HELLINGERIEP: STABLE DOMAIN ADAPTATION UNDER CONTAMINATED COVARIATE SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Importance weighting is a standard approach to domain adaptation under covariate shift, but direct density-ratio estimation can be unstable when the unlabeled target sample is contaminated. Existing robust methods often rely on user-specified trimming or reweighting mechanisms. We study how the divergence geometry of a density-ratio objective affects contamination sensitivity. We introduce HellingerIEP, a nonnegative kernel estimator derived from a variational Hellinger objective. Parameterizing the square root of the density ratio yields a convex optimization problem, and, for a fixed kernel dictionary and regularization level, we establish consistency for the corresponding regularized population solution and asymptotic normality under an interiority condition. Under a common square-root ratio parameterization, the Hellinger, KL, and Pearson target-score contributions contain factors of orders w^{-1}, $w^{-1/2}$, and $w^{1/2}$, respectively. Thus, Hellinger attenuates added-point score contributions more strongly in high-ratio regions but has greater score magnitude in low-ratio regions, so no uniform sensitivity ordering follows. This nonuniformity suggests a stability–adaptation trade-off, since a large fitted ratio may represent either contamination or genuine target shift. Experiments on synthetic and benchmark domain adaptation tasks show that HellingerIEP is competitive in clean settings and more stable under the contamination regimes considered.

1 INTRODUCTION

Domain adaptation addresses a central problem in statistical learning: how to transfer knowledge across domains when the prediction task remains the same but the data distribution changes (Kamath et al., 2019; Farahani et al., 2021). For example, a model for predicting heart disease risk may be trained using labeled patient records from one hospital and then applied at another hospital whose patient characteristics differ. In the second hospital, the patient covariates may be available, but the outcome labels needed to retrain the model or evaluate its performance may be unavailable or costly to obtain. This combination of distribution shift and missing target-domain labels makes direct transfer difficult and motivates unsupervised domain adaptation methods.

To formulate this, let the source domain be $S = (\mathcal{X}, \mathcal{Y}, P_S)$ and the target domain be $T = (\mathcal{X}, \mathcal{Y}, P_T)$, where P_S and P_T are probability measures on $\mathcal{X} \times \mathcal{Y}$. We assume $\mathcal{X} \subseteq \mathbb{R}^d$, and $\mathcal{Y}$ is either $\mathbb{R}$ for regression or a finite label set for classification. We assume P_S and P_T admit joint densities, denoted by $p_S(x, y)$ and $p_T(x, y)$, respectively. Suppose we observe n_S i.i.d. labeled source samples $\{(X_i^S, Y_i^S)\}_{i=1}^{n_S} \sim P_S$, together with n_T i.i.d. unlabeled target covariates $\{X_j^T\}_{j=1}^{n_T} \sim P_T^X$, where P_T^X denotes the marginal distribution of X under P_T. The goal of domain adaptation is to construct a predictor $h : \mathcal{X} \to \mathcal{Y}$ with small target-domain risk given a loss function ℓ:

$$R_T(h) = \mathbb{E}_{(X,Y)\sim P_T}\big[\ell(h(X), Y)\big].$$

Without additional assumptions, the unsupervised domain adaptation problem seems intractable (Kamath et al., 2019). Because the target labels are unobserved, the target risk cannot be estimated from the available data. A standard assumption is *covariate shift* (Shimodaira, 2000; Sugiyama et al., 2007a; Dharani et al., 2019), where the conditional distribution of the response given the covariates is assumed to be invariant across domains while the marginal covariate distribution is allowed to change, i.e., $p_S(y \mid x) = p_T(y \mid x)$. Assuming the source density is strictly positive, $p_S(x) > 0$ for

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

all $x \in \mathcal{X}$, the target risk can be rewritten as a weighted source risk

$$R_T(h) = \mathbb{E}_{(X,Y)\sim P_S}\big[r(X)\,\ell(h(X),Y)\big], \text{ where } r(x) = \frac{p_T(x)}{p_S(x)}.$$

The function $r(x)$ is the density ratio, or importance weighting function (Nguyen et al., 2010; Huang et al., 2006; Sugiyama et al., 2007b). It quantifies how much more or less likely a covariate value x is to appear in the target domain relative to the source domain. Consequently, under the covariate shift model, successful adaptation depends critically on estimating this ratio accurately and reliably. This reweighting approach is commonly referred to as instance-based domain adaptation.

A naive strategy is to estimate the source and target marginal densities separately and then form their ratio. In practice, however, this indirect approach can be highly unstable, especially in moderate dimensions (Härdle et al., 2004). This difficulty has led to a substantial literature on direct density-ratio estimation, which targets $r(x)$ itself without separately estimating p_S and p_T. Widely used methods include Kernel Mean Matching (KMM) (Huang et al., 2006), the Kullback-Leibler Importance Estimation Procedure (KLIEP) (Sugiyama et al., 2007b) and Unconstrained Least-Squares Importance Fitting (uLSIF) (Kanamori et al., 2009). KMM aligns source and target covariate distributions through kernel mean matching; KLIEP estimates the ratio by minimizing a Kullback-Leibler (KL) discrepancy; uLSIF uses a least-squares criterion that is closely connected to Pearson divergence. A log-linear variant of KLIEP further removes the normalization constraints in optimization through a log-linear ratio parameterization, improving scalability for large target samples (Tsuboi et al., 2009). These methods have become standard tools for instance-based adaptation under covariate shift and continue to be used in recent applications across fields such as astronomy, agriculture, and neuroscience (Gilda et al., 2024; Yang et al., 2025; Olza et al., 2025).

In practice, direct density-ratio estimation can be fragile in the presence of outliers, local contamination, or heavy-tailed target covariates (Yamada et al., 2011). Because the estimated ratios are used as importance weights in downstream learning, this instability can substantially degrade target-domain prediction (Yu et al., 2026). However, the impact of target contamination on instance-based domain adaptation has received comparatively limited attention, motivating more robust criteria for importance-weight estimation from contaminated samples.

So far, robustness in density-ratio estimation has been studied through several approaches. Sugiyama et al. (2012b) developed a robust Bregman density-ratio estimator based on positive power divergences, which downweights observations with small fitted ratios. Although their framework can readily accommodate the Hellinger divergence, their robustness analysis focuses on the positive-power class and does not examine the contrasting ratio-dependent behavior of the Hellinger target score, whose added-point contribution decays more rapidly in high-ratio regions. Relative unconstrained least-squares importance fitting (RuLSIF) (Yamada et al., 2011) estimates a relative density ratio, providing a biased but often more stable alternative to uLSIF, although its behavior under contamination remains less well understood. Trimmed-KLIEP (Liu et al., 2017) explicitly removes samples with large estimated density ratios, but requires a pre-specified trimming proportion: under-trimming leaves the estimator sensitive to contamination, whereas over-trimming can discard informative observations. More recently, Nagumo & Fujisawa (2024) proposed Weighted-DRE and γ-DRE, based on the KL divergence and the γ-divergence, respectively. These methods use smooth downweighting to achieve robustness to both source and target contamination, but introduce additional weighting or divergence parameters. Both methods, as well as Trimmed-KLIEP, use a log-linear density-ratio parameterization that enforces strictly positive ratios everywhere.

These limitations motivate a complementary approach to robust adaptation under contaminated covariate shift. Rather than introducing trimming, reweighting, or an additional robustness-specific tuning parameter, we ask whether robustness can emerge from the geometry of the divergence itself. The classical robustness properties of minimum Hellinger distance estimation (Beran, 1977; Lindsay, 1994) make the Hellinger divergence a natural candidate for this investigation. Existing α-divergence density-ratio formulations include the Hellinger divergence as a special case (Kitazawa, 2024), but their analysis focuses on optimization stability and does not examine sensitivity to target contamination or its implications for downstream covariate-shift adaptation.

Our contributions are threefold. First, we develop the Hellinger Importance Estimation Procedure (HellingerIEP), a nonnegative kernel estimator based on a variational Hellinger criterion for stable domain adaptation. Second, we establish consistency and asymptotic normality for a fixed kernel

dictionary and regularization level, and derive the estimator's local sensitivity to target contamination. Under a common square-root parameterization, we show that the Hellinger criterion dampens high-ratio target-score contributions more strongly than the KL and Pearson criteria. Third, we evaluate HellingerIEP on synthetic and benchmark covariate-shift tasks, demonstrating competitive performance in clean settings and greater stability under the contamination settings considered.

2 Hellinger Importance Estimation Procedure

We now introduce HellingerIEP. Its population criterion follows from general variational frameworks for f- and α-divergence density-ratio estimation (Nguyen et al., 2010; Sugiyama et al., 2012a; Kitazawa, 2024). We present a nonnegative kernel formulation of this criterion and analyze its statistical properties in Section 3.

Let $r(x) = p_T(x)/p_S(x)$ denote the true target-to-source density ratio, where p_S and p_T are the source and target marginal densities of X, respectively. Our estimator is defined through a population criterion whose minimizer is exactly r. Consider the objective

$$J(w) = \mathbb{E}_{P_S}\left[\sqrt{w(X)}\right] + \mathbb{E}_{P_T}\left[w(X)^{-1/2}\right], \qquad w : \mathcal{X} \to [0, \infty), \tag{1}$$

where J is an extended-real-valued functional, with $w^{-1/2}$ defined to be $+\infty$ on $\{x : w(x) = 0\}$. The next result shows that the minimizer of this criterion identifies the true density ratio.

Proposition 2.1. *Assume $P_T^X \ll P_S^X$, and let $r(x) = p_T(x)/p_S(x)$. Then $J(w)$ is uniquely minimized, P_S^X-almost surely, at $w^*(x) = r(x)$.*

The proof is provided in Appendix A. Proposition 2.1 shows that the criterion in equation 1 is Fisher consistent for density-ratio estimation. We therefore define HellingerIEP by minimizing an empirical version of J.

Remark 2.1 (Effect of restriction)**.** *The proof also gives the exact excess-criterion identity*

$$J(w) - J(r) = \mathbb{E}_{P_S^X}\left[\left(\sqrt{w(X)} - \sqrt{r(X)}\right)^2 w(X)^{-1/2}\right].$$

Consequently, over a restricted ratio class $\mathcal{W}$, the dual Hellinger criterion minimizes

$$\mathbb{E}_{P_S^X}\left[\left(\sqrt{w(X)} - \sqrt{r(X)}\right)^2 w(X)^{-1/2}\right], \quad w \in \mathcal{W},$$

rather than the primal Hellinger projection $\mathbb{E}_{P_S^X}\left[\left(\sqrt{w(X)} - \sqrt{r(X)}\right)^2\right]$. Thus, variational dualization preserves the unrestricted minimizer $w = r$, but restriction to a model class changes the projection geometry.

For computation, we parameterize the square root of the density ratio, $g(x) = \sqrt{w(x)}$, so that $w(x) = g(x)^2$. This parameterization yields a convex optimization problem in the kernel coefficients. We also consider a direct parameterization of w in Appendix E. The resulting estimator exhibits similar empirical performance; see Appendix F.

Under the square-root parameterization, equation 1 becomes

$$J(g) = \mathbb{E}_{P_S}\left[g(X)\right] + \mathbb{E}_{P_T}\left[g(X)^{-1}\right], \qquad g(x) \geq 0,$$

where $1/g(x)$ is defined to be $+\infty$ on $\{x : g(x) = 0\}$. The population minimizer is $g^*(x) = \sqrt{r(x)}$. Given source observations $\{X_i^S\}_{i=1}^{n_S}$ and target observations $\{X_j^T\}_{j=1}^{n_T}$, we estimate g within a finite-dimensional kernel class

$$\mathcal{G}_M = \left\{ g(\cdot) = \sum_{m=1}^{M} \beta_m K(\cdot, c_m) : \ \beta_m \geq 0 \right\},$$

where K is a nonnegative kernel and $\{c_m\}_{m=1}^M$ are kernel centers selected from the target sample. Kernel expansions are widely used in direct density-ratio estimation, including KLIEP and uLSIF. We use this representation because it provides a flexible and computationally convenient approximation class while preserving nonnegativity through the coefficient constraints.

Within $\mathcal{G}_M$, we define the regularized estimator $\widehat{\beta} = \operatorname{argmin}_{\beta\in[0,\infty)^M} \widehat{J}_{n,\lambda}(\beta)$ with $\widehat{J}_{n,\lambda}$ given by

$$\widehat{J}_{n,\lambda}(\beta) = \frac{1}{n_S}\sum_{i=1}^{n_S}\sum_{m=1}^{M}\beta_m K(X_i^S, c_m) + \frac{1}{n_T}\sum_{j=1}^{n_T}\left(\sum_{m=1}^{M}\beta_m K(X_j^T, c_m)\right)^{-1} + \frac{\lambda}{2}\|\beta\|_2^2, \qquad (2)$$

where $\lambda \geq 0$ controls the amount of ℓ_2 regularization. After obtaining $\widehat{\beta}$, we define $\widehat{g}(x) = \sum_{m=1}^{M}\widehat{\beta}_m K(x, c_m)$ and $\widehat{w}(x) = \widehat{g}(x)^2$. Because both the kernel and the coefficients are non-negative, $\widehat{g}(x) \geq 0$ for all x. In practice, the kernel hyperparameters and regularization parameter λ are selected jointly by cross-validation using the unpenalized empirical objective.

2.1 Connection to Hellinger Divergence

We now provide intuition for the population objective in equation 1 by showing its connection to the Hellinger divergence in the following corollary. The proof is provided in Appendix B.

Corollary 2.1. *Let $\varphi(u) = (\sqrt{u}-1)^2$ and $D_\varphi(P_T^X \| P_S^X)$ denote the Hellinger divergence. Then*

$$D_\varphi(P_T^X \| P_S^X) = \mathbb{E}_{P_S}\left[\left(\sqrt{r(X)}-1\right)^2\right] = 2 - \inf_{w>0} J(w),$$

where the second equality follows from the variational representation of f-divergences (Nguyen et al., 2010) under the change of variable $f(x) = 1 - w(x)^{-1/2}$. Consequently, minimizing equation 1 is equivalent to maximizing the Hellinger variational objective.

The Hellinger divergence is a member of the broader power-divergence family (Cressie & Read, 1984). Appendix H compares HellingerIEP with alternative members of this family and investigates the trade-off between clean-sample efficiency and stability. We also consider additional robust formulations, including the negative-exponential-divergence estimator NEDIEP, in Appendix E, with further empirical comparisons reported in Appendix F.

3 Theoretical Results

We establish statistical guarantees for HellingerIEP and characterize its local sensitivity to contamination in the target covariate distribution. We continue to use the square-root parameterization described in the previous section.

For the following analysis of consistency and asymptotic normality, we treat the kernel dictionary $c_1, \ldots, c_M$ and the regularization level $\lambda > 0$ as fixed. For notational convenience, let

$$\phi(x) = \big(K(x, c_1), \ldots, K(x, c_M)\big)^\top, \qquad g_\beta(x) = \beta^\top \phi(x),$$

and define the parameter space $\mathcal{C} = \big\{\beta \in [0,\infty)^M : \|\beta\|_1 > 0\big\}$. Using this notation, the empirical objective in equation 2 can be written as

$$\widehat{J}_{n,\lambda}(\beta) = \frac{1}{n_S}\sum_{i=1}^{n_S} g_\beta(X_i^S) + \frac{1}{n_T}\sum_{j=1}^{n_T}\frac{1}{g_\beta(X_j^T)} + \frac{\lambda}{2}\|\beta\|_2^2,$$

and we define $\widehat{\beta}_n = \arg\min_{\beta\in\mathcal{C}} \widehat{J}_{n,\lambda}(\beta)$, $\widehat{g}_n = g_{\widehat{\beta}_n}$, and $\widehat{w}_n = (\widehat{g}_n)^2$. The corresponding population objective is

$$J_\lambda(\beta) = \mathbb{E}_{P_S^X}\left[g_\beta(X)\right] + \mathbb{E}_{P_T^X}\left[\frac{1}{g_\beta(X)}\right] + \frac{\lambda}{2}\|\beta\|_2^2. \qquad (3)$$

Similarly, define $\beta_\lambda^\star = \arg\min_{\beta\in\mathcal{C}} J_\lambda(\beta)$, $g_\lambda^\star = g_{\beta_\lambda^\star}$, and $w_\lambda^\star = (g_\lambda^\star)^2$.

Assumption 3.1. *The space $\mathcal{X}$ is compact, and $K : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ is continuous. There exist constants $0 < \kappa_0 < \kappa_1 < \infty$ such that $\kappa_0 \leq K(x, c_m) \leq \kappa_1$ for every $x \in \mathcal{X}$ and $m = 1, \ldots, M$. The centers $c_1, \ldots, c_M$ are fixed.*

Assumption 3.2. *The source and target samples are independent, with $X_i^S \overset{\text{i.i.d.}}{\sim} P_S^X$ and $X_j^T \overset{\text{i.i.d.}}{\sim} P_T^X$. Let $n = n_S + n_T$ and assume $n_S/n \to \rho \in (0,1)$ as $n \to \infty$.*

Assumption 3.3. *$M, \lambda > 0$ are fixed and independent of n.*

Theorem 3.1 (Consistency and asymptotic normality)**.** *Suppose Assumptions 3.1–3.3 hold. Then, as $n \to \infty$, the following hold:*

$$\widehat{\beta}_n \xrightarrow{p} \beta^\star_\lambda, \tag{4}$$

$$\sup_{x\in\mathcal{X}} |\widehat{g}_n(x) - g^\star_\lambda(x)| \xrightarrow{p} 0, \tag{5}$$

$$\sup_{x\in\mathcal{X}} |\widehat{w}_n(x) - w^\star_\lambda(x)| \xrightarrow{p} 0, \tag{6}$$

$$\left\|\widehat{\beta}_n - \beta^\star_\lambda\right\|_2 = O_p\left(\sqrt{1/n}\right). \tag{7}$$

If we further assume $\beta^\star_{\lambda,m} > 0$ for all $m \in \{1, 2, \ldots, M\}$, then

$$\sqrt{n}\left(\widehat{\beta}_n - \beta^\star_\lambda\right) \xrightarrow{d} \mathcal{N}\left(0,\ A_\lambda^{-1}\Sigma_\lambda A_\lambda^{-1}\right),$$

where

$$A_\lambda = \nabla^2 J_\lambda(\beta^\star_\lambda), \quad \Sigma_\lambda = \frac{1}{\rho}\operatorname{Var}_{P_S^X}\left[\phi(X)\right] + \frac{1}{1-\rho}\operatorname{Var}_{P_T^X}\left[\frac{\phi(X)}{(g^\star_\lambda(X))^2}\right].$$

Theorem 3.1 establishes uniform convergence of the estimated importance weights to the corresponding population optimum $w^\star_\lambda$, for a fixed kernel dictionary and $\lambda > 0$. However, the regularized finite-dimensional population target $w^\star_\lambda$ need not equal the unrestricted density ratio r. The difference may reflect both approximation error from the restricted kernel class and regularization bias. Under the additional interiority condition, the coefficient estimator also has a limiting Gaussian distribution that supports standard large-sample inference. The proof is provided in Appendix C. Because the result is conditional on a fixed dictionary and fixed regularization level, it does not cover the data-dependent selection of kernel centers, bandwidths, or λ used in the experiments.

We next study the local sensitivity of HellingerIEP to contamination of the target covariate distribution. For $\lambda \geq 0$, define the population Hellinger criterion

$$Q_{H,\lambda}(\beta; P_S, P_T) = \mathbb{E}_{P_S}[g_\beta(X)] + \mathbb{E}_{P_T}\left[g_\beta(X)^{-1}\right] + \frac{\lambda}{2}\|\beta\|_2^2, \qquad \beta \in \mathcal{C},$$

and let $\beta^\star_{H,\lambda}$ denote its population minimizer. Following the infinitesimal-contamination framework of Hampel (1974), consider

$$P^X_{T,\varepsilon,z} = (1-\varepsilon)P_T^X + \varepsilon\Delta_z, \qquad 0 \leq \varepsilon \ll 1, \tag{8}$$

where Δ_z denotes the point mass at z, and let $\beta^\star_{H,\lambda,\varepsilon}$ denote the corresponding population minimizer.

