min_{\lambda \in \Delta_n} \mathrm{KL}(\lambda \| u) \quad \text{s.t.} \quad \Big\| \sum_{i=1}^n \lambda_i \, \Phi(X_i) - t \Big\| \le \delta_n(t). \tag{11}$$

Indeed, let $B := \{s : \|s - t\| \le \delta_n(t)\}$. By definition of $s^\star$, $B \cap \Phi_n = \{s^\star\}$, so the feasible set of (11) is exactly the affine face $\{\lambda \in \Delta_n : \sum_i \lambda_i \Phi(X_i) = s^\star\}$, and (11) reduces to (10). Consequently the unique optimizer is the same $\lambda^\star$.

Dual viewpoint and computation. Let $\|\cdot\|_*$ denote the dual norm. The Lagrange dual of (11) is the strictly convex problem

$$\min_{\xi \in \mathbb{R}^k} \; H(\xi) - \langle \xi, t \rangle - \delta_n(t) \, \|\xi\|_*,$$

whose minimizer $\xi^\star$ yields $\lambda^\star$ via the exponential tilt above. When $t \notin \Phi_n$ the proximity constraint is active and $\|\xi^\star\|_* > 0$. Practically, one may either (i) solve the projection (9) and then the equality EPA (10), or (ii) solve the single smooth/nonsmooth dual above; both give the same $\lambda^\star$.

Remarks. (i) The choice of norm encodes the geometry you trust: Mahalanobis (with target covariance) is often statistically natural; Euclidean is a robust default. (ii) If $t \in \Phi_n$, then $\delta_n(t) = 0$ and we recover the usual EPA. (iii) Geometrically, $s^\star$ lies on a face of Φ_n exposed by some $\xi^\star$; the EPA weights are the maximum-entropy distribution realizing that face-average.