Proceedings of Machine Learning Research vol 313:1–34, 2026 37th International Conference on Algorithmic Learning Theory

Efficient and Provable Algorithms for Covariate Shift

Deeksha Adil DEEKSHA.ADIL@ETH-ITS.ETHZ.CH
Institute for Theoretical Studies, ETH Zürich

Jarosław Błasiok JAROSLAW.BLASIOK@UNIBOCCONI.IT
Department of Computing Sciences, Bocconi University

Editors: Matus Telgarsky and Jonathan Ullman

Abstract

Covariate shift, a widely used assumption in tackling *distributional shift* (when training and test distributions differ), focuses on scenarios where the distribution of the labels conditioned on the feature vector is the same, but the distribution of features in the training and test data are different. Despite the significance and extensive work on covariate shift, theoretical guarantees for algorithms in this domain remain sparse. In this paper, we distill the essence of the covariate shift problem and focus on estimating the average $\mathbb{E}_{\widetilde{\boldsymbol{x}}\sim p_{\text{test}}} \boldsymbol{f}(\widetilde{\boldsymbol{x}})$, of any unknown and bounded function $\boldsymbol{f}$, given labeled training samples $(\boldsymbol{x}_i, \boldsymbol{f}(\boldsymbol{x}_i))$, and unlabeled test samples $\widetilde{\boldsymbol{x}}_i$; this is a core subroutine for several widely studied learning problems. We give several efficient algorithms, with provable sample complexity and computational guarantees.

Keywords: Distributional Shift, Covariate Shift, Provable Algorithms, Sample Complexity

1. Introduction

A common assumption in several machine learning algorithms is that training and test data come from the same distribution, which does not hold in many real-world applications. *Covariate shift*, first introduced by Shimodaira (2000), is a widely used assumption for addressing such distribution shifts. It assumes that the conditional distribution of labels given the features remains the same across training and test data, i.e., $p_{\text{tr}}(\boldsymbol{y}|\boldsymbol{x}) = p_{\text{te}}(\boldsymbol{y}|\boldsymbol{x})$, while the marginal distributions of the features differ, i.e., $p_{\text{tr}}(\boldsymbol{x}) \neq p_{\text{te}}(\boldsymbol{x})$. Covariate shift has applications across various domains, including Natural Language Processing Yamada et al. (2010), Signal Processing Yamada et al. (2012), Brain-computer interfacing Li et al. (2010), Spam Filtering Bickel and Scheffer (2006b), Human activity recognition Hachiya et al. (2012), speaker identification Yamada et al. (2010), and biomedical engineering Li et al. (2010).

In this paper, we capture the core of covariate shift and introduce the following general problem, which we refer to as *covariate-shifted mean estimation*. Let $\boldsymbol{f} : \mathbb{R}^d \to [-1, 1]$ be a bounded function unknown to the algorithm. Given $\epsilon > 0$, along with n_{tr} labeled samples $(\boldsymbol{x}_i, \boldsymbol{f}(\boldsymbol{x}_i))$ for $\boldsymbol{x}_i \sim p_{\text{tr}}$ and n_{te} unlabeled samples $\widetilde{\boldsymbol{x}}_i \sim p_{\text{te}}$, the goal is to compute Z such that $|Z - \mathbb{E}_{\boldsymbol{x}\in p_{\text{te}}} \boldsymbol{f}(\boldsymbol{x})| \leq \epsilon$. The question is: for a target accuracy ε, how many samples n_{tr}, n_{te} are sufficient to achieve this guarantee? Notably, if both p_{tr} and p_{te} were known exactly, solving this problem would be significantly easier. In this case

$$\mathbb{E}_{\boldsymbol{x}\sim p_{\text{tr}}} \frac{p_{\text{te}}(\boldsymbol{x})}{p_{\text{tr}}(\boldsymbol{x})} \boldsymbol{f}(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{te}}} \boldsymbol{f}(\boldsymbol{x}),$$

© 2026 D. Adil & J. Błasiok.