Proposition 3.1 (Local target-contamination sensitivity)**.** *Suppose that $Q_{H,\lambda}$ is twice continuously differentiable in β in a neighborhood of $\beta^\star_{H,\lambda}$, and that differentiation may be interchanged with expectation. Assume that $\beta^\star_{H,\lambda} \in (0,\infty)^M$. Define*

$$A_{H,\lambda} = \nabla^2_\beta Q_{H,\lambda}(\beta^\star_{H,\lambda}; P_S, P_T) = 2\mathbb{E}_{P_T}\left[\frac{\phi(X)\phi(X)^\top}{g_{\beta^\star_{H,\lambda}}(X)^3}\right] + \lambda I_M.$$

When $\lambda = 0$, assume additionally that $A_{H,0}$ is nonsingular. Fix a contamination point z such that $g_{\beta^\star_{H,\lambda}}(z) > 0$. Then, for all sufficiently small $\varepsilon \geq 0$, the contaminated criterion has a unique interior minimizer $\beta^\star_{H,\lambda,\varepsilon}$, and the map $\varepsilon \longmapsto \beta^\star_{H,\lambda,\varepsilon}$ is differentiable at $\varepsilon = 0$, with

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\beta^\star_{H,\lambda,\varepsilon}\bigg|_{\varepsilon=0} = -A^{-1}_{H,\lambda}\left\{\psi_H(z;\beta^\star_{H,\lambda}) - \mathbb{E}_{P_T}\left[\psi_H(X;\beta^\star_{H,\lambda})\right]\right\},$$

where $\psi_H(x;\beta) = -\phi(x)/g_\beta(x)^2 = -\phi(x)/w_\beta(x)$. Moreover, for every evaluation point $x \in \mathcal{X}$,

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}w_{\beta^\star_{H,\lambda,\varepsilon}}(x)\bigg|_{\varepsilon=0} = 2g_{\beta^\star_{H,\lambda}}(x)\phi(x)^\top \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\beta^\star_{H,\lambda,\varepsilon}\bigg|_{\varepsilon=0}.$$

For comparison with KLIEP and uLSIF, consider all three methods under the common square-root parameterization $w_\beta = g_\beta^2$. Their target-score contributions are

$$\psi_H(x;\beta) = -\frac{\phi(x)}{w_\beta(x)}, \qquad \psi_K(x;\beta) = -\frac{2\phi(x)}{\sqrt{w_\beta(x)}}, \qquad \psi_U(x;\beta) = -2\sqrt{w_\beta(x)}\,\phi(x).$$

The proof is given in Appendix D. Under the same parameterization, the target score scales as w^{-1} for HellingerIEP, $w^{-1/2}$ for KLIEP, and $w^{1/2}$ for uLSIF. With fixed method-specific curvature, the added-point Hellinger score contribution is attenuated more strongly in high-ratio regions than the corresponding KL contribution, whereas the least-squares (Pearson) score increases with the fitted ratio. This is a score-level comparison, not a global ordering of influence functions, since the resulting parameter perturbations also depend on the centering term, the constraints, and method-specific curvature.

As $w_\beta(x) \downarrow 0$, the ordering reverses. The Hellinger score is most singular, scaling as w^{-1}, compared with $w^{-1/2}$ for KLIEP, while the uLSIF score vanishes as $w^{1/2}$. The Hellinger criterion therefore imposes the strongest local barrier against collapsing a target point toward zero fitted weight. Hence, the score-level comparison is not uniform: HellingerIEP provides stronger high-ratio attenuation but greater low-ratio sensitivity. This nonuniformity suggests a potential stability–adaptation trade-off, since a large fitted ratio may represent either contamination or genuine target-domain shift.

Remark 3.1. *When P_S^X is nonatomic, the contaminated target law equation 8 is generally not absolutely continuous with respect to P_S^X. The proposition therefore describes the local sensitivity of the fitted projection functional to an added target observation; it does not assert that the contaminated target law has an ordinary density ratio with respect to P_S^X.*

4 Experiments

In the following experiments, we compare HellingerIEP with KLIEP, uLSIF, and RuLSIF to evaluate density-ratio estimation and downstream prediction under target contamination. Throughout, tunable hyperparameters are selected by cross-validation. Additional comparisons with robust alternatives, including Trimmed-KLIEP, WeightedDRE, and γ-DRE, are provided in Appendix G, where HellingerIEP achieves comparable or superior performance without requiring an additional robustness-specific tuning parameter.

Density-Ratio Estimation. We first study density-ratio estimation performance in the d-dimensional Gaussian distribution shift used by Sugiyama et al. (2007b) and Kanamori et al. (2009). The source distribution is $p_S(x) = \mathcal{N}(0, I_d)$, and the clean target distribution is $p_T(x) = \mathcal{N}((1, 0, \ldots, 0)^\top, I_d)$. The goal is to estimate the importance ratio at the source sample points.

In addition to reporting performance on clean data, we assess robustness by contaminating the target sample used for fitting. Specifically, with the contamination level fixed at $\alpha = 0.3$, an α-fraction of target observations is replaced by the point $(1, 0, \ldots, 0)^\top$, i.e., the mean of the clean target distribution. This yields a simple robustness check based on a localized spike perturbation.

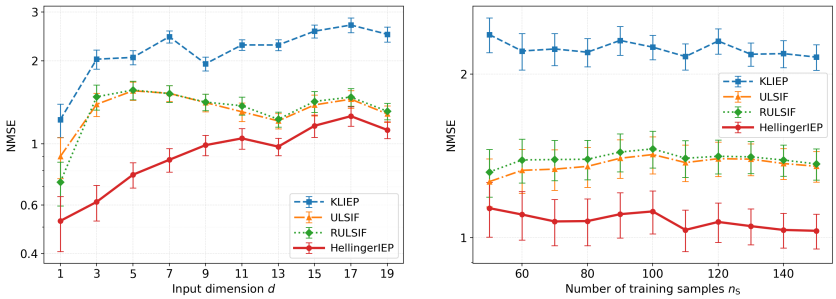
In this setting with known source and target distributions, we can evaluate the clean oracle density ratio exactly at the source sample points. Following the original KLIEP and uLSIF studies (Sugiyama et al., 2007b; Kanamori et al., 2009; 2012), we assess density-ratio estimation accuracy using the normalized mean squared error (NMSE),

$$\mathrm{NMSE} = \frac{1}{n_S} \sum_{i=1}^{n_S} \left(\frac{\hat{w}_i}{\frac{1}{n_S}\sum_{j=1}^{n_S} \hat{w}_j} - \frac{w_i}{\frac{1}{n_S}\sum_{j=1}^{n_S} w_j} \right)^2,$$

where w_i denotes the oracle density ratio between the clean target distribution and the source distribution, and $\hat{w}_i$ denotes the estimated density ratio obtained from the possibly contaminated target sample. Normalizing both vectors removes overall scale differences and focuses the comparison on recovery of the relative weighting structure for the clean target population.

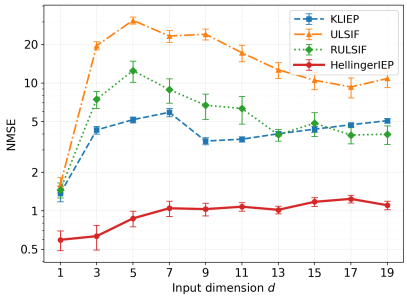
We consider two experimental regimes. In the first, we fix the source sample size $n_S = 100$ and vary the dimension over $d \in \{1, 3, 5, \ldots, 19\}$, while holding the target sample size at $n_T = 1000$. In the second, we fix $d = 10$ and vary the source sample size over $n_S \in \{50, 60, \ldots, 150\}$, again with $n_T = 1000$. Each configuration is repeated over 50 Monte Carlo replicates and all methods use cross-validation to select hyperparameters. Throughout our experiments, the RuLSIF relative-density-ratio parameter is fixed at the commonly used value 0.1.

The results in Figure 1 show that in the uncontaminated setting, HellingerIEP is competitive and typically attains the smallest NMSE across both experimental regimes. Under 30% target contami-

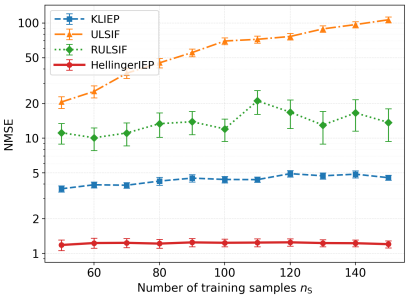


(a) Varying d with no contamination.

(b) Varying n_S with no contamination.



(c) Varying d with 30% contamination.



(d) Varying n_S with 30% contamination.

Figure 1: Density-ratio estimation NMSE for KLIEP, uLSIF, RuLSIF, and HellingerIEP when varying dimension d (left panels) or source sample size n_S (right panels), with no contamination ($\alpha = 0$; top row) and 30% target contamination ($\alpha = 0.3$; bottom row). Points and error bars denote means and standard errors over 50 replicates (lower is better).

nation, its advantage becomes more pronounced, whereas the competing estimators deteriorate more substantially. The pattern is consistent as the dimension or the source sample size increases.

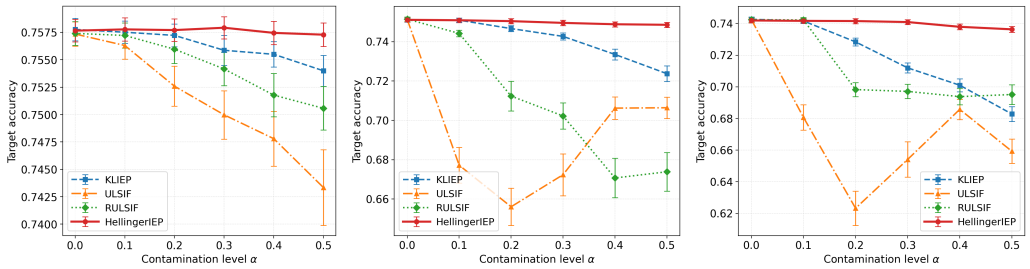
Downstream Predictive Performance. We next check if robustness in density-ratio estimation translates into improved downstream prediction under contamination. We consider both binary classification and regression tasks in dimensions $d \in \{1, 5, 10\}$ and across varying levels of target contamination. To separate the effect of dimensionality from the difficulty of the prediction problem, only the first covariate contains predictive information and exhibits covariate shift. All remaining coordinates are non-informative and identically distributed in the source and target domains. We report the classification results in the main text, with regression results in Appendix F.2 and additional experiments under model misspecification in Appendix F.3.

For classification tasks, the class label is first sampled in both domains as $Y \sim \text{Bernoulli}(0.5)$. Conditional on Y, the first coordinate is generated according to

$$X_1^S \mid Y \sim \begin{cases} \mathcal{N}(2,4), & Y = 1, \\ \mathcal{N}(-2,4), & Y = 0, \end{cases} \qquad X_1^T \mid Y \sim \begin{cases} \mathcal{N}(1,2), & Y = 1, \\ \mathcal{N}(-1,2), & Y = 0 \end{cases}.$$

For $j = 2, \ldots, d$ if $d \geq 2$, the remaining coordinates are generated independently as $X_j^S, X_j^T \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, independently of Y. Under this construction, the target distribution of the first coordinate is more concentrated around the origin than the source distribution, while the other coordinates contain neither class information nor source-target shift. Moreover, the conditional distribution $P(Y \mid X)$ is identical across the two domains, and the difference between them arises solely through $P(X)$.

To assess robustness, we contaminate the unlabeled target sample used for adaptation. For each contamination level $\alpha \in \{0, 0.1, \ldots, 0.5\}$, an α-fraction of the target training covariates is replaced by draws from the Gaussian spike $\mathcal{N}(0, 0.001^2 I_d)$. The downstream test set, however, remains uncontaminated, so the experiment measures how contamination in the unlabeled target sample affects adaptation to the true target population.



(a) $d = 1$ (b) $d = 5$ (c) $d = 10$

Figure 2: Target classification accuracy under contaminated covariate shift for $d \in \{1, 5, 10\}$, comparing HellingerIEP, KLIEP, uLSIF, and RuLSIF. Points denote Monte Carlo means and error bars denote standard errors over 50 replicates. Higher values indicate better predictive performance.

For each replicate, we generate $n_S = 300$ labeled source observations, $n_T = 300$ unlabeled target observations for adaptation, and an independent clean labeled target test set of size 3000. We first estimate importance weights and then fit a weighted logistic regression model to the source sample for each method. Figure 2 compares downstream target-domain classification performance across contamination levels and dimensions. Results are averaged over 50 replicates.

Across all dimensions, HellingerIEP degrades more slowly as the contamination level increases and maintains higher target-domain accuracy than the competing methods. The differences are more pronounced under heavier contamination, suggesting that HellingerIEP is less sensitive to target contamination in this setting.

We also examine whether this sensitivity depends on the location of the contamination in Appendix I. In a separate one-dimensional experiment, we move the contamination spike across the covariate space while holding its proportion fixed. The results show that its impact varies with the location of the contamination relative to the source and target distributions, while HellingerIEP achieves better downstream performance than KLIEP, uLSIF, and RuLSIF across the range considered.

Benchmark Datasets. We also evaluate HellingerIEP on benchmark regression and classification datasets under controlled covariate shift and target contamination. For each dataset, we choose the source, unlabeled target, and test sample sizes based on the number of available observations. Following the KLIEP and uLSIF studies (Sugiyama et al., 2007b; Kanamori et al., 2009), we rescale each feature to $[0, 1]$ and, in each repetition, select a feature c uniformly at random. We inspect observations in random order, accepting each into the target pool with probability $\min\{1, 4(x^{(c)})^2\}$ and removing it from the sampling pool whether accepted or rejected. Once the desired target-pool size is reached, we draw the source sample uniformly from the remaining observations. This procedure induces covariate shift because observations with higher values of the selected feature are more likely to appear in the target domain. We then split the shifted target pool into an unlabeled target sample for density-ratio estimation and a separate holdout target sample for downstream evaluation.

After sampling the clean source and target, we add contamination only into the unlabeled target covariates. For each contamination level $\alpha \in \{0, 0.1, \ldots, 0.5\}$, we randomly replace an α-fraction of these covariate vectors with the feature-wise mean of the clean unlabeled target sample. The resulting contaminated target sample is used only for estimating importance weights, while downstream prediction is always evaluated on the separate clean holdout target sample. For downstream prediction, we use the estimated density ratios as sample weights in weighted logistic regression for classification and weighted linear regression for regression. Each benchmark configuration is repeated over 100 independently generated replicates.

Table 1 reports the mean and standard deviation of classification accuracy and root mean squared error (RMSE) for regression. It shows HellingerIEP delivers strong downstream performance across both regression and classification benchmarks. In the clean setting ($\alpha = 0$), its downstream performance is comparable to that of the other direct density-ratio estimators. As contamination increases, HellingerIEP generally degrades less and achieves better downstream performance relative to the competing methods.

Table 1: Benchmark performance of HellingerIEP and competing methods under increasing contamination level α. For the regression datasets, lower RMSE is better; for the classification datasets, higher accuracy is better. Results are reported as mean (standard deviation) across 100 replicates. The best mean performance within each dataset and contamination level is shown in bold.

Dataset	Method	$\alpha = 0.0$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$
Regression							
Diabetes (RMSE)	**HellingerIEP**	**60.18 (4.61)**	**60.39 (4.61)**	**61.02 (5.28)**	**60.34 (4.59)**	**61.10 (5.30)**	**61.03 (5.69)**
	KLIEP	66.32 (7.79)	69.61 (8.90)	71.37 (8.30)	73.60 (10.62)	77.33 (11.80)	80.62 (13.13)
	uLSIF	63.08 (8.21)	74.22 (26.97)	100.61 (40.46)	68.98 (8.97)	86.45 (38.59)	120.43 (54.13)
	RuLSIF	75.83 (29.91)	80.23 (30.62)	63.80 (5.32)	67.55 (10.49)	71.03 (17.81)	76.87 (33.15)
Abalone (RMSE)	**HellingerIEP**	**2.70 (0.54)**	**2.72 (0.65)**	**2.71 (0.55)**	**2.74 (0.63)**	**2.72 (0.61)**	**2.75 (0.65)**
	KLIEP	2.79 (0.58)	2.78 (0.62)	2.81 (0.65)	2.87 (0.86)	2.89 (0.79)	2.83 (0.71)
	uLSIF	2.87 (0.82)	2.87 (0.76)	2.93 (0.74)	3.01 (0.75)	3.09 (0.75)	3.15 (0.75)
	RuLSIF	2.84 (0.79)	2.87 (0.75)	2.90 (0.77)	3.04 (0.74)	3.12 (0.75)	3.18 (0.75)
Classification							
Breast Cancer (Accuracy)	**HellingerIEP**	**0.9459 (0.0197)**	0.9459 (0.0197)	**0.9456 (0.0204)**	**0.9461 (0.0209)**	**0.9454 (0.0209)**	**0.9456 (0.0254)**
	KLIEP	0.9447 (0.0187)	**0.9464 (0.0168)**	0.9377 (0.0196)	0.9279 (0.0294)	0.9066 (0.0386)	0.8729 (0.0499)
	uLSIF	0.9328 (0.0237)	0.8619 (0.1464)	0.8300 (0.1619)	0.9163 (0.0508)	0.9084 (0.0629)	0.8782 (0.1168)
	RuLSIF	0.9259 (0.0587)	0.9024 (0.1033)	0.9266 (0.0388)	0.8559 (0.1451)	0.7364 (0.1884)	0.7786 (0.1779)
Wine (Accuracy)	**HellingerIEP**	**0.9190 (0.0457)**	**0.9138 (0.0486)**	**0.9092 (0.0501)**	**0.9098 (0.0520)**	**0.8998 (0.0588)**	**0.8972 (0.0591)**
	KLIEP	0.9122 (0.0511)	0.8966 (0.0662)	0.8728 (0.0728)	0.8184 (0.0915)	0.7644 (0.1036)	0.7028 (0.1219)
	uLSIF	0.8848 (0.0761)	0.7838 (0.1530)	0.7890 (0.1516)	0.7344 (0.1542)	0.6958 (0.1500)	0.6150 (0.1823)
	RuLSIF	0.8870 (0.0968)	0.8002 (0.1349)	0.8320 (0.1365)	0.7674 (0.1504)	0.7106 (0.1360)	0.6860 (0.1520)

5 Discussion and Conclusion

We introduce HellingerIEP, a nonnegative kernel density-ratio estimator for instance-based domain adaptation with a contaminated unlabeled target sample. Its unrestricted variational criterion is Fisher consistent for the target-to-source density ratio, and the square-root parameterization yields a convex optimization problem. At the same time, restriction to a kernel class changes the primal–dual projection geometry. Consequently, the classical robustness properties of minimum Hellinger estimation do not automatically transfer to HellingerIEP, and the stability of the restricted dual estimator must be studied directly.

For a fixed kernel dictionary and regularization level, we establish consistency for the corresponding regularized finite-dimensional population minimizer and asymptotic normality under an interiority condition. Our local target-contamination analysis uses the same square-root parameterization as the estimator. Under this common parameterization, the Hellinger, KL, and Pearson target-score factors scale as w^{-1}, $w^{-1/2}$, and $w^{1/2}$, respectively. Thus, the Hellinger criterion attenuates added-point score contributions more strongly in high-ratio regions but is more singular in low-ratio regions. This comparison is nonuniform, and the complete parameter perturbations also depend on centering, constraints, and method-specific curvature.

Across the contamination mechanisms considered, HellingerIEP is competitive in clean settings and generally more stable in density-ratio estimation and downstream prediction. The misspecified-model experiments in the appendix, which include source-only and oracle-weighted baselines, further indicate that this stability can coexist with a genuine adaptation benefit in the settings studied. The accompanying NEDIEP comparison also illustrates that other variational criteria can provide different forms of contamination control. Taken together, the results suggest that HellingerIEP provides a computationally attractive alternative for stable importance weighting under contaminated covariate shift, while remaining competitive in the clean settings considered here.

Several limitations remain. Our contamination analysis is local and does not establish finite-contamination robustness or downstream target-risk guarantees. The asymptotic theory assumes a fixed finite-dimensional kernel dictionary, a fixed positive regularization level, and a compact covariate domain; it therefore does not cover data-dependent center selection, cross-validation, or the unbounded-support Gaussian experiments. In addition, the empirical analysis uses controlled contamination mechanisms. Developing finite-contamination guarantees, target-risk bounds, data-adaptive asymptotic theory, and extensions to broader forms of distribution shift are important directions for future work.

AI USE STATEMENT

Large language models were used for literature searches, language editing, and writing/debugging experimental code. The authors reviewed all references, AI-assisted code, and manuscript revisions. All experiments were conducted by the authors, who take full responsibility for the final content of the paper.

REFERENCES

Ayanendranath Basu, Sahadeb Sarkar, and A. N. Vidyashankar. Minimum negative exponential disparity estimation in parametric models. *Journal of Statistical Planning and Inference*, 58(2): 349–370, 1997. doi: 10.1016/S0378-3758(96)00078-X.

Rudolf Beran. Minimum Hellinger distance estimates for parametric models. *The Annals of Statistics*, 5(3):445–463, 1977. URL http://www.jstor.org/stable/2958896.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Noel Cressie and Timothy RC Read. Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 46(3):440–464, 1984.

Geeta Y. Dharani, Nimisha G. Nair, Pallavi Satpathy, and Jabez Christopher. Covariate shift: A review and analysis on classifiers. In *2019 Global Conference for Advancement in Technology (GCAT)*, pp. 1–6, 2019. doi: 10.1109/GCAT47503.2019.8978471.

Abolfazl Farahani, Sahar Voghoei, Khaled Rasheed, and Hamid R. Arabnia. A brief review of domain adaptation. In Robert Stahlbock, Gary M. Weiss, Mahmoud Abou-Nasr, Cheng-Ying Yang, Hamid R. Arabnia, and Leonidas Deligiannidis (eds.), *Advances in Data Science and Information Engineering*, pp. 877–894, Cham, 2021. Springer International Publishing. doi: 10.1007/978-3-030-71704-9_65.

Sankalp Gilda, Antoine de Mathelin, Sabine Bellstedt, and Guillaume Richard. Unsupervised domain adaptation for constraining star formation histories. *Astronomy*, 3(3):189–207, 2024. doi: 10.3390/astronomy3030012.

Frank R. Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393, 1974. doi: 10.1080/01621459.1974.10482962.

Wolfgang Härdle, Axel Werwatz, Marlene Müller, and Stefan Sperlich. *Nonparametric and Semiparametric Models*. Springer Series in Statistics. Springer, Berlin, Heidelberg, 2004. doi: 10.1007/978-3-642-17146-8.

Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex Smola. Correcting sample selection bias by unlabeled data. In B. Schölkopf, J. Platt, and T. Hoffman (eds.), *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006. URL https://proceedings.neurips.cc/paper_files/paper/2006/file/a2186aa7c086b46ad4e8bf81e2a3a19b-Paper.pdf.

Uday Kamath, John Liu, and James Whitaker. Transfer learning: Domain adaptation. In *Deep Learning for NLP and Speech Recognition*, pp. 495–535. Springer International Publishing, Cham, 2019. doi: 10.1007/978-3-030-14596-5_11.

Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10(48):1391–1445, 2009. URL http://jmlr.org/papers/v10/kanamori09a.html.

Takafumi Kanamori, Taiji Suzuki, and Masashi Sugiyama. Statistical analysis of kernel-based least-squares density-ratio estimation. *Machine Learning*, 86(3):335–367, 2012. doi: 10.1007/s10994-011-5266-3.

Yoshiaki Kitazawa. Alpha-divergence loss function for neural density ratio estimation, 2024. URL https://arxiv.org/abs/2402.02041.

Bruce G. Lindsay. Efficiency versus robustness: The case for minimum Hellinger distance and related methods. *The Annals of Statistics*, 22(2):1081–1114, 1994. URL http://www.jstor.org/stable/2242307.

Song Liu, Akiko Takeda, Taiji Suzuki, and Kenji Fukumizu. Trimmed density ratio estimation. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/ea204361fe7f024b130143eb3e189a18-Paper.pdf.

Ryosuke Nagumo and Hironori Fujisawa. Density ratio estimation with doubly strong robustness. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 37260–37276. PMLR, 2024.

XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010. doi: 10.1109/TIT.2010.2068870.

Alexander Olza, David Soto, and Roberto Santana. Domain adaptation-enhanced searchlight: Enabling classification of brain states from visual perception to mental imagery. *Brain Informatics*, 12(1):17, 2025. doi: 10.1186/s40708-025-00263-0.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000. doi: 10.1016/S0378-3758(00)00115-4.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(35):985–1005, 2007a. URL http://jmlr.org/papers/v8/sugiyama07a.html.

Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In J. Platt, D. Koller, Y. Singer, and S. Roweis (eds.), *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007b. URL https://proceedings.neurips.cc/paper_files/paper/2007/file/be83ab3ecd0db773eb2dc1b0a17836a1-Paper.pdf.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density ratio estimation in machine learning*. Cambridge University Press, 2012a.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. Density-ratio matching under the Bregman divergence: A unified framework of density-ratio estimation. *Annals of the Institute of Statistical Mathematics*, 64(5):1009–1044, 2012b. doi: 10.1007/s10463-011-0343-8.

Yuta Tsuboi, Hisashi Kashima, Shohei Hido, Steffen Bickel, and Masashi Sugiyama. Direct density ratio estimation for large-scale covariate shift adaptation. *Journal of Information Processing*, 17: 138–155, 2009. doi: 10.2197/ipsjjip.17.138.

Makoto Yamada, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Masashi Sugiyama. Relative density-ratio estimation for robust distribution comparison. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper_files/paper/2011/file/d1f255a373a3cef72e03aa9d980c7eca-Paper.pdf.

Ming-Der Yang, Yu-Chun Hsu, Tsai-Ting Liu, and Han-Hui Huang. Enhancing grain moisture prediction in multiple crop seasons using domain adaptation AI. *Computers and Electronics in Agriculture*, 231:110058, 2025. doi: 10.1016/j.compag.2025.110058.

Chenglin Yu, Zhengyu Zhou, and Weiwei Liu. Quantitative analysis of the effect of density ratio estimation in covariate shift adaptation. *Transactions on Machine Learning Research*, 2026.

A Proof of Proposition 2.1

Proof. Using $\mathrm{d}P_T^X = r\,\mathrm{d}P_S^X$, we obtain

$$J(w) - J(r) = \mathbb{E}_{P_S^X}\left[\sqrt{w(X)} + \frac{r(X)}{\sqrt{w(X)}} - 2\sqrt{r(X)}\right]$$
$$= \mathbb{E}_{P_S^X}\left[\frac{\left\{\sqrt{w(X)} - \sqrt{r(X)}\right\}^2}{\sqrt{w(X)}}\right] \geq 0.$$

On the set where $w(X) = 0$, the integrand is interpreted as 0 if $r(X) = 0$ and $+\infty$ otherwise. Equality holds if and only if $w = r$, P_S^X-almost surely. Therefore, J is uniquely minimized at $w = r$. □

B Proof of Corollary 2.1

Proof. Let $\varphi(u) = (\sqrt{u} - 1)^2$ for $u \geq 0$. Its Fenchel conjugate is

$$\varphi^*(t) = \sup_{u \geq 0}\{tu - \varphi(u)\}.$$

Writing $v = \sqrt{u}$, we obtain

$$\varphi^*(t) = \sup_{v \geq 0}\left\{(t-1)v^2 + 2v - 1\right\}.$$

If $t \geq 1$, this supremum is $+\infty$. If $t < 1$, the expression is a concave quadratic in v, and its maximum is attained at $v = (1-t)^{-1}$. Substitution gives

$$\varphi^*(t) = \frac{t}{1-t}, \qquad t < 1.$$

Therefore,

$$\varphi^*(t) = \begin{cases} \dfrac{t}{1-t}, & t < 1, \\ +\infty, & t \geq 1 \end{cases}.$$

The variational representation of f-divergences (Nguyen et al., 2010) now gives

$$D_\varphi(P_T^X \| P_S^X) = \sup_{f<1}\left\{\mathbb{E}_{P_T^X}[f(X)] - \mathbb{E}_{P_S^X}\left[\frac{f(X)}{1-f(X)}\right]\right\},$$

where the supremum is over measurable functions for which the expectations are well defined. For every positive measurable function w, set

$$f(x) = 1 - \frac{1}{\sqrt{w(x)}}.$$

This transformation is a bijection between $w(x) > 0$ and $f(x) < 1$, with inverse $w(x) = (1 - f(x))^{-2}$. Moreover,

$$\frac{f(x)}{1-f(x)} = \sqrt{w(x)} - 1.$$

Consequently,

$$\mathbb{E}_{P_T^X}[f(X)] - \mathbb{E}_{P_S^X}\left[\frac{f(X)}{1-f(X)}\right] = \mathbb{E}_{P_T^X}\left[1 - \frac{1}{\sqrt{w(X)}}\right] - \mathbb{E}_{P_S^X}\left[\sqrt{w(X)} - 1\right]$$
$$= 2 - \mathbb{E}_{P_S^X}\left[\sqrt{w(X)}\right] - \mathbb{E}_{P_T^X}\left[w(X)^{-1/2}\right]$$
$$= 2 - J(w).$$

Taking the supremum over $f < 1$, equivalently over $w > 0$, yields

$$D_\varphi(P_T^X \| P_S^X) = \sup_{w>0}\{2 - J(w)\} = 2 - \inf_{w>0} J(w),$$

which proves the result. □

C Proof of Theorem 3.1

We first establish several auxiliary lemmas and then use them to prove the main theorem. Throughout this section, define

$$\Phi_\beta(x) = \frac{\phi(x)}{g_\beta(x)^2}.$$

C.1 Proofs of the Auxiliary Lemmas

Lemma C.1. *Under Assumption 3.1, there exist constants $0 < \kappa_0 < \kappa_1 < \infty$ such that*

$$\kappa_0\|\beta\|_1 \le g_\beta(x) \le \kappa_1\|\beta\|_1$$

for every $x \in \mathcal{X}$ and $\beta \in [0,\infty)^M$.

Moreover, there exist constants $0 < C_1 \le C_2 < \infty$, independent of n, such that

$$\widehat{\beta}_n, \beta^\star_\lambda \in \mathcal{B}_0,$$

where

$$\mathcal{B}_0 = \left\{\beta \in [0,\infty)^M : C_1 \le \|\beta\|_1 \le C_2\right\}.$$

Proof. By definition,

$$g_\beta(x) = \beta^\top\phi(x) = \sum_{m=1}^M \beta_m K(x, c_m).$$

Since $\beta_m \ge 0$ and

$$\kappa_0 \le K(x, c_m) \le \kappa_1$$

for every $x \in \mathcal{X}$ and $m = 1, \dots, M$, we have

$$\kappa_0 \sum_{m=1}^M \beta_m \le g_\beta(x) \le \kappa_1 \sum_{m=1}^M \beta_m.$$

Because $\beta \ge 0$,

$$\sum_{m=1}^M \beta_m = \|\beta\|_1,$$

and therefore

$$\kappa_0\|\beta\|_1 \le g_\beta(x) \le \kappa_1\|\beta\|_1.$$

Next, we prove the second part of the lemma. First note that $\beta = 0$ cannot be a minimizer: under the extended-real convention for the reciprocal term,

$$J_\lambda(0) = \widehat{J}_{n,\lambda}(0) = +\infty.$$

Thus, both minimizers necessarily satisfy $\|\beta\|_1 > 0$. Based on the population objective in equation 3, for any $\beta \in \mathcal{C}$,

$$\begin{aligned} J_\lambda(\beta) &= \mathbb{E}_{P_S^X}\left[g_\beta(X)\right] + \mathbb{E}_{P_T^X}\left[\frac{1}{g_\beta(X)}\right] + \frac{\lambda}{2}\|\beta\|_2^2 \\ &\ge \mathbb{E}_{P_S^X}\left[g_\beta(X)\right] + \mathbb{E}_{P_T^X}\left[\frac{1}{g_\beta(X)}\right] \\ &\ge \kappa_0\|\beta\|_1 + \frac{1}{\kappa_1\|\beta\|_1}. \end{aligned} \tag{9}$$

Similarly, we obtain the upper bound

$$J_\lambda(\beta) \le \kappa_1\|\beta\|_1 + \frac{1}{\kappa_0\|\beta\|_1} + \frac{\lambda}{2}\|\beta\|_2^2. \tag{10}$$

Now let $\widetilde{\beta} \in \mathcal{C}$ be an arbitrary fixed reference point and define

$$\widetilde{C} = \kappa_1 \|\widetilde{\beta}\|_1 + \frac{1}{\kappa_0 \|\widetilde{\beta}\|_1} + \frac{\lambda}{2}\|\widetilde{\beta}\|_2^2.$$

By equation 10 we have $J_\lambda(\widetilde{\beta}) \le \widetilde{C}$. Set

$$C_1 = \frac{1}{\widetilde{C}\kappa_1}, \qquad C_2 = \frac{\widetilde{C}}{\kappa_0},$$

and define

$$\mathcal{B}_0 = \left\{\beta \in [0,\infty)^M : C_1 \le \|\beta\|_1 \le C_2\right\}.$$

If $\|\beta\|_1 < C_1$, then equation 9 gives

$$J_\lambda(\beta) \ge \frac{1}{\kappa_1 \|\beta\|_1} > \widetilde{C},$$

whereas if $\|\beta\|_1 > C_2$, equation 9 gives

$$J_\lambda(\beta) \ge \kappa_0 \|\beta\|_1 > \widetilde{C}.$$

Consequently, the set

$$\left\{\beta \in \mathcal{C} : J_\lambda(\beta) \le \widetilde{C}\right\}$$

is nonempty and contained in $\mathcal{B}_0$. Since $\mathcal{B}_0$ is compact and J_λ is continuous on $\mathcal{B}_0$, the Weierstrass theorem implies that J_λ attains its global minimum at some $\beta_\lambda^\star \in \mathcal{B}_0$.

For the empirical objective, the analogous upper and lower bounds give

$$\begin{aligned}\widehat{J}_{n,\lambda}(\beta) &= \frac{1}{n_S}\sum_{i=1}^{n_S} g_\beta(X_i^S) + \frac{1}{n_T}\sum_{j=1}^{n_T}\frac{1}{g_\beta(X_j^T)} + \frac{\lambda}{2}\|\beta\|_2^2 \\ &\ge \kappa_0\|\beta\|_1 + \frac{1}{\kappa_1\|\beta\|_1}.\end{aligned}$$

Moreover,

$$\widehat{J}_{n,\lambda}(\widetilde{\beta}) \le \kappa_1\|\widetilde{\beta}\|_1 + \frac{1}{\kappa_0\|\widetilde{\beta}\|_1} + \frac{\lambda}{2}\|\widetilde{\beta}\|_2^2 = \widetilde{C}.$$

Thus,

$$\left\{\beta \in \mathcal{C} : \widehat{J}_{n,\lambda}(\beta) \le \widetilde{C}\right\} \subseteq \mathcal{B}_0.$$

By continuity and compactness, $\widehat{J}_{n,\lambda}$ attains its global minimum at some $\widehat{\beta}_n \in \mathcal{B}_0$.

The constants C_1 and C_2 depend only on the fixed reference point $\widetilde{\beta}$, κ_0, κ_1, and λ, and are therefore independent of n.

□

Lemma C.2. *Under Assumption 3.1 and $\lambda > 0$, for every $\beta \in \mathcal{C}$,*

$$\nabla^2 \widehat{J}_{n,\lambda}(\beta) = \frac{2}{n_T}\sum_{j=1}^{n_T}\frac{\phi(X_j^T)\phi(X_j^T)^\top}{g_\beta(X_j^T)^3} + \lambda I_M \succeq \lambda I_M.$$

Similarly,

$$\nabla^2 J_\lambda(\beta) = 2\mathbb{E}_{X\sim P_T^X}\left[\frac{\phi(X)\phi(X)^\top}{g_\beta(X)^3}\right] + \lambda I_M \succeq \lambda I_M.$$

Therefore both objectives are λ-strongly convex and have unique global minimizers in $\mathcal{B}_0$.

Proof. Since $g_\beta(x) = \beta^\top \phi(x)$,

$$\nabla_\beta g_\beta(x) = \phi(x), \qquad \nabla^2_\beta g_\beta(x) = 0,$$

and

$$\nabla^2_\beta \left\{ g_\beta(x)^{-1} \right\} = 2\frac{\phi(x)\phi(x)^\top}{g_\beta(x)^3}.$$

Therefore,

$$\begin{aligned}\nabla^2 \widehat{J}_{n,\lambda}(\beta) &= \frac{1}{n_S}\sum_{i=1}^{n_S} \nabla^2_\beta g_\beta(X_i^S) + \frac{1}{n_T}\sum_{j=1}^{n_T} \nabla^2_\beta \{ g_\beta(X_j^T)^{-1} \} + \nabla^2_\beta \left(\frac{\lambda}{2}\|\beta\|_2^2 \right) \\ &= \frac{2}{n_T}\sum_{j=1}^{n_T} \frac{\phi(X_j^T)\phi(X_j^T)^\top}{g_\beta(X_j^T)^3} + \lambda I_M.\end{aligned}$$

For every $v \in \mathbb{R}^M$,

$$v^\top \frac{\phi(x)\phi(x)^\top}{g_\beta(x)^3} v = \frac{\{v^\top \phi(x)\}^2}{g_\beta(x)^3} \geq 0,$$

because $g_\beta(x) > 0$ on $\mathcal{C}$ following Lemma C.1. Hence

$$\nabla^2 \widehat{J}_{n,\lambda}(\beta) \succeq \lambda I_M.$$

For the population objective, the same calculation applies. The uniform bounds established in Lemma C.1 justify differentiation under the expectation by the dominated convergence theorem, yielding

$$\nabla^2 J_\lambda(\beta) = 2\mathbb{E}_{X \sim P_T^X}\left[\frac{\phi(X)\phi(X)^\top}{g_\beta(X)^3} \right] + \lambda I_M \succeq \lambda I_M.$$

Therefore both objectives are λ-strongly convex on the convex set $\mathcal{C}$. Existence of their minimizers in $\mathcal{B}_0$ follows from Lemma C.1, and strong convexity implies that each minimizer is unique. □

Lemma C.3. *The empirical and population minimizers satisfy*

$$\|\widehat{\beta}_n - \beta^\star_\lambda\|_2 \leq \frac{1}{\lambda}\left\| \nabla \widehat{J}_{n,\lambda}(\beta^\star_\lambda) - \nabla J_\lambda(\beta^\star_\lambda) \right\|_2.$$

Consequently,

$$\|\widehat{\beta}_n - \beta^\star_\lambda\|_2 \leq \frac{1}{\lambda}\sup_{\beta \in \mathcal{B}_0}\left\| \nabla \widehat{J}_{n,\lambda}(\beta) - \nabla J_\lambda(\beta) \right\|_2.$$

These bounds remain valid when either minimizer lies on the boundary of $\mathcal{C}$.

Proof. Let $d_n = \widehat{\beta}_n - \beta^\star_\lambda$. Since J_λ and $\widehat{J}_{n,\lambda}$ are λ-strongly convex and $\mathcal{C}$ is a convex set, $\langle \nabla J_\lambda(\beta^\star_\lambda), \beta - \beta^\star_\lambda \rangle \geq 0$ and $\left\langle \nabla \widehat{J}_{n,\lambda}(\widehat{\beta}_n), \beta - \widehat{\beta}_n \right\rangle \geq 0$ for all $\beta \in \mathcal{C}$ (cf. Boyd & Vandenberghe, 2004, §4.2.3).