providing direct access to an unbiased estimator of the desired quantity. Moreover, if the ratio $p_{\mathrm{te}}(\boldsymbol{x})/p_{\mathrm{tr}}(\boldsymbol{x})$ is bounded for a typical $\boldsymbol{x} \sim p_{\mathrm{tr}}$, we can control the variance of this estimator[1], and therefore estimate $\mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{te}}} \boldsymbol{f}(\boldsymbol{x})$ from a finite sample drawn from p_{tr}. As a result, most of the interest is focused on approximately learning the density ratio $p_{\mathrm{te}}(\boldsymbol{x})/p_{\mathrm{tr}}(\boldsymbol{x})$, also called *importance weights*. Independently, there is a rich body of work on importance weight estimation in both the theoretical and applied literature, for example, Cortes et al. (2010); Wen et al. (2015); Hachiya et al. (2012); Gopalan et al. (2022); Sugiyama et al. (2012b); see also Sugiyama et al. (2012a) for a comprehensive review. However, surprisingly, relatively few studies attempt to address the following two fundamental questions systematically.

1. What approximation guarantees can be obtained for a specific algorithm attempting to estimate the density ratio with a given number of samples?

2. What kind of approximation guarantees do we need for the estimated density ratio to serve as a good proxy of the true density ratio for covariate shift?

Since the covariate-shifted mean estimation problem cannot be efficiently solved in full generality, certain regularity conditions must be imposed — either on the unknown function $\boldsymbol{f}$ or on the unknown densities $p_{\mathrm{tr}}, p_{\mathrm{te}}$. Most prior work has focused on the development of algorithms under specific restrictions on $\boldsymbol{f}$, with the most significant theoretical results based on *kernel mean matching*. In this work, we primarily focus on the latter approach, where we assume constraints on the densities instead.

1.1. Related Works

Before presenting our results, we summarize the current state of theoretical results in this space.

Kernel Mean Matching based Approaches The most popular algorithm for the covariate shift problem is the celebrated *Kernel Mean Matching* (KMM), introduced by Huang et al. (2006) — one of the few results in this space with end-to-end theoretical guarantees. We discuss it in more detail here, as it shares a superficial resemblance to our most general result due to its foundation in *Reproducing Kernel Hilbert Space* (RKHS) theory.

Consider an RKHS $\mathcal{H}$ with a reproducing kernel map Φ (see Section 5.1 for a brief overview of RKHS notation), where the kernel is bounded such that $\|\Phi_{\boldsymbol{x}}\|_{\mathcal{H}} \leq 1$ for all $\boldsymbol{x}$. The standard analysis of KMM assumes that $\boldsymbol{f} \in \mathcal{H}$ is bounded in the $\mathcal{H}$ norm, i.e., $\|\boldsymbol{f}\|_{\mathcal{H}} \leq M$, and the density ratio satisfies $p_{\mathrm{te}}(\boldsymbol{x})/p_{\mathrm{tr}}(\boldsymbol{x}) \leq B$. Theorem 1 of Yu and Szepesvári (2012) shows that, under these conditions, collecting $n_{tr} \gtrsim (MB/\varepsilon)^2$ samples from p_{tr} and $n_{te} \gtrsim (M/\varepsilon)^2$ samples from p_{te} suffices to estimate $\mathbb{E}_{\tilde{\boldsymbol{x}}\sim p_{\mathrm{te}}} \boldsymbol{f}(\tilde{\boldsymbol{x}})$ within an error of ε. More detailed description of the algorithm can be found in Appendix F.

We further observe that if $\boldsymbol{f}$ belongs to a function class $\mathcal{F}$ that allows it to be learned with arbitrarily small error — i.e., one can find $\hat{\boldsymbol{f}}$ such that, $\mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{tr}}}|\boldsymbol{f}(\boldsymbol{x}) - \hat{\boldsymbol{f}}(\boldsymbol{x})| \leq \varepsilon'$, with respect to p_{tr} — there is a naive approach to solve the covariate-shifted mean estimation problem. Specifically, one can first approximate $\boldsymbol{f}$ with $\hat{\boldsymbol{f}}$ up to an error $\varepsilon' \ll \varepsilon$, and use $\hat{\boldsymbol{f}}$ in place of $\boldsymbol{f}$ when estimating $\mathbb{E}_{\tilde{\boldsymbol{x}}\sim p_{\mathrm{te}}} \boldsymbol{f}(\tilde{\boldsymbol{x}})$.