Applying these conditions to $\widehat{J}_{n,\lambda}$ with $\beta = \beta^\star_\lambda$, and to J_λ with $\beta = \widehat{\beta}_n$ gives

$$\left\langle \nabla \widehat{J}_{n,\lambda}(\widehat{\beta}_n), d_n \right\rangle \leq 0, \qquad \langle \nabla J_\lambda(\beta^\star_\lambda), d_n \rangle \geq 0,$$

and combining these inequalities,

$$\begin{aligned}\lambda\|d_n\|_2^2 &\leq \left\langle \nabla \widehat{J}_{n,\lambda}(\widehat{\beta}_n) - \nabla \widehat{J}_{n,\lambda}(\beta^\star_\lambda), d_n \right\rangle \\ &\leq -\left\langle \nabla \widehat{J}_{n,\lambda}(\beta^\star_\lambda), d_n \right\rangle \\ &= -\langle \nabla J_\lambda(\beta^\star_\lambda), d_n \rangle - \left\langle \nabla \widehat{J}_{n,\lambda}(\beta^\star_\lambda) - \nabla J_\lambda(\beta^\star_\lambda), d_n \right\rangle \\ &\leq \left\| \nabla \widehat{J}_{n,\lambda}(\beta^\star_\lambda) - \nabla J_\lambda(\beta^\star_\lambda) \right\|_2 \|d_n\|_2,\end{aligned}$$

where the last inequality follows from the population first-order condition and the Cauchy-Schwarz inequality. If $d_n = 0$, the first claim is obvious. Otherwise, dividing by $\lambda\|d_n\|_2$ yields

$$\|\widehat{\beta}_n - \beta^\star_\lambda\|_2 \le \frac{1}{\lambda}\left\|\nabla \widehat{J}_{n,\lambda}(\beta^\star_\lambda) - \nabla J_\lambda(\beta^\star_\lambda)\right\|_2.$$

Finally, Lemma C.1 gives $\beta^\star_\lambda \in \mathcal{B}_0$, so

$$\left\|\nabla \widehat{J}_{n,\lambda}(\beta^\star_\lambda) - \nabla J_\lambda(\beta^\star_\lambda)\right\|_2 \le \sup_{\beta\in\mathcal{B}_0}\left\|\nabla \widehat{J}_{n,\lambda}(\beta) - \nabla J_\lambda(\beta)\right\|_2.$$

□

Lemma C.4. *Under Assumption 3.1, there exist finite constants $C_\Phi, L_\Phi > 0$ such that, uniformly over $x \in \mathcal{X}$ and $\beta, \beta' \in \mathcal{B}_0$,*

$$\|\Phi_\beta(x)\|_2 \le C_\Phi$$

and

$$\|\Phi_\beta(x) - \Phi_{\beta'}(x)\|_2 \le L_\Phi\|\beta - \beta'\|_2.$$

Proof. By definition,

$$\Phi_\beta(x) = \frac{\phi(x)}{g_\beta(x)^2}.$$

Thus,

$$\begin{aligned}\|\Phi_\beta(x)\|_2 &= \frac{1}{g_\beta(x)^2}\|\phi(x)\|_2 \\ &= \frac{\left(\sum_{m=1}^M K(x,c_m)^2\right)^{1/2}}{g_\beta(x)^2}.\end{aligned}$$

Assumption 3.1 gives

$$\|\phi(x)\|_2^2 = \sum_{m=1}^M K(x,c_m)^2 \le M\kappa_1^2,$$

and hence $\|\phi(x)\|_2 \le \sqrt{M}\kappa_1$, while Lemma C.1 gives $g_\beta(x) \ge \kappa_0 C_1$ for every $\beta \in \mathcal{B}_0$. Therefore,

$$\|\Phi_\beta(x)\|_2 \le \frac{\sqrt{M}\kappa_1}{(\kappa_0 C_1)^2} =: C_\Phi.$$

For the Lipschitz bound, differentiating $\Phi_\beta(x)$ with respect to β gives

$$D_\beta\Phi_\beta(x) = -2\frac{\phi(x)\phi(x)^\top}{g_\beta(x)^3}.$$

We first verify that $\|\phi(x)\phi(x)^\top\|_{\mathrm{op}} = \|\phi(x)\|_2^2$. For any $\nu \in \mathbb{R}^M$, the Cauchy-Schwarz inequality gives

$$\begin{aligned}\|\phi(x)\phi(x)^\top\nu\|_2 &= |\phi(x)^\top\nu|\,\|\phi(x)\|_2 \\ &\le \|\phi(x)\|_2^2\|\nu\|_2.\end{aligned}$$

Taking the supremum over $\|\nu\|_2 = 1$ yields $\|\phi(x)\phi(x)^\top\|_{\mathrm{op}} \le \|\phi(x)\|_2^2$. Conversely, since $\phi(x) \ne 0$, choosing $\nu = \phi(x)/\|\phi(x)\|_2$ gives $\|\nu\|_2 = 1$ and

$$\|\phi(x)\phi(x)^\top\nu\|_2 = \|\phi(x)\|_2^2.$$

Therefore, $\|\phi(x)\phi(x)^\top\|_{\mathrm{op}} = \|\phi(x)\|_2^2$. Substituting this identity into the expression for the Jacobian and using $\|\phi(x)\|_2^2 \le M\kappa_1^2$ and $g_\beta(x) \ge \kappa_0 C_1$, we obtain

$$\|D_\beta\Phi_\beta(x)\|_{\mathrm{op}} = 2\frac{\|\phi(x)\|_2^2}{g_\beta(x)^3} \le \frac{2M\kappa_1^2}{(\kappa_0 C_1)^3} =: L_\Phi.$$

The set $\mathcal{B}_0$ is convex. For any $\beta, \beta' \in \mathcal{B}_0$ and $t \in [0,1]$, the vector $\beta_t = (1-t)\beta' + t\beta$ is nonnegative and satisfies

$$\|\beta_t\|_1 = (1-t)\|\beta'\|_1 + t\|\beta\|_1 \in [C_1, C_2].$$

Therefore, $\beta_t \in \mathcal{B}_0$. Applying the fundamental theorem of calculus, we get

$$\Phi_\beta(x) - \Phi_{\beta'}(x) = \int_0^1 D_\beta \Phi_{\beta_t}(x)(\beta - \beta')\, dt.$$

Hence

$$\|\Phi_\beta(x) - \Phi_{\beta'}(x)\|_2 \le \int_0^1 \|D_\beta \Phi_{\beta_t}(x)\|_{\mathrm{op}} \|\beta - \beta'\|_2\, dt \le L_\Phi \|\beta - \beta'\|_2.$$

Both constants are finite, positive, and independent of x, β, β', which proves the result. □

Lemma C.5. *Under Assumptions 3.1–3.3,*

$$\sup_{\beta \in \mathcal{B}_0} \left\| \nabla \widehat{J}_{n,\lambda}(\beta) - \nabla J_\lambda(\beta) \right\|_2 \xrightarrow{p} 0.$$

Proof. Let $P_{n,S}$ and $P_{n,T}$ denote the empirical source and target measures, and let P_S and P_T denote their population counterparts. For any integrable vector-valued function f, write

$$P_{n,S} f \coloneqq \frac{1}{n_S} \sum_{i=1}^{n_S} f(X_i^S), \qquad P_{n,T} f \coloneqq \frac{1}{n_T} \sum_{j=1}^{n_T} f(X_j^T),$$

and

$$P_S f \coloneqq \mathbb{E}_{X \sim P_S^X}[f(X)], \qquad P_T f \coloneqq \mathbb{E}_{X \sim P_T^X}[f(X)].$$

Since $\nabla_\beta g_\beta(x) = \phi(x)$ and

$$\nabla_\beta \left\{ \frac{1}{g_\beta(x)} \right\} = -\frac{\phi(x)}{g_\beta(x)^2} = -\Phi_\beta(x),$$

differentiating the empirical and population objectives gives

$$\nabla \widehat{J}_{n,\lambda}(\beta) = \frac{1}{n_S} \sum_{i=1}^{n_S} \phi(X_i^S) - \frac{1}{n_T} \sum_{j=1}^{n_T} \Phi_\beta(X_j^T) + \lambda\beta$$

and

$$\nabla J_\lambda(\beta) = \mathbb{E}_{P_S^X}[\phi(X)] - \mathbb{E}_{P_T^X}[\Phi_\beta(X)] + \lambda\beta.$$

Equivalently,

$$\nabla \widehat{J}_{n,\lambda}(\beta) = P_{n,S}\phi - P_{n,T}\Phi_\beta + \lambda\beta, \qquad \nabla J_\lambda(\beta) = P_S\phi - P_T\Phi_\beta + \lambda\beta.$$

The regularization terms cancel when the two gradients are subtracted. Therefore, the triangle inequality gives

$$\sup_{\beta \in \mathcal{B}_0} \left\| \nabla \widehat{J}_{n,\lambda}(\beta) - \nabla J_\lambda(\beta) \right\|_2 \le \|(P_{n,S} - P_S)\phi\|_2 + \sup_{\beta \in \mathcal{B}_0} \|(P_{n,T} - P_T)\Phi_\beta\|_2 .$$

We first control the source term. Let $\mu_S = P_S\phi = \mathbb{E}_{P_S^X}[\phi(X)]$ and $Z_i = \phi(X_i^S) - \mu_S$. Then

$$(P_{n,S} - P_S)\phi = \frac{1}{n_S} \sum_{i=1}^{n_S} Z_i.$$

The random vectors $Z_1, \ldots, Z_{n_S}$ are independent and satisfy $\mathbb{E}[Z_i] = 0$. Hence, for $i \neq \ell$,

$$\mathbb{E}[Z_i^\top Z_\ell] = \mathbb{E}[Z_i]^\top \mathbb{E}[Z_\ell] = 0.$$

Expanding the squared norm therefore gives

$$\begin{aligned} \mathbb{E}\left[\|(P_{n,S} - P_S)\phi\|_2^2 \right] &= \frac{1}{n_S^2} \sum_{i=1}^{n_S} \sum_{\ell=1}^{n_S} \mathbb{E}[Z_i^\top Z_\ell] \\ &= \frac{1}{n_S} \mathbb{E}[\|Z_1\|_2^2]. \end{aligned}$$

Expanding $\|\phi(X) - \mu_S\|_2^2$ and using $\mathbb{E}_{P_S^X}[\phi(X)] = \mu_S$ yields

$$\mathbb{E}[\|Z_1\|_2^2] = \mathbb{E}_{P_S^X}[\|\phi(X)\|_2^2] - \|\mu_S\|_2^2 \leq \mathbb{E}_{P_S^X}[\|\phi(X)\|_2^2].$$

Assumption 3.1 gives

$$\|\phi(x)\|_2^2 = \sum_{m=1}^{M} K(x, c_m)^2 \leq M\kappa_1^2,$$

and consequently

$$\mathbb{E}\left[\|(P_{n,S} - P_S)\phi\|_2^2\right] \leq \frac{M\kappa_1^2}{n_S}.$$

Applying Markov's inequality to the nonnegative random variable $\|(P_{n,S}-P_S)\phi\|_2^2$, for every $\varepsilon > 0$,

$$\mathbb{P}\left(\|(P_{n,S} - P_S)\phi\|_2 > \varepsilon\right) \leq \frac{\mathbb{E}[\|(P_{n,S} - P_S)\phi\|_2^2]}{\varepsilon^2} \leq \frac{M\kappa_1^2}{n_S\varepsilon^2} \longrightarrow 0.$$

Thus,

$$\|(P_{n,S} - P_S)\phi\|_2 \xrightarrow{p} 0.$$

We next control the target term uniformly over $\mathcal{B}_0$. Fix $\varepsilon > 0$ and set $\delta = \varepsilon/(4L_\Phi)$. Since $\mathcal{B}_0$ is compact, there exist $\beta_1, \ldots, \beta_N \in \mathcal{B}_0$ such that

$$\mathcal{B}_0 \subseteq \bigcup_{k=1}^{N} B(\beta_k, \delta),$$

where $B(\beta, \delta) \coloneqq \left\{\beta' \in \mathbb{R}^M : \|\beta' - \beta\|_2 < \delta\right\}$ is the open Euclidean δ-ball centered at β. For each $\beta \in \mathcal{B}_0$, choose k such that $\|\beta - \beta_k\|_2 \leq \delta$. We can decompose

$$(P_{n,T} - P_T)\Phi_\beta = (P_{n,T} - P_T)\Phi_{\beta_k} + P_{n,T}(\Phi_\beta - \Phi_{\beta_k}) - P_T(\Phi_\beta - \Phi_{\beta_k}).$$

Using the triangle inequality for empirical averages and Jensen's inequality for expectations, we have

$$\begin{aligned}
\|(P_{n,T} - P_T)\Phi_\beta\|_2 &\leq \|(P_{n,T} - P_T)\Phi_{\beta_k}\|_2 + P_{n,T}\|\Phi_\beta - \Phi_{\beta_k}\|_2 + P_T\|\Phi_\beta - \Phi_{\beta_k}\|_2 \\
&\leq \|(P_{n,T} - P_T)\Phi_{\beta_k}\|_2 + 2L_\Phi\|\beta - \beta_k\|_2 \\
&\leq \|(P_{n,T} - P_T)\Phi_{\beta_k}\|_2 + \frac{\varepsilon}{2},
\end{aligned}$$

where the second inequality follows from Lemma C.4. Taking the supremum over $\beta \in \mathcal{B}_0$ yields

$$\sup_{\beta \in \mathcal{B}_0} \|(P_{n,T} - P_T)\Phi_\beta\|_2 \leq \max_{1 \leq k \leq N} \|(P_{n,T} - P_T)\Phi_{\beta_k}\|_2 + \frac{\varepsilon}{2}.$$

It remains to control the finitely many terms on the right. For each k, let $\mu_{T,k} = P_T\Phi_{\beta_k}$. By the same centered-sum calculation used for the source term,

$$\mathbb{E}\left[\|(P_{n,T} - P_T)\Phi_{\beta_k}\|_2^2\right] = \frac{1}{n_T}\mathbb{E}_{P_T^X}\left[\|\Phi_{\beta_k}(X) - \mu_{T,k}\|_2^2\right] \leq \frac{1}{n_T}\mathbb{E}_{P_T^X}\left[\|\Phi_{\beta_k}(X)\|_2^2\right].$$

Lemma C.4 gives $\|\Phi_{\beta_k}(x)\|_2 \leq C_\Phi$, so

$$\mathbb{E}\left[\|(P_{n,T} - P_T)\Phi_{\beta_k}\|_2^2\right] \leq \frac{C_\Phi^2}{n_T}.$$

Markov's inequality and the union bound now give

$$\begin{aligned}
\mathbb{P}\left(\sup_{\beta \in \mathcal{B}_0} \|(P_{n,T} - P_T)\Phi_\beta\|_2 > \varepsilon\right) &\leq \mathbb{P}\left(\max_{1 \leq k \leq N} \|(P_{n,T} - P_T)\Phi_{\beta_k}\|_2 > \frac{\varepsilon}{2}\right) \\
&\leq \sum_{k=1}^{N} \mathbb{P}\left(\|(P_{n,T} - P_T)\Phi_{\beta_k}\|_2 > \frac{\varepsilon}{2}\right) \\
&\leq \frac{4NC_\Phi^2}{n_T\varepsilon^2} \longrightarrow 0.
\end{aligned}$$

Here N depends on the fixed value of ε, but not on n_T. Therefore,

$$\sup_{\beta \in \mathcal{B}_0} \|(P_{n,T} - P_T)\Phi_\beta\|_2 \xrightarrow{p} 0.$$

Both terms in the initial upper bound converge to zero in probability. It follows that

$$\sup_{\beta \in \mathcal{B}_0} \left\| \nabla \widehat{J}_{n,\lambda}(\beta) - \nabla J_\lambda(\beta) \right\|_2 \xrightarrow{p} 0.$$

□

Lemma C.6. *Under Assumptions 3.1–3.3,*

$$\left\| \nabla \widehat{J}_{n,\lambda}(\beta^\star_\lambda) - \nabla J_\lambda(\beta^\star_\lambda) \right\|_2 = O_p\left(n^{-1/2}\right).$$

Proof. Recall that

$$\nabla \widehat{J}_{n,\lambda}(\beta) = \frac{1}{n_S}\sum_{i=1}^{n_S} \phi(X_i^S) - \frac{1}{n_T}\sum_{j=1}^{n_T} \Phi_\beta(X_j^T) + \lambda\beta,$$

where

$$\Phi_\beta(x) = \frac{\phi(x)}{g_\beta(x)^2}.$$

Similarly,

$$\nabla J_\lambda(\beta) = \mathbb{E}_{P_S^X}[\phi(X)] - \mathbb{E}_{P_T^X}[\Phi_\beta(X)] + \lambda\beta.$$

Therefore, at the fixed population minimizer $\beta^\star_\lambda$,

$$\begin{aligned} &\nabla \widehat{J}_{n,\lambda}(\beta^\star_\lambda) - \nabla J_\lambda(\beta^\star_\lambda) \\ &\quad = \left\{ \frac{1}{n_S}\sum_{i=1}^{n_S} \phi(X_i^S) - \mathbb{E}_{P_S^X}[\phi(X)] \right\} \\ &\qquad - \left\{ \frac{1}{n_T}\sum_{j=1}^{n_T} \Phi_{\beta^\star_\lambda}(X_j^T) - \mathbb{E}_{P_T^X}[\Phi_{\beta^\star_\lambda}(X)] \right\}. \end{aligned} \tag{11}$$

By Assumption 3.1, for every $m = 1, \ldots, M$,

$$0 < \kappa_0 \le \phi_m(x) \le \kappa_1.$$

Moreover, Lemma C.1 implies

$$g_{\beta^\star_\lambda}(x) \ge \kappa_0 \|\beta^\star_\lambda\|_1 \ge \kappa_0 C_1.$$

Consequently,

$$0 \le \Phi_{\beta^\star_\lambda, m}(x) = \frac{\phi_m(x)}{g_{\beta^\star_\lambda}(x)^2} \le \frac{\kappa_1}{(\kappa_0 C_1)^2} =: B_\Phi$$

uniformly over $x \in \mathcal{X}$ and $m = 1, \ldots, M$.

For the source term, Hoeffding's inequality gives, for every $t > 0$,

$$\mathbb{P}\left(\left| \frac{1}{n_S}\sum_{i=1}^{n_S} \phi_m(X_i^S) - \mathbb{E}_{P_S^X}[\phi_m(X)] \right| > t \right) \le 2\exp\left(-\frac{2n_S t^2}{(\kappa_1 - \kappa_0)^2} \right).$$

Applying the union bound over $m = 1, \ldots, M$,

$$\mathbb{P}\left(\max_{1 \le m \le M} \left| \frac{1}{n_S}\sum_{i=1}^{n_S} \phi_m(X_i^S) - \mathbb{E}_{P_S^X}[\phi_m(X)] \right| > t \right) \le 2M\exp\left(-\frac{2n_S t^2}{(\kappa_1 - \kappa_0)^2} \right). \tag{12}$$

Likewise, since $0 \le \Phi_{\beta^\star_\lambda,m}(X) \le B_\Phi$,

$$\mathbb{P}\left(\max_{1\le m\le M}\left|\frac{1}{n_T}\sum_{j=1}^{n_T}\Phi_{\beta^\star_\lambda,m}(X_j^T) - \mathbb{E}_{P_T^X}[\Phi_{\beta^\star_\lambda,m}(X)]\right| > t\right) \le 2M\exp\left(-\frac{2n_T t^2}{B_\Phi^2}\right). \tag{13}$$

Since

$$\frac{n_S}{n} \longrightarrow \rho \in (0,1),$$

we also have

$$\frac{n_T}{n} = 1 - \frac{n_S}{n} \longrightarrow 1-\rho \in (0,1).$$

Hence, taking

$$c_\rho = \frac{1}{2}\min\{\rho, 1-\rho\} > 0,$$

there exists $N < \infty$ such that, for all $n \ge N$,

$$n_S \ge c_\rho n, \qquad n_T \ge c_\rho n.$$

Define

$$S_n = \frac{1}{n_S}\sum_{i=1}^{n_S}\phi(X_i^S) - \mathbb{E}_{P_S^X}[\phi(X)]$$

and

$$T_n = \frac{1}{n_T}\sum_{j=1}^{n_T}\Phi_{\beta^\star_\lambda}(X_j^T) - \mathbb{E}_{P_T^X}[\Phi_{\beta^\star_\lambda}(X)].$$

Then, by equation 11,

$$\nabla\widehat{J}_{n,\lambda}(\beta^\star_\lambda) - \nabla J_\lambda(\beta^\star_\lambda) = S_n - T_n.$$

We first control the source term. Let

$$t_n = \frac{C}{\sqrt{n}}, \qquad C > 0.$$

Substituting $t = t_n$ into equation 12 and using $n_S \ge c_\rho n$, we obtain, for all sufficiently large n,

$$\begin{aligned}\mathbb{P}\left(\|S_n\|_\infty > \frac{C}{\sqrt{n}}\right) &\le 2M\exp\left(-\frac{2n_S C^2}{n(\kappa_1-\kappa_0)^2}\right)\\ &\le 2M\exp\left(-\frac{2c_\rho C^2}{(\kappa_1-\kappa_0)^2}\right).\end{aligned}$$

The right-hand side does not depend on n and converges to zero as $C \to \infty$. Hence, for every $\varepsilon > 0$, there exists $C_\varepsilon < \infty$ such that

$$\sup_{n\ge N}\mathbb{P}\left(\sqrt{n}\,\|S_n\|_\infty > C_\varepsilon\right) < \varepsilon.$$

Therefore,

$$\|S_n\|_\infty = O_p(n^{-1/2}).$$

Similarly, substituting $t = t_n$ into equation 13 and using $n_T \ge c_\rho n$ gives, for all sufficiently large n,

$$\begin{aligned}\mathbb{P}\left(\|T_n\|_\infty > \frac{C}{\sqrt{n}}\right) &\le 2M\exp\left(-\frac{2n_T C^2}{nB_\Phi^2}\right)\\ &\le 2M\exp\left(-\frac{2c_\rho C^2}{B_\Phi^2}\right).\end{aligned}$$

Again, the right-hand side converges to zero as $C \to \infty$, uniformly for all sufficiently large n. Thus,

$$\|T_n\|_\infty = O_p(n^{-1/2}).$$

Now, by the triangle inequality,

$$\|S_n - T_n\|_\infty \le \|S_n\|_\infty + \|T_n\|_\infty.$$

More explicitly, for every $C > 0$ and all sufficiently large n,

$$\begin{aligned}\mathbb{P}\left(\|S_n - T_n\|_\infty > \frac{2C}{\sqrt{n}}\right) &\le \mathbb{P}\left(\|S_n\|_\infty > \frac{C}{\sqrt{n}}\right) + \mathbb{P}\left(\|T_n\|_\infty > \frac{C}{\sqrt{n}}\right) \\ &\le 2M \exp\left(-\frac{2c_\rho C^2}{(\kappa_1 - \kappa_0)^2}\right) + 2M \exp\left(-\frac{2c_\rho C^2}{B_\Phi^2}\right).\end{aligned}$$

The right-hand side can be made arbitrarily small by choosing C sufficiently large. Therefore,

$$\left\|\nabla \widehat{J}_{n,\lambda}(\beta^\star_\lambda) - \nabla J_\lambda(\beta^\star_\lambda)\right\|_\infty = O_p(n^{-1/2}).$$

Finally, for every $v \in \mathbb{R}^M$,

$$\|v\|_2 \le \sqrt{M}\,\|v\|_\infty.$$

Since M is fixed under Assumption 3.3,

$$\sqrt{M} = O(1).$$

Applying this inequality to the score difference yields

$$\begin{aligned}\left\|\nabla \widehat{J}_{n,\lambda}(\beta^\star_\lambda) - \nabla J_\lambda(\beta^\star_\lambda)\right\|_2 &\le \sqrt{M}\left\|\nabla \widehat{J}_{n,\lambda}(\beta^\star_\lambda) - \nabla J_\lambda(\beta^\star_\lambda)\right\|_\infty \\ &= O_p(n^{-1/2}).\end{aligned}$$

□

Lemma C.7. *Under Assumptions 3.1–3.3,*

$$\sup_{\beta \in \mathcal{B}_0}\left\|\nabla^2 \widehat{J}_{n,\lambda}(\beta) - \nabla^2 J_\lambda(\beta)\right\|_{\mathrm{op}} \xrightarrow{p} 0.$$

Proof. Let

$$H_\beta(x) = \frac{\phi(x)\phi(x)^\top}{g_\beta(x)^3}.$$

We first show that the family $\{H_\beta : \beta \in \mathcal{B}_0\}$ is uniformly bounded in operator norm and Lipschitz in β. By Assumption 3.1 and Lemma C.1,

$$\|\phi(x)\|_2^2 \le M\kappa_1^2, \qquad g_\beta(x) \ge \kappa_0 C_1$$

uniformly over $x \in \mathcal{X}$ and $\beta \in \mathcal{B}_0$. As shown in the proof of Lemma C.4, $\|\phi(x)\phi(x)^\top\|_{\mathrm{op}} = \|\phi(x)\|_2^2$. Therefore,

$$\|H_\beta(x)\|_{\mathrm{op}} = \frac{\|\phi(x)\phi(x)^\top\|_{\mathrm{op}}}{g_\beta(x)^3} = \frac{\|\phi(x)\|_2^2}{g_\beta(x)^3} \le \frac{M\kappa_1^2}{(\kappa_0 C_1)^3} =: C_H.$$

For $\beta, \beta' \in \mathcal{B}_0$,

$$\|H_\beta(x) - H_{\beta'}(x)\|_{\mathrm{op}} = \|\phi(x)\phi(x)^\top\|_{\mathrm{op}}\left|\frac{1}{g_\beta(x)^3} - \frac{1}{g_{\beta'}(x)^3}\right|.$$

The mean value theorem gives

$$\left|\frac{1}{g_\beta(x)^3} - \frac{1}{g_{\beta'}(x)^3}\right| = \frac{3}{\xi^4}|g_\beta(x) - g_{\beta'}(x)|$$

for some ξ between $g_\beta(x)$ and $g_{\beta'}(x)$. Since both quantities are at least $\kappa_0 C_1$, we have $\xi \ge \kappa_0 C_1$. Moreover, the Cauchy-Schwarz inequality gives

$$|g_\beta(x) - g_{\beta'}(x)| = \left|(\beta - \beta')^\top \phi(x)\right| \le \|\beta - \beta'\|_2 \|\phi(x)\|_2 \le \sqrt{M}\kappa_1 \|\beta - \beta'\|_2.$$

Consequently,

$$\left|\frac{1}{g_\beta(x)^3}-\frac{1}{g_{\beta'}(x)^3}\right|\le\frac{3\sqrt{M}\kappa_1}{(\kappa_0 C_1)^4}\|\beta-\beta'\|_2,$$

and hence

$$\|H_\beta(x)-H_{\beta'}(x)\|_{\mathrm{op}}\le\frac{3M^{3/2}\kappa_1^3}{(\kappa_0 C_1)^4}\|\beta-\beta'\|_2=:L_H\|\beta-\beta'\|_2.$$

Note that $C_H>0$ and $L_H>0$. We are now ready to prove the uniform convergence. Using the empirical-measure notation introduced in the proof of Lemma C.5,

$$P_{n,T}H_\beta=\frac{1}{n_T}\sum_{j=1}^{n_T}H_\beta(X_j^T),\qquad P_TH_\beta=\mathbb{E}_{P_T^X}[H_\beta(X)].$$

Because the source term in the objective is linear in β,

$$\nabla^2\widehat{J}_{n,\lambda}(\beta)=2P_{n,T}H_\beta+\lambda I_M,\qquad\nabla^2J_\lambda(\beta)=2P_TH_\beta+\lambda I_M.$$

The regularization terms cancel each other, so we get

$$\nabla^2\widehat{J}_{n,\lambda}(\beta)-\nabla^2J_\lambda(\beta)=2(P_{n,T}-P_T)H_\beta.$$

It is thus enough to prove

$$\sup_{\beta\in\mathcal{B}_0}\|(P_{n,T}-P_T)H_\beta\|_{\mathrm{op}}\xrightarrow{p}0.$$

Fix $\varepsilon>0$ and set $\delta=\varepsilon/(4L_H)$. Since $\mathcal{B}_0$ is compact, there exist $\beta_1,\dots,\beta_N\in\mathcal{B}_0$ such that

$$\mathcal{B}_0\subseteq\bigcup_{k=1}^{N}B(\beta_k,\delta),$$

where $B(\beta_k,\delta)\coloneqq\{\beta\in\mathbb{R}^M:\|\beta-\beta_k\|_2<\delta\}$. For each $\beta\in\mathcal{B}_0$, choose k such that $\|\beta-\beta_k\|_2\le\delta$. We can then decompose

$$(P_{n,T}-P_T)H_\beta=(P_{n,T}-P_T)H_{\beta_k}+P_{n,T}(H_\beta-H_{\beta_k})-P_T(H_\beta-H_{\beta_k}).$$

For any matrix-valued function A, the triangle inequality for empirical averages and Jensen's inequality for expectations give

$$\|P_{n,T}A\|_{\mathrm{op}}\le P_{n,T}\|A\|_{\mathrm{op}},\qquad\|P_TA\|_{\mathrm{op}}\le P_T\|A\|_{\mathrm{op}}.$$

Applying these inequalities and the Lipschitz bound established above, we get

$$\begin{aligned}\|(P_{n,T}-P_T)H_\beta\|_{\mathrm{op}}&\le\|(P_{n,T}-P_T)H_{\beta_k}\|_{\mathrm{op}}+P_{n,T}\|H_\beta-H_{\beta_k}\|_{\mathrm{op}}+P_T\|H_\beta-H_{\beta_k}\|_{\mathrm{op}}\\&\le\|(P_{n,T}-P_T)H_{\beta_k}\|_{\mathrm{op}}+2L_H\|\beta-\beta_k\|_2\\&\le\|(P_{n,T}-P_T)H_{\beta_k}\|_{\mathrm{op}}+\frac{\varepsilon}{2}.\end{aligned}$$

Taking the supremum over $\beta\in\mathcal{B}_0$, we have

$$\sup_{\beta\in\mathcal{B}_0}\|(P_{n,T}-P_T)H_\beta\|_{\mathrm{op}}\le\max_{1\le k\le N}\|(P_{n,T}-P_T)H_{\beta_k}\|_{\mathrm{op}}+\frac{\varepsilon}{2}.$$

It remains to control the finitely many terms on the right. Let $\|\cdot\|_F$ denote the Frobenius norm. For every $M\times M$ matrix A,

$$\|A\|_{\mathrm{op}}\le\|A\|_F\le\sqrt{M}\|A\|_{\mathrm{op}}.$$

For each k, define

$$\mu_{T,k}=P_TH_{\beta_k},\qquad Y_{j,k}=H_{\beta_k}(X_j^T)-\mu_{T,k}.$$

Then $Y_{1,k},\dots,Y_{n_T,k}$ are independent, identically distributed, mean-zero random matrices, and

$$(P_{n,T}-P_T)H_{\beta_k}=\frac{1}{n_T}\sum_{j=1}^{n_T}Y_{j,k}.$$

Writing $\langle A, B\rangle_F = \mathrm{tr}(A^\top B)$, independence and mean zero imply that $\mathbb{E}[\langle Y_{j,k}, Y_{\ell,k}\rangle_F] = 0$ whenever $j \neq \ell$. Therefore,

$$\begin{aligned}\mathbb{E}\left[\|(P_{n,T} - P_T)H_{\beta_k}\|_F^2\right] &= \frac{1}{n_T^2}\sum_{j=1}^{n_T}\sum_{\ell=1}^{n_T}\mathbb{E}[\langle Y_{j,k}, Y_{\ell,k}\rangle_F] \\ &= \frac{1}{n_T}\mathbb{E}[\|Y_{1,k}\|_F^2].\end{aligned}$$

Expanding the squared Frobenius norm and using $\mathbb{E}[Y_{1,k}] = 0$ gives

$$\mathbb{E}[\|Y_{1,k}\|_F^2] = \mathbb{E}_{P_T^X}[\|H_{\beta_k}(X)\|_F^2] - \|\mu_{T,k}\|_F^2 \leq \mathbb{E}_{P_T^X}[\|H_{\beta_k}(X)\|_F^2].$$

The uniform operator-norm bound and $\|A\|_F^2 \leq M\|A\|_{\mathrm{op}}^2$ imply

$$\mathbb{E}\left[\|(P_{n,T} - P_T)H_{\beta_k}\|_F^2\right] \leq \frac{MC_H^2}{n_T}.$$

Since the operator norm is bounded by the Frobenius norm, Markov's inequality gives

$$\mathbb{P}\left(\|(P_{n,T} - P_T)H_{\beta_k}\|_{\mathrm{op}} > \frac{\varepsilon}{2}\right) \leq \mathbb{P}\left(\|(P_{n,T} - P_T)H_{\beta_k}\|_F > \frac{\varepsilon}{2}\right) \leq \frac{4MC_H^2}{n_T\varepsilon^2}.$$

Using the finite cover and the union bound,

$$\begin{aligned}\mathbb{P}\left(\sup_{\beta\in\mathcal{B}_0}\|(P_{n,T} - P_T)H_\beta\|_{\mathrm{op}} > \varepsilon\right) &\leq \mathbb{P}\left(\max_{1\leq k\leq N}\|(P_{n,T} - P_T)H_{\beta_k}\|_{\mathrm{op}} > \frac{\varepsilon}{2}\right) \\ &\leq \sum_{k=1}^{N}\mathbb{P}\left(\|(P_{n,T} - P_T)H_{\beta_k}\|_{\mathrm{op}} > \frac{\varepsilon}{2}\right) \\ &\leq \frac{4NMC_H^2}{n_T\varepsilon^2} \longrightarrow 0.\end{aligned}$$

Here N depends on the fixed value of ε, but not on n_T. Thus,

$$\sup_{\beta\in\mathcal{B}_0}\|(P_{n,T} - P_T)H_\beta\|_{\mathrm{op}} \xrightarrow{p} 0.$$

Finally,

$$\sup_{\beta\in\mathcal{B}_0}\left\|\nabla^2\widehat{J}_{n,\lambda}(\beta) - \nabla^2 J_\lambda(\beta)\right\|_{\mathrm{op}} = 2\sup_{\beta\in\mathcal{B}_0}\|(P_{n,T} - P_T)H_\beta\|_{\mathrm{op}} \xrightarrow{p} 0.$$

□

C.2 Proof of the Main Theorem

Proof of Theorem 3.1. The first claim equation 4 follows immediately from Lemmas C.3 and C.5,

$$\|\widehat{\beta}_n - \beta_\lambda^\star\|_2 \leq \frac{1}{\lambda}\sup_{\beta\in\mathcal{B}_0}\|\nabla\widehat{J}_{n,\lambda}(\beta) - \nabla J_\lambda(\beta)\|_2 \xrightarrow{p} 0.$$

For the second claim equation 5, Assumption 3.1 gives $\sup_{x\in\mathcal{X}}\|\phi(x)\|_2 \leq \sqrt{M}\kappa_1$ and hence

$$\sup_{x\in\mathcal{X}}|\widehat{g}_n(x) - g_\lambda^\star(x)| = \sup_{x\in\mathcal{X}}|(\widehat{\beta}_n - \beta_\lambda^\star)^\top\phi(x)| \leq \sqrt{M}\kappa_1\|\widehat{\beta}_n - \beta_\lambda^\star\|_2 \xrightarrow{p} 0.$$

For the third claim equation 6, Lemma C.1 also gives $g_\beta(x) \leq \kappa_1 C_2$ uniformly over $\beta \in \mathcal{B}_0$ and $x \in \mathcal{X}$. Therefore,

$$\begin{aligned}\sup_{x\in\mathcal{X}}|\widehat{w}_n(x) - w_\lambda^\star(x)| &= \sup_{x\in\mathcal{X}}|\widehat{g}_n(x) - g_\lambda^\star(x)|\,|\widehat{g}_n(x) + g_\lambda^\star(x)| \\ &\leq 2\kappa_1 C_2\sup_{x\in\mathcal{X}}|\widehat{g}_n(x) - g_\lambda^\star(x)| \xrightarrow{p} 0.\end{aligned}$$

For the fourth claim equation 7, by Lemmas C.3 and C.6, we have

$$\|\widehat{\beta}_n - \beta^\star_\lambda\|_2 \le \frac{1}{\lambda}\left\|\nabla \widehat{J}_{n,\lambda}(\beta^\star_\lambda) - \nabla J_\lambda(\beta^\star_\lambda)\right\|_2 = O_p\left(\sqrt{1/n}\right).$$

It remains to prove asymptotic normality under the additional assumption that $\beta^\star_{\lambda,m} > 0$ for every $m = 1, \dots, M$. Define

$$b_{\min} \coloneqq \min_{1\le m\le M} \beta^\star_{\lambda,m} > 0.$$

If $\min_m \widehat{\beta}_{n,m} \le b_{\min}/2$, then for some m,

$$|\widehat{\beta}_{n,m} - \beta^\star_{\lambda,m}| \ge \frac{b_{\min}}{2}.$$

However, from the consistency in equation 4,

$$\begin{aligned}\mathbb{P}\left(\min_{1\le m\le M} \widehat{\beta}_{n,m} \le \frac{b_{\min}}{2}\right) &\le \mathbb{P}\left(\|\widehat{\beta}_n - \beta^\star_\lambda\|_\infty \ge \frac{b_{\min}}{2}\right)\\ &\le \mathbb{P}\left(\|\widehat{\beta}_n - \beta^\star_\lambda\|_2 \ge \frac{b_{\min}}{2}\right) \longrightarrow 0,\end{aligned}$$

and therefore $\widehat{\beta}_n$ is an interior point of $\mathcal{C}$ with probability tending to one. The population minimizer $\beta^\star_\lambda$ is an interior point by assumption. At an interior minimizer $x^\star$ of a differentiable function F, both $x^\star + te_m$ and $x^\star - te_m$ are feasible for sufficiently small $t > 0$. Minimality in the two directions implies

$$\frac{\partial F(x^\star)}{\partial x_m} \ge 0, \qquad -\frac{\partial F(x^\star)}{\partial x_m} \ge 0,$$

and therefore $\partial F(x^\star)/\partial x_m = 0$. Applying this argument to every coordinate gives

$$\nabla J_\lambda(\beta^\star_\lambda) = 0$$

and

$$\mathbb{P}\left(\nabla \widehat{J}_{n,\lambda}(\widehat{\beta}_n) = 0\right) \longrightarrow 1.$$

Let

$$d_n = \widehat{\beta}_n - \beta^\star_\lambda, \qquad \beta_{n,t} = \beta^\star_\lambda + td_n, \quad t \in [0,1].$$

Since $\mathrm{d}\beta_{n,t}/\mathrm{d}t = d_n$, the chain rule gives

$$\frac{\mathrm{d}}{\mathrm{d}t}\nabla \widehat{J}_{n,\lambda}(\beta_{n,t}) = \nabla^2 \widehat{J}_{n,\lambda}(\beta_{n,t}) d_n.$$

The fundamental theorem of calculus therefore yields

$$\nabla \widehat{J}_{n,\lambda}(\widehat{\beta}_n) - \nabla \widehat{J}_{n,\lambda}(\beta^\star_\lambda) = \left\{\int_0^1 \nabla^2 \widehat{J}_{n,\lambda}(\beta_{n,t})\, dt\right\} d_n. \tag{14}$$

Define

$$\widehat{A}_n \coloneqq \int_0^1 \nabla^2 \widehat{J}_{n,\lambda}(\beta_{n,t})\, dt.$$

By Lemma C.2, for every $v \in \mathbb{R}^M$,

$$v^\top \widehat{A}_n v = \int_0^1 v^\top \nabla^2 \widehat{J}_{n,\lambda}(\beta_{n,t}) v\, dt \ge \lambda \|v\|_2^2.$$

Thus, $\widehat{A}_n \succeq \lambda I_M$. If $\widehat{A}_n$ were singular, there would exist a nonzero v such that $\widehat{A}_n v = 0$, contradicting

$$0 = v^\top \widehat{A}_n v \ge \lambda \|v\|_2^2 > 0.$$

Hence, $\widehat{A}_n$ is invertible. Moreover,

$$\lambda \|v\|_2^2 \le v^\top \widehat{A}_n v \le \|v\|_2 \|\widehat{A}_n v\|_2,$$

so $\|\widehat{A}_n v\|_2 \geq \lambda \|v\|_2$. For any $y \in \mathbb{R}^M$, set $v = \widehat{A}_n^{-1} y$. The inequality $\|\widehat{A}_n v\|_2 \geq \lambda \|v\|_2$ then gives

$$\|y\|_2 \geq \lambda \|\widehat{A}_n^{-1} y\|_2,$$

and hence

$$\|\widehat{A}_n^{-1} y\|_2 \leq \frac{1}{\lambda} \|y\|_2.$$

Taking the supremum over $\|y\|_2 = 1$ yields

$$\|\widehat{A}_n^{-1}\|_{\mathrm{op}} = \sup_{\|y\|_2 = 1} \|\widehat{A}_n^{-1} y\|_2 \leq \frac{1}{\lambda}.$$

We have shown the invertibility of $\widehat{A}_n$. Rearranging equation 14, we get

$$\sqrt{n}\, d_n = -\widehat{A}_n^{-1} \sqrt{n}\, \nabla \widehat{J}_{n,\lambda}(\beta_\lambda^\star) + \widehat{A}_n^{-1} \sqrt{n}\, \nabla \widehat{J}_{n,\lambda}(\widehat{\beta}_n).$$

Notice that the second term is zero with probability tending to one. That is, for every $\varepsilon > 0$,

$$\mathbb{P}\left(\left\| \widehat{A}_n^{-1} \sqrt{n}\, \nabla \widehat{J}_{n,\lambda}(\widehat{\beta}_n) \right\|_2 > \varepsilon \right) \leq \mathbb{P}\left(\nabla \widehat{J}_{n,\lambda}(\widehat{\beta}_n) \neq 0 \right) \longrightarrow 0.$$

Consequently,

$$\sqrt{n}\, d_n = -\widehat{A}_n^{-1} \sqrt{n}\, \nabla \widehat{J}_{n,\lambda}(\beta_\lambda^\star) + o_p(1). \tag{15}$$

We next show that $\widehat{A}_n \xrightarrow{p} A_\lambda$, where

$$A_\lambda = \nabla^2 J_\lambda(\beta_\lambda^\star).$$

Since $\widehat{\beta}_n, \beta_\lambda^\star \in \mathcal{B}_0$ and $\mathcal{B}_0$ is convex (shown in the proof of Lemma C.4), we have $\beta_{n,t} \in \mathcal{B}_0$ for every $t \in [0,1]$. Therefore,

$$\begin{aligned} \|\widehat{A}_n - A_\lambda\|_{\mathrm{op}} \leq & \int_0^1 \left\| \nabla^2 \widehat{J}_{n,\lambda}(\beta_{n,t}) - \nabla^2 J_\lambda(\beta_{n,t}) \right\|_{\mathrm{op}} dt \\ & + \int_0^1 \left\| \nabla^2 J_\lambda(\beta_{n,t}) - \nabla^2 J_\lambda(\beta_\lambda^\star) \right\|_{\mathrm{op}} dt. \end{aligned}$$

The first term is bounded by

$$\sup_{\beta \in \mathcal{B}_0} \left\| \nabla^2 \widehat{J}_{n,\lambda}(\beta) - \nabla^2 J_\lambda(\beta) \right\|_{\mathrm{op}} \xrightarrow{p} 0$$

by Lemma C.7. For the second term, the Lipschitz bound established in the proof of Lemma C.7 gives

$$\begin{aligned} \|\nabla^2 J_\lambda(\beta) - \nabla^2 J_\lambda(\beta')\|_{\mathrm{op}} &= 2\|P_T(H_\beta - H_{\beta'})\|_{\mathrm{op}} \\ &\leq 2P_T \|H_\beta - H_{\beta'}\|_{\mathrm{op}} \\ &\leq 2L_H \|\beta - \beta'\|_2. \end{aligned}$$

Since $\|\beta_{n,t} - \beta_\lambda^\star\|_2 = t\|d_n\|_2$,

$$\int_0^1 \left\| \nabla^2 J_\lambda(\beta_{n,t}) - \nabla^2 J_\lambda(\beta_\lambda^\star) \right\|_{\mathrm{op}} dt \leq 2L_H \int_0^1 t \|d_n\|_2 \, dt = L_H \|d_n\|_2 \xrightarrow{p} 0.$$

Hence,

$$\widehat{A}_n \xrightarrow{p} A_\lambda.$$

Lemma C.2 also gives $A_\lambda \succeq \lambda I_M$, so A_λ is invertible and $\|A_\lambda^{-1}\|_{\mathrm{op}} \leq 1/\lambda$. Using the identity

$$\widehat{A}_n^{-1} - A_\lambda^{-1} = \widehat{A}_n^{-1}(A_\lambda - \widehat{A}_n) A_\lambda^{-1}$$

and the submultiplicative property of the operator norm, i.e., $\|AB\|_{\mathrm{op}} \leq \|A\|_{\mathrm{op}} \|B\|_{\mathrm{op}}$, we obtain

$$\|\widehat{A}_n^{-1} - A_\lambda^{-1}\|_{\mathrm{op}} \leq \frac{1}{\lambda^2} \|\widehat{A}_n - A_\lambda\|_{\mathrm{op}} \xrightarrow{p} 0.$$

Therefore,

$$\widehat{A}_n^{-1} \xrightarrow{p} A_\lambda^{-1}. \tag{16}$$

It remains to derive the limiting distribution of the empirical score. Define

$$U_i = \phi(X_i^S) - \mathbb{E}_{P_S^X}[\phi(X)], \qquad V_j = \Phi_{\beta_\lambda^\star}(X_j^T) - \mathbb{E}_{P_T^X}[\Phi_{\beta_\lambda^\star}(X)].$$

Because $\nabla J_\lambda(\beta_\lambda^\star) = 0$,

$$\begin{aligned}\sqrt{n}\,\nabla \widehat{J}_{n,\lambda}(\beta_\lambda^\star) &= \sqrt{n}\left\{\nabla \widehat{J}_{n,\lambda}(\beta_\lambda^\star) - \nabla J_\lambda(\beta_\lambda^\star)\right\} \\ &= \sqrt{\frac{n}{n_S}}\frac{1}{\sqrt{n_S}}\sum_{i=1}^{n_S} U_i - \sqrt{\frac{n}{n_T}}\frac{1}{\sqrt{n_T}}\sum_{j=1}^{n_T} V_j.\end{aligned}$$

Assumption 3.1 and Lemma C.4 imply

$$\|\phi(X)\|_2 \le \sqrt{M}\kappa_1, \qquad \|\Phi_{\beta_\lambda^\star}(X)\|_2 \le C_\Phi,$$

so U_i and V_j have finite second moments. The multivariate central limit theorem therefore gives

$$\frac{1}{\sqrt{n_S}}\sum_{i=1}^{n_S} U_i \xrightarrow{d} Z_S, \qquad Z_S \sim \mathcal{N}(0, \Gamma_S),$$

where

$$\Gamma_S = \mathrm{Var}_{P_S^X}[\phi(X)],$$

and

$$\frac{1}{\sqrt{n_T}}\sum_{j=1}^{n_T} V_j \xrightarrow{d} Z_T, \qquad Z_T \sim \mathcal{N}(0, \Gamma_T),$$

where

$$\Gamma_T = \mathrm{Var}_{P_T^X}[\Phi_{\beta_\lambda^\star}(X)].$$

The source and target samples are independent, so Z_S and Z_T are independent. Since $n_S/n \to \rho$ and $n_T/n \to 1-\rho$,

$$\sqrt{\frac{n}{n_S}} \longrightarrow \frac{1}{\sqrt{\rho}}, \qquad \sqrt{\frac{n}{n_T}} \longrightarrow \frac{1}{\sqrt{1-\rho}}.$$

Slutsky's theorem gives

$$\sqrt{n}\,\nabla \widehat{J}_{n,\lambda}(\beta_\lambda^\star) \xrightarrow{d} \frac{1}{\sqrt{\rho}} Z_S - \frac{1}{\sqrt{1-\rho}} Z_T.$$

Because Z_S and Z_T are independent centered Gaussian vectors, their linear combination is centered Gaussian with covariance

$$\Sigma_\lambda = \frac{1}{\rho}\Gamma_S + \frac{1}{1-\rho}\Gamma_T.$$

Since

$$\Phi_{\beta_\lambda^\star}(X) = \frac{\phi(X)}{(g_\lambda^\star(X))^2},$$

this covariance is

$$\Sigma_\lambda = \frac{1}{\rho}\,\mathrm{Var}_{P_S^X}[\phi(X)] + \frac{1}{1-\rho}\,\mathrm{Var}_{P_T^X}\left[\frac{\phi(X)}{(g_\lambda^\star(X))^2}\right].$$

Therefore,

$$\sqrt{n}\,\nabla \widehat{J}_{n,\lambda}(\beta_\lambda^\star) \xrightarrow{d} \mathcal{N}(0, \Sigma_\lambda). \tag{17}$$

Now combining equation 15, equation 16, and equation 17, we have

$$\sqrt{n}\,(\widehat{\beta}_n - \beta_\lambda^\star) \xrightarrow{d} -A_\lambda^{-1} Z, \qquad Z \sim \mathcal{N}(0, \Sigma_\lambda).$$

A linear transformation of a centered Gaussian vector is centered Gaussian, and its covariance is

$$(-A_\lambda^{-1})\Sigma_\lambda(-A_\lambda^{-1})^\top = A_\lambda^{-1}\Sigma_\lambda A_\lambda^{-\top}.$$

Because A_λ is a symmetric Hessian, $A_\lambda^{-\top} = A_\lambda^{-1}$. Hence,

$$\sqrt{n}\,(\widehat{\beta}_n - \beta_\lambda^\star) \xrightarrow{d} \mathcal{N}\left(0, A_\lambda^{-1}\Sigma_\lambda A_\lambda^{-1}\right).$$

This completes the proof. □

D LOCAL-CONTAMINATION SENSITIVITY AND PROOF OF PROPOSITION 3.1

Throughout this section, we use the square-root parameterization from Section 2,

$$g_\beta(x) = \beta^\top \phi(x), \qquad w_\beta(x) = g_\beta(x)^2.$$

We derive the local-contamination sensitivity of HellingerIEP, KLIEP, and uLSIF separately, since KLIEP requires a KKT argument due to its source-normalization constraint.

Proof. **HellingerIEP.**

For HellingerIEP, define the uncontaminated population criterion

$$Q_{H,\lambda}(\beta; P_S, P_T) = \mathbb{E}_{P_S}[g_\beta(X)] + \mathbb{E}_{P_T}[g_\beta(X)^{-1}] + \frac{\lambda}{2}\|\beta\|_2^2,$$

and let $\beta^\star_{H,\lambda}$ denote the interior population minimizer appearing in the proposition.

Under point-mass contamination

$$P^X_{T,\varepsilon,z} = (1-\varepsilon)P^X_T + \varepsilon\Delta_z,$$

the corresponding contaminated criterion is

$$Q_{H,\lambda,\varepsilon,z}(\beta) = \mathbb{E}_{P_S}[g_\beta(X)] + (1-\varepsilon)\mathbb{E}_{P_T}[g_\beta(X)^{-1}] + \varepsilon g_\beta(z)^{-1} + \frac{\lambda}{2}\|\beta\|_2^2.$$

Thus $Q_{H,\lambda,0,z} = Q_{H,\lambda}$, and contamination affects only the target part of the criterion.

Let

$$\Psi_H(\beta, \varepsilon, z) := \nabla_\beta Q_{H,\lambda,\varepsilon,z}(\beta)$$

denote the population score. Since $g_\beta(x) = \beta^\top\phi(x)$,

$$\nabla_\beta g_\beta(x)^{-1} = -\frac{\phi(x)}{g_\beta(x)^2},$$

and hence

$$\Psi_H(\beta, \varepsilon, z) = \mathbb{E}_{P_S}[\phi(X)] - (1-\varepsilon)\mathbb{E}_{P_T}\left[\frac{\phi(X)}{g_\beta(X)^2}\right] - \varepsilon\frac{\phi(z)}{g_\beta(z)^2} + \lambda\beta.$$

Because $\beta^\star_{H,\lambda}$ is an interior minimizer of $Q_{H,\lambda}$, its first-order condition is

$$\Psi_H(\beta^\star_{H,\lambda}, 0, z) = 0.$$

Moreover,

$$A_{H,\lambda} = \nabla^2_\beta Q_{H,\lambda}(\beta^\star_{H,\lambda}; P_S, P_T) = 2\mathbb{E}_{P_T}\left[\frac{\phi(X)\phi(X)^\top}{g_{\beta^\star_{H,\lambda}}(X)^3}\right] + \lambda I_M,$$

where the source term contributes no Hessian because g_β is linear in β.

We first establish that the stationary point $\beta^\star_{H,\lambda}$ continues to a unique interior minimizer under sufficiently small contamination. On the feasible region where $g_\beta > 0$, the Hessian of the contaminated criterion is

$$\nabla^2_\beta Q_{H,\lambda,\varepsilon,z}(\beta) = 2(1-\varepsilon)\mathbb{E}_{P_T}\left[\frac{\phi(X)\phi(X)^\top}{g_\beta(X)^3}\right] + 2\varepsilon\frac{\phi(z)\phi(z)^\top}{g_\beta(z)^3} + \lambda I_M.$$

For $\varepsilon \in [0,1]$, each term is positive semidefinite, so $Q_{H,\lambda,\varepsilon,z}$ is convex on this region. If $\lambda > 0$, then $A_{H,\lambda} \succ 0$ immediately. If $\lambda = 0$, then $A_{H,0} \succeq 0$, and the assumed nonsingularity of $A_{H,0}$ implies $A_{H,0} \succ 0$. Thus, in either case,

$$A_{H,\lambda} \succ 0.$$

By the assumed twice continuous differentiability of the uncontaminated criterion, its score is continuously differentiable in β near $\beta^\star_{H,\lambda}$. The condition $g_{\beta^\star_{H,\lambda}}(z) > 0$, together with continuity of $g_\beta(z)$, implies that $g_\beta(z)^{-1}$ is smooth in a sufficiently small neighborhood of $\beta^\star_{H,\lambda}$. Moreover,

$$\Psi_H(\beta, \varepsilon, z) = \Psi_H(\beta, 0, z) + \varepsilon \left\{ \mathbb{E}_{P_T}\left[\frac{\phi(X)}{g_\beta(X)^2}\right] - \frac{\phi(z)}{g_\beta(z)^2} \right\},$$

so $\Psi_H(\beta, \varepsilon, z)$ is continuously differentiable jointly in (β, ε) near $(\beta^\star_{H,\lambda}, 0)$.

To apply the implicit function theorem, extend the algebraic expression for Ψ_H to ε in an open neighborhood of zero. This is well defined because the score depends affinely on ε. At $(\beta^\star_{H,\lambda}, 0)$,

$$\Psi_H(\beta^\star_{H,\lambda}, 0, z) = 0, \qquad \nabla_\beta \Psi_H(\beta^\star_{H,\lambda}, 0, z) = A_{H,\lambda},$$

and $A_{H,\lambda}$ is nonsingular. The implicit function theorem therefore yields a neighborhood of zero and a unique continuously differentiable function

$$\varepsilon \longmapsto \widetilde{\beta}_{H,\lambda}(\varepsilon)$$

satisfying

$$\widetilde{\beta}_{H,\lambda}(0) = \beta^\star_{H,\lambda}, \qquad \Psi_H(\widetilde{\beta}_{H,\lambda}(\varepsilon), \varepsilon, z) = 0.$$

Because $\beta^\star_{H,\lambda}$ is interior and $\widetilde{\beta}_{H,\lambda}(\varepsilon) \to \beta^\star_{H,\lambda}$ as $\varepsilon \to 0$, the stationary points remain interior for all sufficiently small ε. Continuity of the Hessian together with $A_{H,\lambda} \succ 0$ further implies

$$\nabla^2_\beta Q_{H,\lambda,\varepsilon,z}\big(\widetilde{\beta}_{H,\lambda}(\varepsilon)\big) \succ 0$$

for all sufficiently small ε. Thus $\widetilde{\beta}_{H,\lambda}(\varepsilon)$ is a strict local minimizer.

Now restrict to sufficiently small $\varepsilon \geq 0$. Since $Q_{H,\lambda,\varepsilon,z}$ is convex, this interior stationary point is also a global minimizer. It is unique: if there were another distinct global minimizer, convexity would imply that every point on the line segment joining the two minimizers is also a minimizer, contradicting strict local minimality of $\widetilde{\beta}_{H,\lambda}(\varepsilon)$. We may therefore define

$$\beta^\star_{H,\lambda,\varepsilon} := \widetilde{\beta}_{H,\lambda}(\varepsilon).$$

This proves that, for all sufficiently small $\varepsilon \geq 0$, the contaminated criterion has a unique interior minimizer and that $\varepsilon \mapsto \beta^\star_{H,\lambda,\varepsilon}$ is differentiable at zero.

It remains to compute the derivative of $\beta^\star_{H,\lambda,\varepsilon}$ at $\varepsilon = 0$. Since

$$\Psi_H(\beta^\star_{H,\lambda,\varepsilon}, \varepsilon, z) = 0,$$

the chain rule gives

$$0 = D_\beta \Psi_H(\beta^\star_{H,\lambda,\varepsilon}, \varepsilon, z) \frac{d}{d\varepsilon} \beta^\star_{H,\lambda,\varepsilon} + \partial_\varepsilon \Psi_H(\beta^\star_{H,\lambda,\varepsilon}, \varepsilon, z).$$

Evaluating at $\varepsilon = 0$ gives

$$0 = A_{H,\lambda} \left. \frac{d}{d\varepsilon} \beta^\star_{H,\lambda,\varepsilon} \right|_{\varepsilon=0} + \partial_\varepsilon \Psi_H(\beta^\star_{H,\lambda}, \varepsilon, z)\big|_{\varepsilon=0}.$$

With β held fixed,

$$\partial_\varepsilon \Psi_H(\beta, \varepsilon, z) = \mathbb{E}_{P_T}\left[\frac{\phi(X)}{g_\beta(X)^2}\right] - \frac{\phi(z)}{g_\beta(z)^2}.$$

Define

$$\psi_H(x; \beta) := -\frac{\phi(x)}{g_\beta(x)^2} = -\frac{\phi(x)}{w_\beta(x)}.$$

Then

$$\partial_\varepsilon \Psi_H(\beta^\star_{H,\lambda}, \varepsilon, z)\big|_{\varepsilon=0} = \psi_H(z; \beta^\star_{H,\lambda}) - \mathbb{E}_{P_T}\left[\psi_H(X; \beta^\star_{H,\lambda})\right].$$

Therefore,

$$\frac{d}{d\varepsilon}\beta^\star_{H,\lambda,\varepsilon}\Big|_{\varepsilon=0} = -A^{-1}_{H,\lambda}\left\{\psi_H(z;\beta^\star_{H,\lambda}) - \mathbb{E}_{P_T}\left[\psi_H(X;\beta^\star_{H,\lambda})\right]\right\}.$$

Finally, since

$$w_\beta(x) = g_\beta(x)^2 = \{\beta^\top\phi(x)\}^2,$$

we have

$$\nabla_\beta w_\beta(x) = 2g_\beta(x)\phi(x).$$

A final application of the chain rule therefore gives

$$\frac{d}{d\varepsilon}w_{\beta^\star_{H,\lambda,\varepsilon}}(x)\Big|_{\varepsilon=0} = 2g_{\beta^\star_{H,\lambda}}(x)\phi(x)^\top\ \frac{d}{d\varepsilon}\beta^\star_{H,\lambda,\varepsilon}\Big|_{\varepsilon=0},$$

as claimed.

KLIEP.

For KLIEP, we retain the same square-root parameterization $w_\beta = g_\beta^2$ so that the target score contributions can be compared under a common parameterization. The population objective is

$$Q_K(\beta;P_T) = -\mathbb{E}_{P_T}[\log w_\beta(X)],$$

subject to the normalization constraint

$$c_K(\beta;P_S) := \mathbb{E}_{P_S}[w_\beta(X)] - 1 = 0.$$

Let $\beta^\star_K$ denote an uncontaminated constrained minimizer, and assume that $\beta^\star_{K,m} > 0$ for every $m = 1,\dots,M$, so the coefficient nonnegativity constraints are locally inactive. Let $\eta^\star_K$ denote the Lagrange multiplier associated with the normalization constraint. The corresponding Lagrangian is

$$\mathcal{L}_K(\beta,\eta;P_S,P_T) = Q_K(\beta;P_T) + \eta\, c_K(\beta;P_S).$$

Under target contamination,

$$P_{T,\varepsilon,z} = (1-\varepsilon)P_T + \varepsilon\Delta_z,$$

the objective becomes

$$Q_{K,\varepsilon,z}(\beta) = -(1-\varepsilon)\mathbb{E}_{P_T}[\log w_\beta(X)] - \varepsilon\log w_\beta(z),$$

while the normalization constraint is unchanged because it depends only on P_S.

We first identify the contribution of a single target observation to the coefficient score. Since $w_\beta(x) = g_\beta(x)^2$ and $g_\beta(x) > 0$,

$$-\log w_\beta(x) = -2\log g_\beta(x),$$

and therefore

$$\psi_K(x;\beta) := \nabla_\beta\{-\log w_\beta(x)\} = -\frac{2\phi(x)}{g_\beta(x)} = -\frac{2\phi(x)}{\sqrt{w_\beta(x)}}.$$

Thus the target part of the contaminated objective gradient is

$$(1-\varepsilon)\mathbb{E}_{P_T}[\psi_K(X;\beta)] + \varepsilon\psi_K(z;\beta).$$

Because KLIEP is constrained, its local sensitivity is obtained by differentiating both the stationarity and normalization equations. Define the KKT map

$$F_K(\beta,\eta,\varepsilon,z) := \begin{pmatrix}\nabla_\beta\mathcal{L}_K(\beta,\eta;P_S,P_{T,\varepsilon,z})\\ c_K(\beta;P_S)\end{pmatrix}.$$

At the uncontaminated solution,

$$F_K(\beta^\star_K,\eta^\star_K,0,z) = 0.$$

Let

$$H_K := \nabla^2_{\beta\beta}\mathcal{L}_K(\beta^\star_K,\eta^\star_K;P_S,P_T), \qquad a_K := \nabla_\beta c_K(\beta^\star_K;P_S).$$

Direct differentiation gives

$$H_K = 2\mathbb{E}_{P_T}\left[\frac{\phi(X)\phi(X)^\top}{g_{\beta_K^\star}(X)^2}\right] + 2\eta_K^\star \mathbb{E}_{P_S}[\phi(X)\phi(X)^\top],$$

and

$$a_K = 2\mathbb{E}_{P_S}[g_{\beta_K^\star}(X)\phi(X)].$$

Hence the Jacobian of the KKT map with respect to (β, η) is

$$\mathcal{K}_K = \begin{pmatrix} H_K & a_K \\ a_K^\top & 0 \end{pmatrix}.$$

Under the usual local regularity conditions for the constrained problem, including nonsingularity of $\mathcal{K}_K$, the implicit function theorem gives a differentiable local solution branch

$$\varepsilon \longmapsto \big(\widetilde{\beta}_K(\varepsilon), \widetilde{\eta}_K(\varepsilon)\big)$$

through $(\beta_K^\star, \eta_K^\star)$. Under the constrained second-order sufficient condition, the coefficient component remains a strict constrained local minimizer for sufficiently small ε. We denote this local continuation by $\beta_{K,\varepsilon}^\star$, with associated multiplier $\eta_{K,\varepsilon}^\star$.

The KKT equations satisfy

$$F_K(\beta_{K,\varepsilon}^\star, \eta_{K,\varepsilon}^\star, \varepsilon, z) = 0.$$

Differentiating with respect to ε at zero gives

$$\mathcal{K}_K \begin{pmatrix} \frac{d}{d\varepsilon}\beta_{K,\varepsilon}^\star\Big|_{\varepsilon=0} \\ \frac{d}{d\varepsilon}\eta_{K,\varepsilon}^\star\Big|_{\varepsilon=0} \end{pmatrix} + \begin{pmatrix} \psi_K(z;\beta_K^\star) - \mathbb{E}_{P_T}[\psi_K(X;\beta_K^\star)] \\ 0 \end{pmatrix} = 0.$$

The zero in the second component appears because the normalization constraint depends only on P_S and therefore has no direct dependence on ε. Since $\mathcal{K}_K$ is nonsingular,

$$\begin{pmatrix} \frac{d}{d\varepsilon}\beta_{K,\varepsilon}^\star\Big|_{\varepsilon=0} \\ \frac{d}{d\varepsilon}\eta_{K,\varepsilon}^\star\Big|_{\varepsilon=0} \end{pmatrix} = -\mathcal{K}_K^{-1} \begin{pmatrix} \psi_K(z;\beta_K^\star) - \mathbb{E}_{P_T}[\psi_K(X;\beta_K^\star)] \\ 0 \end{pmatrix}.$$

Thus the target-contamination term entering the KLIEP sensitivity is governed by

$$\psi_K(x;\beta) = -\frac{2\phi(x)}{\sqrt{w_\beta(x)}}.$$

uLSIF.

For uLSIF, we again use the common parameterization $w_\beta = g_\beta^2$. Its uncontaminated population criterion is

$$Q_{U,0}(\beta) := Q_U(\beta; P_S, P_T) = \frac{1}{2}\mathbb{E}_{P_S}[w_\beta(X)^2] - \mathbb{E}_{P_T}[w_\beta(X)].$$

Under target contamination, the criterion becomes

$$Q_{U,\varepsilon,z}(\beta) = \frac{1}{2}\mathbb{E}_{P_S}[w_\beta(X)^2] - (1-\varepsilon)\mathbb{E}_{P_T}[w_\beta(X)] - \varepsilon w_\beta(z).$$

Let $\beta_U^\star$ be an interior local minimizer of the uncontaminated criterion and write

$$A_U := \nabla_\beta^2 Q_{U,0}(\beta_U^\star).$$

For the corresponding local sensitivity calculation, suppose A_U is nonsingular.

The contribution of a single target observation is simply $-w_\beta(x) = -g_\beta(x)^2$. Therefore

$$\psi_U(x;\beta) := \nabla_\beta\{-w_\beta(x)\} = -2g_\beta(x)\phi(x) = -2\sqrt{w_\beta(x)}\,\phi(x).$$

The population score is consequently

$$\Psi_U(\beta, \varepsilon, z) := \nabla_\beta Q_{U,\varepsilon,z}(\beta) = 2\mathbb{E}_{P_S}[g_\beta(X)^3\phi(X)] + (1-\varepsilon)\mathbb{E}_{P_T}[\psi_U(X;\beta)] + \varepsilon\psi_U(z;\beta).$$

At the uncontaminated minimizer,

$$\Psi_U(\beta_U^\star, 0, z) = 0,$$

and direct differentiation gives

$$A_U = 6\mathbb{E}_{P_S}[g_{\beta_U^\star}(X)^2\phi(X)\phi(X)^\top] - 2\mathbb{E}_{P_T}[\phi(X)\phi(X)^\top].$$

Since $\beta_U^\star$ is a local minimizer, its Hessian is positive semidefinite; together with the assumed nonsingularity of A_U, this gives $A_U \succ 0$. The score is continuously differentiable near $(\beta_U^\star, 0)$, so the implicit function theorem yields a differentiable local stationary branch through $\beta_U^\star$. By continuity of the Hessian, this branch remains a strict local minimizer for sufficiently small ε. Denote it by $\beta_{U,\varepsilon}^\star$.

Its first-order condition is

$$\Psi_U(\beta_{U,\varepsilon}^\star, \varepsilon, z) = 0.$$

Differentiating with respect to ε at zero gives

$$0 = A_U \left.\frac{d}{d\varepsilon}\beta_{U,\varepsilon}^\star\right|_{\varepsilon=0} + \partial_\varepsilon \Psi_U(\beta_U^\star, \varepsilon, z)|_{\varepsilon=0}.$$

Since only the target distribution depends directly on ε,

$$\partial_\varepsilon \Psi_U(\beta_U^\star, \varepsilon, z)|_{\varepsilon=0} = \psi_U(z;\beta_U^\star) - \mathbb{E}_{P_T}[\psi_U(X;\beta_U^\star)].$$

Therefore

$$\left.\frac{d}{d\varepsilon}\beta_{U,\varepsilon}^\star\right|_{\varepsilon=0} = -A_U^{-1}\left\{\psi_U(z;\beta_U^\star) - \mathbb{E}_{P_T}[\psi_U(X;\beta_U^\star)]\right\}.$$

In particular, the target score contribution entering the uLSIF sensitivity is

$$\psi_U(x;\beta) = -2\sqrt{w_\beta(x)}\,\phi(x),$$

as claimed. □

E Additional Density-Ratio Estimator Formulations

For completeness, we briefly describe two additional estimators included in the appendix experiments: a direct parameterization of the Hellinger objective and an estimator based on the negative exponential divergence.

Direct-HellingerIEP. The HellingerIEP formulation in Section 2 parameterizes the square root of the density ratio, i.e., $g(x) = \sqrt{w(x)}$, which yields a convex objective in the kernel coefficients. As an ablation, we also consider a direct parameterization

$$w_\beta(x) = \sum_{m=1}^{M} \beta_m K(x, c_m), \qquad \beta_m \geq 0.$$

Substituting this representation directly into the Hellinger criterion gives

$$\widehat{J}_{\mathrm{DH}}(\beta) = \frac{1}{n_S}\sum_{i=1}^{n_S}\sqrt{w_\beta(X_i^S)} + \frac{1}{n_T}\sum_{j=1}^{n_T}\frac{1}{\sqrt{w_\beta(X_j^T)}} + \frac{\lambda}{2}\|\beta\|_2^2.$$

We refer to the resulting estimator as Direct-HellingerIEP. It uses the same Hellinger population criterion as HellingerIEP but removes the square-root parameterization; consequently, the optimization problem is generally nonconvex.

Negative-Exponential-Divergence IEP (NEDIEP). We also consider an alternative estimator based on the negative exponential divergence (Basu et al., 1997). We define the population criterion

$$J_{\mathrm{NE}}(w) = \mathbb{E}_{P_T}\left[e^{1-w(X)}\right] - \mathbb{E}_{P_S}\left[(1+w(X))e^{1-w(X)}\right].$$

Since $p_T(x) = r(x)p_S(x)$, the pointwise integrand is

$$\{r(x) - 1 - w(x)\}e^{1-w(x)},$$

whose derivative with respect to $w(x)$ is

$$\{w(x) - r(x)\}e^{1-w(x)}.$$

Thus the population criterion is minimized at $w(x) = r(x)$. Using the same nonnegative kernel model

$$w_\beta(x) = \sum_{m=1}^{M} \beta_m K(x, c_m), \qquad \beta_m \geq 0,$$

the empirical objective becomes

$$\widehat{J}_{\mathrm{NE}}(\beta) = \frac{1}{n_T}\sum_{j=1}^{n_T} e^{1-w_\beta(X_j^T)} - \frac{1}{n_S}\sum_{i=1}^{n_S}\left(1 + w_\beta(X_i^S)\right)e^{1-w_\beta(X_i^S)}.$$

We refer to the resulting estimator as the Negative-Exponential-Divergence Importance Estimation Procedure (NEDIEP). Like Direct-HellingerIEP, its empirical optimization problem is generally nonconvex, and we include it as an additional comparison.

A useful property of this criterion is that its change under finite contamination is uniformly bounded. For every $u \geq 0$,

$$0 < e^{1-u} \leq e, \qquad 0 < (1+u)e^{1-u} \leq e.$$

Consider contaminated source and target distributions

$$P_{S,\varepsilon_S}^X = (1-\varepsilon_S)P_S^X + \varepsilon_S Q_S^X,$$
$$P_{T,\varepsilon_T}^X = (1-\varepsilon_T)P_T^X + \varepsilon_T Q_T^X,$$

where Q_S^X and Q_T^X are arbitrary probability distributions on the covariate space. Let $J_{\mathrm{NE},\varepsilon}$ denote the criterion obtained by replacing P_S^X and P_T^X with these contaminated distributions, and write $J_{\mathrm{NE},0} = J_{\mathrm{NE}}$. Because each observation-level term takes values in an interval of length at most e, any class $\mathcal{W}$ of nonnegative functions satisfies

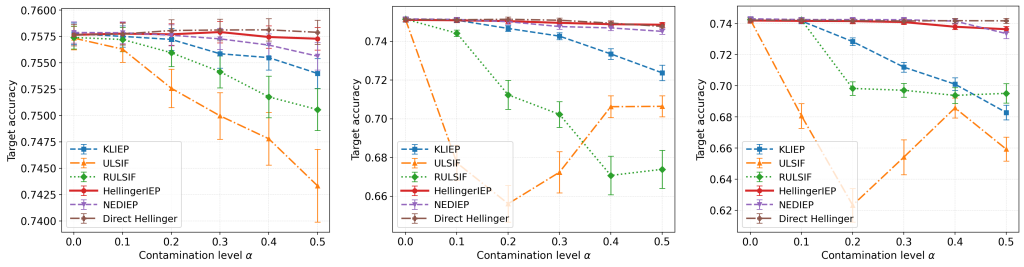
$$\sup_{w\in\mathcal{W}} |J_{\mathrm{NE},\varepsilon}(w) - J_{\mathrm{NE},0}(w)| \leq e(\varepsilon_S + \varepsilon_T).$$

This bound holds uniformly over $w \in \mathcal{W}$ and the contaminating distributions Q_S and Q_T. In contrast, the observation-level terms $\sqrt{w}$ and $w^{-1/2}$ in the dual Hellinger criterion are unbounded over $w > 0$. However, the bound establishes stability of the negative-exponential criterion under contamination and does not imply a bounded influence function for the resulting estimator.

F Additional Downstream Prediction Performance Results

F.1 Binary Classification

We supplement the binary classification results in Section 4 by including NEDIEP and Direct-HellingerIEP in the comparison, while keeping the experimental setup unchanged. Figure 3 reports the results. HellingerIEP, Direct-HellingerIEP, and NEDIEP remain comparatively stable as the contamination level increases and achieve stronger predictive performance than KLIEP, uLSIF and RuLSIF.



(a) $d = 1$ (b) $d = 5$ (c) $d = 10$

Figure 3: Target classification accuracy under contaminated covariate shift for $d \in \{1, 5, 10\}$, comparing HellingerIEP, KLIEP, uLSIF, RuLSIF, NEDIEP, and Direct-HellingerIEP. Points denote Monte Carlo means and error bars denote standard errors over 50 replicates. Higher values indicate better predictive performance.

F.2 Regression

For the regression experiment, we use the same source-target covariate distributions as in the classification experiment, with only the first coordinate exhibiting covariate shift. Specifically,

$$X_1^S \sim \frac{1}{2}\mathcal{N}(2,4) + \frac{1}{2}\mathcal{N}(-2,4), \qquad X_1^T \sim \frac{1}{2}\mathcal{N}(1,2) + \frac{1}{2}\mathcal{N}(-1,2),$$

while, for $j = 2, \ldots, d$,

$$X_j^S, X_j^T \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1).$$

The response is generated in both domains according to

$$Y = 1.5X_1 + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0,1), \qquad \varepsilon \perp\!\!\!\perp X,$$

so that only the first coordinate is predictive and the conditional distribution $P(Y \mid X)$ is identical across the source and target domains.

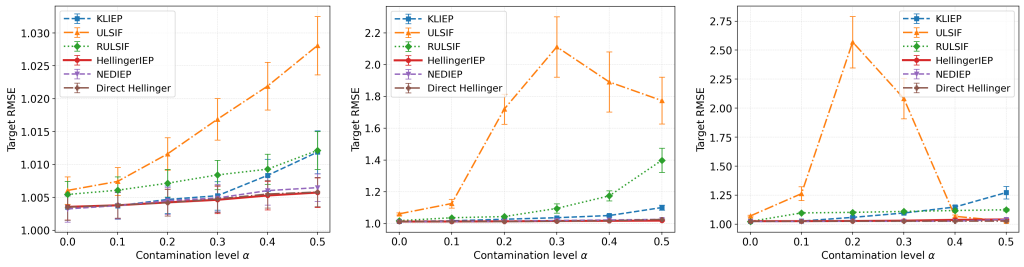
For each replicate, we generate $n_S = 300$ labeled source observations, $n_T = 300$ unlabeled target observations, and an independent clean target test set of size 3000. For each contamination level $\alpha \in \{0, 0.1, \ldots, 0.5\}$, an α-fraction of the unlabeled target covariate vectors is replaced by draws from $\mathcal{N}(0, 0.001^2 I_d)$. The contaminated target sample is used only for estimating the importance weights.

We compare KLIEP, uLSIF, RuLSIF, HellingerIEP, Direct-HellingerIEP, and NEDIEP, with tunable hyperparameters for all methods selected by cross-validation. For each method, a weighted linear regression model is fitted to the labeled source sample using the estimated importance weights, and predictive performance is measured by RMSE on the clean target test set. Each configuration is repeated over 50 independent Monte Carlo replicates.

Performance results are presented in Figure 4. Consistent with the classification results, HellingerIEP, Direct-HellingerIEP, and NEDIEP outperform KLIEP, uLSIF, and RuLSIF, particularly under contamination.

F.3 Downstream Performance under Model Misspecification

The synthetic downstream experiments in the main text and the preceding appendix sections use predictive models that are correctly specified for the common conditional distribution $P(Y \mid X)$. We therefore consider an additional setting in which the covariate-shift assumption remains exactly satisfied, but the downstream prediction model is deliberately misspecified. This distinction is important because, under model misspecification, the best approximation within a restricted model class can depend on the marginal distribution of X. Consequently, the source- and target-optimal predictors need not coincide, and importance weighting has a genuine role in adapting the fitted predictor to the target population.



(a) $d = 1$ (b) $d = 5$ (c) $d = 10$

Figure 4: Target regression RMSE under contaminated covariate shift for $d \in \{1, 5, 10\}$, comparing HellingerIEP, KLIEP, uLSIF, RuLSIF, NEDIEP, and Direct-HellingerIEP. Points denote Monte Carlo means and error bars denote standard errors over 50 replicates. Lower values indicate better predictive performance.

We use the same covariate distributions for both classification and regression. Only the first coordinate exhibits distribution shift, with

$$X_1^S \sim \tfrac{1}{2}\mathcal{N}(2,4) + \tfrac{1}{2}\mathcal{N}(-2,4), \qquad X_1^T \sim \tfrac{1}{2}\mathcal{N}(1,2) + \tfrac{1}{2}\mathcal{N}(-1,2).$$

For $j = 2, \ldots, d$, the remaining coordinates are generated independently from $\mathcal{N}(0,1)$ in both domains and contain no predictive information. We consider $d \in \{1, 5, 10\}$.

For classification, the source and target domains share the nonlinear conditional model

$$P(Y = 1 \mid X = x) = \mathrm{sigmoid}\big(0.5 + x_1 - 0.18x_1^2\big),$$

while the downstream predictor is an ordinary linear logistic regression and therefore omits the quadratic term. For regression, we similarly generate

$$Y = 1.5X_1 + 0.10X_1^2 + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0,1),$$

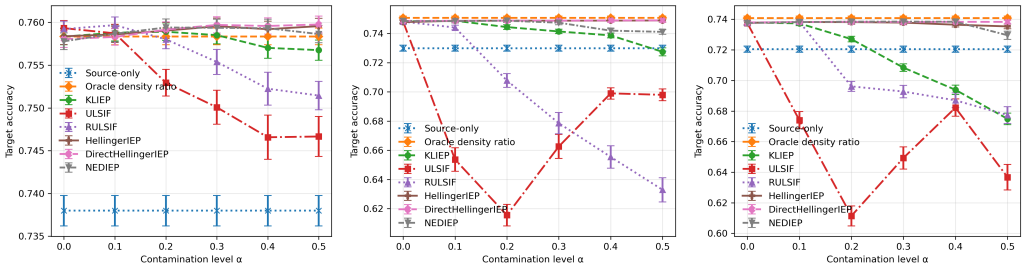
and fit a linear regression model. Thus, in both cases, $P(Y \mid X)$ is identical across source and target domains, but the fitted prediction model is mildly misspecified.

For each replicate, we generate $n_S = 300$ labeled source observations, $n_T = 300$ unlabeled target observations, and an independent clean target test set of size 3000. For contamination levels $\alpha \in \{0, 0.1, \ldots, 0.5\}$, an α-fraction of the unlabeled target sample is replaced by draws from $\mathcal{N}(0, 0.001^2 I_d)$, while the target test set remains uncontaminated. We compare the source-only estimator, the oracle clean density ratio, KLIEP, uLSIF, RuLSIF, HellingerIEP, Direct-HellingerIEP, and NEDIEP. Downstream performance is evaluated using target accuracy for classification and target RMSE for regression, with results averaged over 100 Monte Carlo replicates.

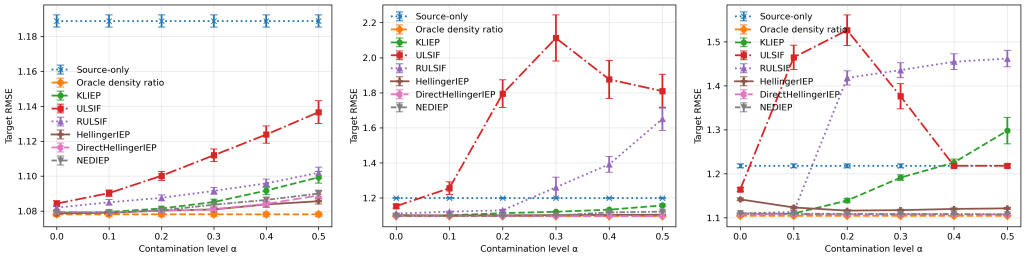
Figure 5 shows that the advantage of robust importance weighting persists when the downstream prediction model is misspecified. The oracle density ratio provides a benchmark for ideal importance weighting under the clean target distribution. The robust divergence-based methods, including HellingerIEP, Direct-HellingerIEP, and NEDIEP, remain close to this benchmark in both classification and regression. In contrast, KLIEP, uLSIF, and RuLSIF become increasingly unstable as contamination grows, with the deterioration becoming especially pronounced in $d = 5$ and $d = 10$. These results indicate that the robustness of the Hellinger-based criterion is not limited to the well-specified downstream models considered in the main experiments, but continues to translate into stable target prediction when the fitted predictive model only approximates the true conditional relationship.

G Comparison with Additional Robust Density-Ratio Estimators

We supplement the main experiments with comparisons against Trimmed-KLIEP, WeightedDRE, and γ-DRE. These methods use log-linear density-ratio parameterizations and introduce additional robustness through trimming, observation weighting, or a robust divergence objective. These comparisons therefore provide a complementary evaluation of HellingerIEP against estimators specifically designed to mitigate the effects of contamination.



(a) Classification, $d = 1$ (b) Classification, $d = 5$ (c) Classification, $d = 10$



(d) Regression, $d = 1$ (e) Regression, $d = 5$ (f) Regression, $d = 10$

Figure 5: Downstream predictive performance under model misspecification for $d \in \{1, 5, 10\}$. The top row reports target classification accuracy and the bottom row reports target regression RMSE as the contamination level increases. The covariate-shift assumption remains satisfied, but the downstream prediction model is deliberately misspecified. Points denote means and error bars denote standard errors over 100 Monte Carlo replicates.

G.1 Shared Experimental Setup

Alongside HellingerIEP, we include a log-linear variant, denoted LL-HellingerIEP, with

$$r_\theta(x) = \exp\{\theta^\top h(x)\}.$$

This provides a comparison within the log-linear model class and helps distinguish the effects of the estimation objective from those of the ratio parameterization. All methods use RBF kernel features, with tunable hyperparameters selected separately by cross-validation using their corresponding density-ratio objectives.

For density-ratio estimation, we use the same Gaussian distribution-shift setting as in Section 4 and report NMSE under both clean target samples and 30% target contamination. We consider two experimental designs. In the dimension experiment, we fix $n_S = 100$ and $n_T = 1000$ and vary

$$d \in \{1, 4, 7, 10, 13, 16, 19\}.$$

In the sample-size experiment, we fix $d = 10$ and $n_T = 1000$ and vary

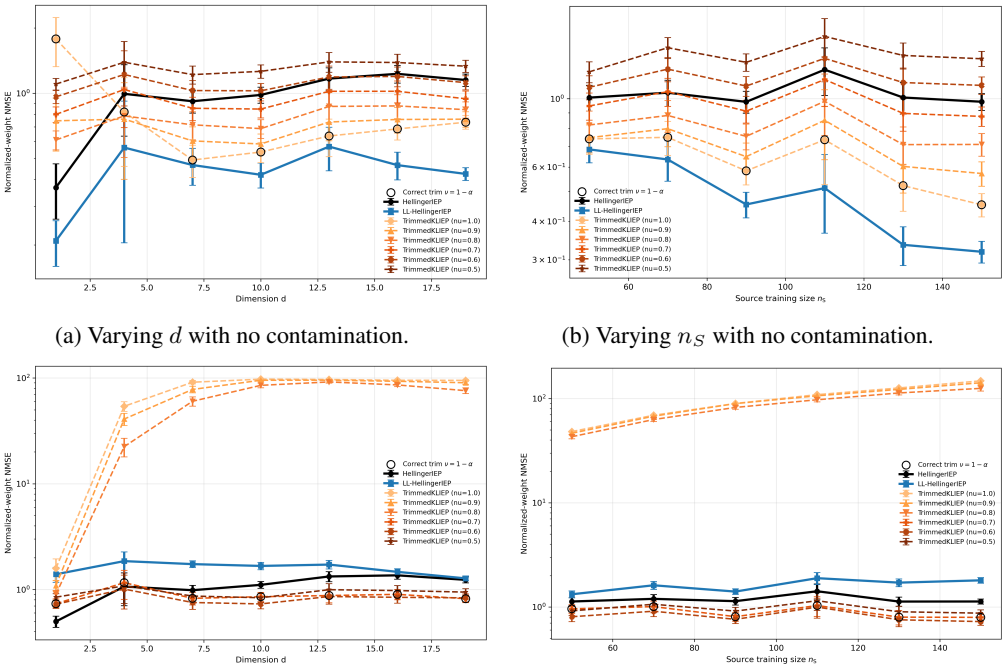
$$n_S \in \{50, 70, 90, 110, 130, 150\}.$$

Under contamination, 30% of the target observations are replaced by the same concentrated contamination component used in the corresponding main-text experiment. Density-ratio estimation accuracy is evaluated by NMSE relative to the oracle density ratio.

For downstream prediction, we use the same binary classification setting as in Section 4, with $d \in \{1, 5, 10\}$, $n_S = n_T = 300$, and an independent clean target test set of size 3000. For each contamination level $\alpha \in \{0, 0.1, \ldots, 0.5\}$, an α-fraction of the unlabeled target covariates is replaced by draws from

$$\mathcal{N}(0, 0.001^2 I_d).$$

The contaminated target sample is used only to estimate the importance weights. For each method, a weighted logistic regression model is then fitted to the labeled source sample, and predictive performance is measured by classification accuracy on the clean target test set. Each configuration is repeated over 50 independent Monte Carlo replicates.



(a) Varying d with no contamination.

(b) Varying n_S with no contamination.

(c) Varying d with 30% contamination.

(d) Varying n_S with 30% contamination.

Figure 6: Density-ratio estimation NMSE for HellingerIEP, LL-HellingerIEP, and Trimmed-KLIEP at different retained target fractions. The top row shows the clean setting, and the bottom row shows 30% target contamination.

G.2 Comparison with Trimmed-KLIEP

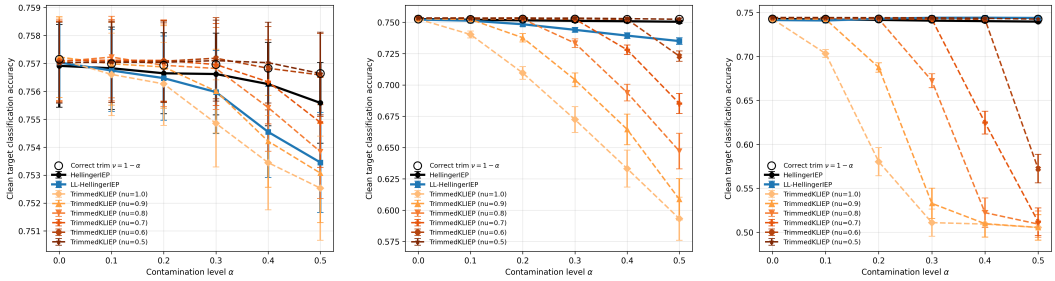
Trimmed-KLIEP applies trimming to a log-linear KLIEP objective. Its robustness depends on a user-specified retained target fraction ν, with trimming proportion $1 - \nu$. We compare HellingerIEP and LL-HellingerIEP with Trimmed-KLIEP over

$$\nu \in \{1.0, 0.9, 0.8, 0.7, 0.6, 0.5\}.$$

Density-ratio estimation. Figure 6 reports NMSE under the shared density-ratio estimation protocol. Trimmed-KLIEP can perform well when the retained fraction is appropriately chosen, but its performance depends on this tuning parameter. Insufficient trimming can leave contaminated observations influential. In contrast, HellingerIEP and LL-HellingerIEP remain stable across the settings considered without requiring a pre-specified trimming proportion.

Downstream predictive performance. Figure 7 reports target classification accuracy under the shared downstream protocol. We show the full retained-fraction grid and separately indicate the contamination-matched choice $\nu = 1 - \alpha$, allowing us to assess the sensitivity of Trimmed-KLIEP to the trimming proportion.

Across all three dimensions, HellingerIEP and LL-HellingerIEP remain relatively stable as the contamination level increases. In $d = 1$, the methods perform similarly and the choice of retained fraction has little effect. The differences become more pronounced in $d = 5$ and $d = 10$. Trimmed-KLIEP performs competitively when the retained fraction is well calibrated to the contamination level, but its performance can deteriorate when ν is misspecified, particularly in higher dimensions. HellingerIEP, in contrast, does not require specification of a trimming proportion or prior knowledge of the contamination level.



(a) $d = 1$ (b) $d = 5$ (c) $d = 10$

Figure 7: Target classification accuracy for HellingerIEP, LL-HellingerIEP, and Trimmed-KLIEP under increasing target contamination. Trimmed-KLIEP uses retained fractions $\nu \in \{1.0, 0.9, 0.8, 0.7, 0.6, 0.5\}$, with the contamination-matched choice $\nu = 1 - \alpha$ indicated separately. Points denote means and error bars denote standard errors over 50 Monte Carlo replicates.

G.3 Comparison with WeightedDRE and γ-DRE

WeightedDRE and γ-DRE (Nagumo & Fujisawa, 2024) use log-linear density-ratio models together with the observation weight

$$\omega_\tau(x) = \exp\left\{-\frac{1}{\tau}\left\|\frac{x - \mathrm{Med}}{\mathrm{MADN}}\right\|_4^4\right\},$$

where Med and MADN denote the coordinatewise median and normalized median absolute deviation computed from the pooled source and target covariates. This weighting scheme downweights observations that lie far from the center of the pooled covariate distribution, with smaller values of τ producing more aggressive downweighting. WeightedDRE incorporates these weights into a weighted KL-based criterion, whereas γ-DRE combines them with a γ-divergence objective.

We compare WeightedDRE and γ-DRE with HellingerIEP and LL-HellingerIEP under the shared experimental protocols described above. For both methods, we consider

$$\tau \in \{5, 10, 50, 100\},$$

and fix the divergence parameter of γ-DRE at $\gamma = 0.01$, following the settings considered in Nagumo & Fujisawa (2024). Results are reported separately for each value of τ, rather than selecting τ based on test performance.

Density-ratio estimation. Figure 8 reports NMSE under the shared density-ratio estimation experiment protocol. In the clean setting, HellingerIEP and LL-HellingerIEP perform comparably to WeightedDRE and γ-DRE and achieve lower NMSE in several configurations. Under 30% target contamination, HellingerIEP and LL-HellingerIEP remain stable across the dimensions and source sample sizes considered, whereas WeightedDRE and γ-DRE exhibit greater degradation under contamination.

Downstream predictive performance. Figure 9 reports target classification accuracy under the shared downstream protocol. WeightedDRE and γ-DRE deteriorate more rapidly as contamination increases, particularly in higher dimensions. One possible explanation is that their weighting mechanism primarily attenuates observations that are far from the center of the pooled covariate distribution, whereas the contamination considered here is concentrated within the main data support. Such observations may therefore receive relatively little downweighting while still perturbing the estimated density ratio.

H Choice of the Hellinger Divergence within the Power-Divergence Family

We briefly discuss why we focus on the Hellinger divergence in our density-ratio objective. The Hellinger divergence is a member of the broader power-divergence family, which provides a one-

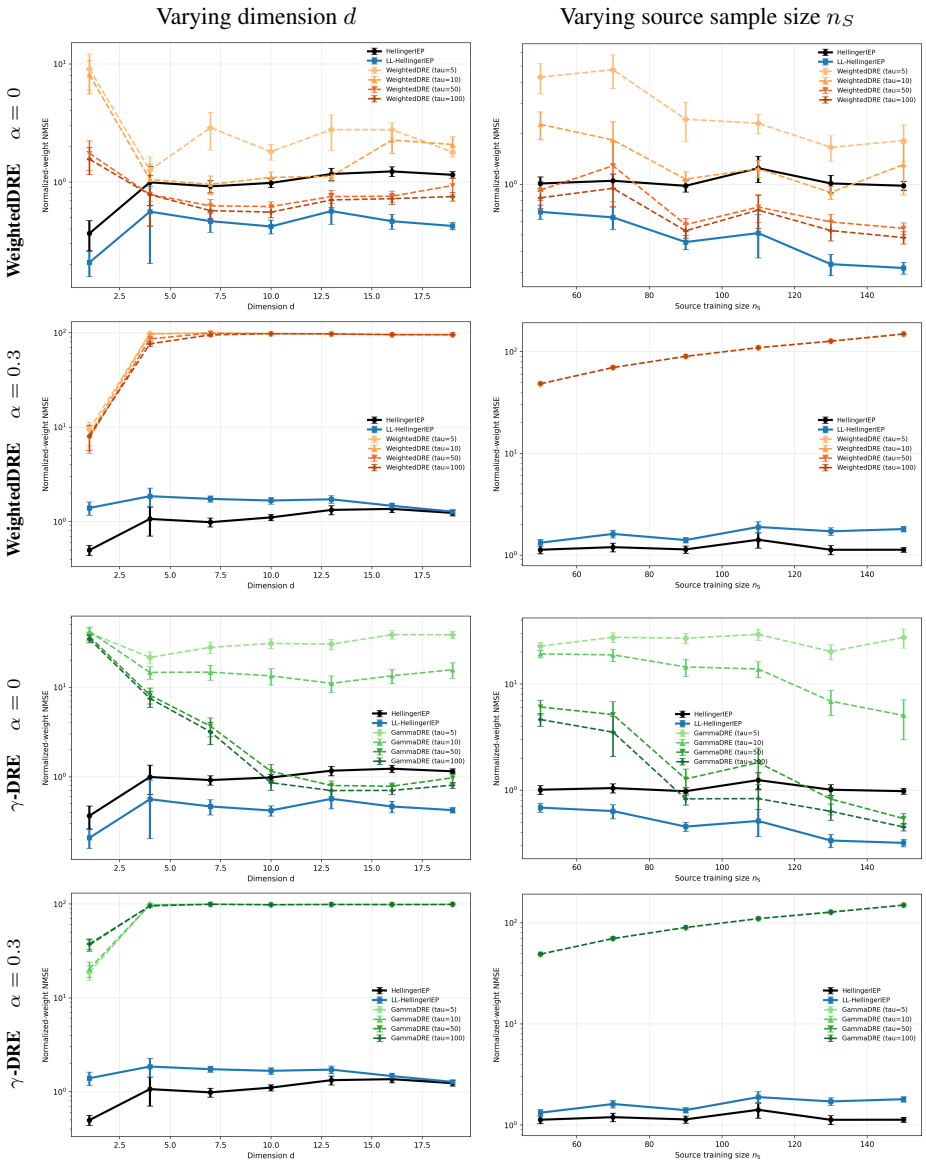


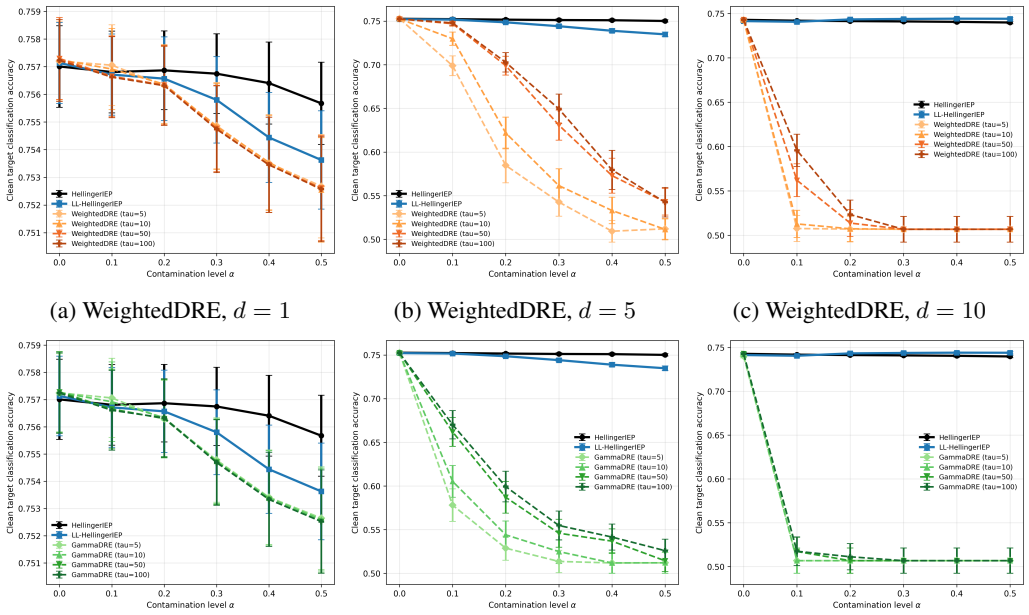
Figure 8: Density-ratio estimation NMSE for HellingerIEP, LL-HellingerIEP, WeightedDRE, and γ-DRE at $\tau \in \{5, 10, 50, 100\}$, with divergence parameter $\gamma = 0.01$. Columns distinguish varying dimension and varying source sample size; rows distinguish the comparison method and contamination level.

parameter class of f-divergences. Let $r(x) = p(x)/q(x)$. Under the convention used here, the power-divergence is indexed by η and can be written as

$$D_\eta(P\|Q) = \frac{1}{\eta(\eta-1)}\mathbb{E}_Q\left[r(X)^\eta - \eta r(X) + \eta - 1\right], \qquad \eta \neq 0, 1,$$

with the cases $\eta = 0$ and $\eta = 1$ obtained by continuity. Different choices of η recover familiar divergences, including forward KL, reverse KL, and χ^2-type divergences. In this parameterization, the Hellinger divergence corresponds to the intermediate value $\eta = 1/2$, up to a constant scaling.

This perspective gives a natural ablation study: it lets us ask whether the robustness of HellingerIEP is specific to the Hellinger geometry or is simply a generic consequence of using a power-divergence objective. To examine this question, we use the same one-dimensional Gaussian density-ratio estimation setup as in the main experiment and compare IEP estimators obtained from several choices



(a) WeightedDRE, $d = 1$ (b) WeightedDRE, $d = 5$ (c) WeightedDRE, $d = 10$

(d) γ-DRE, $d = 1$ (e) γ-DRE, $d = 5$ (f) γ-DRE, $d = 10$

Figure 9: Target classification accuracy for HellingerIEP, LL-HellingerIEP, WeightedDRE, and γ-DRE under increasing target contamination. The top row compares HellingerIEP and LL-HellingerIEP with WeightedDRE, and the bottom row compares them with γ-DRE. WeightedDRE and γ-DRE use $\tau \in \{5, 10, 50, 100\}$, with the divergence parameter of γ-DRE fixed at $\gamma = 0.01$. Points denote means and error bars denote standard errors over 50 Monte Carlo replicates.

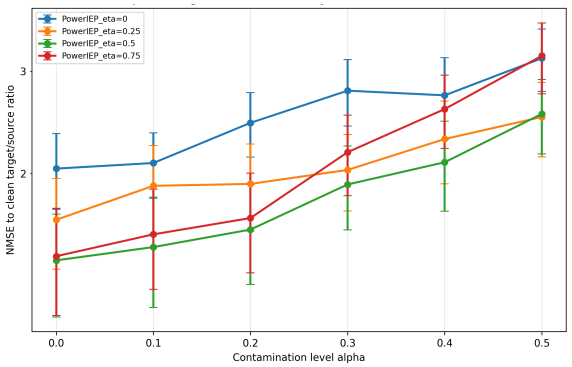


Figure 10: Comparison of power-divergence IEP estimators in the one-dimensional Gaussian density-ratio estimation experiment under target contamination. The parameter η controls the divergence geometry, with $\eta = 1/2$ corresponding to the Hellinger divergence up to constant scaling. The Hellinger choice provides competitive clean-sample performance and remains stable as target contamination increases.

of η, using a common direct kernel parameterization. Figure 10 reports the resulting NMSE under different levels of target contamination.

The results show that the choice of divergence has a substantial effect on robustness. More aggressive choices of η can improve performance under severe contamination but may sacrifice accuracy in clean samples, whereas less robust choices behave more similarly to KL-type objectives and remain sensitive to contaminated target covariates. In this experiment, the Hellinger choice $\eta = 1/2$ provides a balanced trade-off between clean-sample performance and robustness across contamination levels, motivating our focus on HellingerIEP in the main paper.

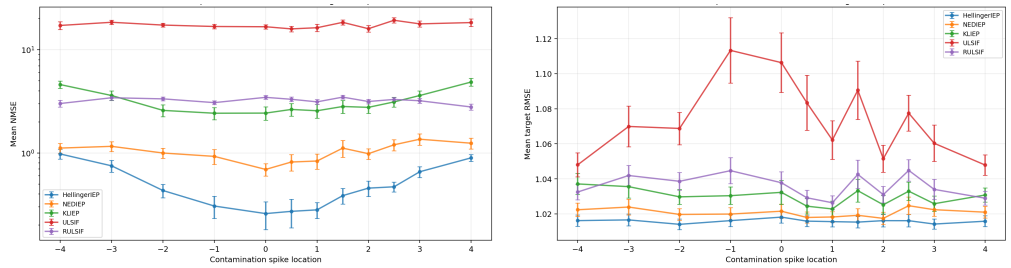
I Effect of Contamination Location

We next examine how the location of target contamination affects density-ratio estimation and downstream performance. We use the same 1D regression setting as in Section F.2, with source and target sample sizes $n_S = n_T = 100$ and a separate test set of size 2000. The target sample is contaminated by replacing 30% of the target covariates with samples from a narrow Gaussian distribution, $\mathcal{N}(\mu, 0.001^2)$, where the contamination location μ is varied from -4 to 4. For each value of μ, we evaluate density-ratio estimation accuracy using NMSE against the oracle clean weights on the source sample points, and we evaluate downstream performance using regression RMSE. In this experiment, we compare HellingerIEP, NEDIEP, KLIEP, uLSIF, and RuLSIF.

Figure 11 shows that the effect of contamination strongly depends on its location. For downstream prediction, contamination in the tails has relatively little effect, whereas contamination near regions of substantial source and target probability mass is more damaging. One possible explanation is the locality of the RBF kernel representation: contamination far from most source observations has low kernel similarity to the source sample and therefore exerts less influence on the estimated density ratios at the source points used for model training. In contrast, contamination near the central region can have a stronger effect on the fitted density ratio where the source distribution places substantial probability mass.

Across contamination locations, HellingerIEP maintains low density-ratio NMSE and downstream RMSE relative to the competing methods. These results complement the main experiments by indicating that the relative stability of HellingerIEP is not confined to a single contamination location.

Comparing the NMSE performance of HellingerIEP and NEDIEP, we see that NEDIEP is more stable when contamination occurs in low-source-density regions, whereas HellingerIEP performs better when contamination occurs in high-source-density regions. The two methods perform similarly in the downstream prediction task considered, with HellingerIEP showing a slight advantage.



(a) Density-ratio estimation NMSE versus spike contamination location.

(b) Downstream regression RMSE versus spike contamination location.

Figure 11: Effect of target contamination location on density-ratio estimation and downstream regression performance. We replace 30% of the target covariates with samples from $\mathcal{N}(\mu, 0.001^2)$ and vary μ from -4 to 4.