1. A variant of this condition is necessary. For instance, if p_{te} and p_{tr} have disjoint support, estimating $\mathbb{E}_{\boldsymbol{x}\in p_{\mathrm{te}}} \boldsymbol{f}(\boldsymbol{x})$ is impossible, even if we know both densities p_{te} and p_{tr} exactly since we have no information about $\boldsymbol{f}$ on the support of p_{te}.

In this case, we obtain:

$$\left|\mathbb{E}_{\tilde{\boldsymbol{x}}\sim p_{\mathrm{te}}}\hat{\boldsymbol{f}}(\tilde{\boldsymbol{x}}) - \mathbb{E}_{\tilde{\boldsymbol{x}}\sim p_{\mathrm{te}}}\boldsymbol{f}(\tilde{\boldsymbol{x}})\right| \leq \mathbb{E}_{\tilde{\boldsymbol{x}}\sim p_{\mathrm{te}}}|\hat{\boldsymbol{f}}(\tilde{\boldsymbol{x}}) - \boldsymbol{f}(\tilde{\boldsymbol{x}})| \leq B\,\mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{tr}}}|\hat{\boldsymbol{f}}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x})| \leq B\varepsilon'.$$

Thus, setting $\varepsilon' := \varepsilon/B$ ensures the desired accuracy. The estimate $\hat{\boldsymbol{f}}$ can be found using L_1 kernel regression with $n_{tr} \approx M^2/\varepsilon'^2 = (MB/\varepsilon)^2$ samples. Then $n_{te} \approx M^2/\varepsilon^2$ samples suffice to estimate $\mathbb{E}_{\tilde{\boldsymbol{x}}\sim p_{\mathrm{te}}}\hat{\boldsymbol{f}}(\boldsymbol{x})$ within an error of ε. See Appendix F for more details.

This highlights that the standard analysis of the KMM algorithm in a simple realizable scenario falls short of providing an asymptotic improvement in sample complexity over the naive plug-in method. Therefore, a more fine-grained analysis is necessary to justify the widely observed superior empirical performance of KMM.

Importance Weights Estimation via Classification A common approach to estimating the density ratio is to reduce it to a classification problem. Specifically, consider a distribution over pairs $(\boldsymbol{x}, y)$ where $y \in \{0, 1\}$ is uniform, and $\boldsymbol{x}$ is drawn from p_{tr} when $y = 0$, and from p_{te} when $y = 1$. The Bayes-optimal classifier for this problem outputs $\Pr[y = 1|\boldsymbol{x}]$, which can be transformed into the density ratio $p_{\mathrm{te}}(\boldsymbol{x})/p_{\mathrm{tr}}(\boldsymbol{x})$.

Given the extensive literature on classification, one can train a model to approximate $\Pr[y = 1|\boldsymbol{x}]$ and then apply the same transformation to obtain an estimate of the density ratio. This estimated ratio can, in turn, serve as importance weights for solving the covariate-shifted mean estimation problem.

Logistic regression is a widely used method for estimating $\Pr[y = 1|\boldsymbol{x}]$ in this setting, while the use of more powerful models, such as *kernel logistic regression* has also been explored Bickel et al. (2009); Sugiyama et al. (2012b).

Despite the popularity of this approach to handle the covariate shift, no end-to-end theoretical guarantees are known. In this work, we bridge this gap by providing sample complexity and approximation guarantees for covariate-shifted mean estimation using logistic regression and kernel logistic regression under specific model assumptions.

Separately Estimating Both Densities A widely held belief in importance weight estimation is that separately learning the densities p_{tr} and p_{te}, and then using their ratio, $\widehat{p_{\mathrm{te}}}/\widehat{p_{\mathrm{tr}}}$, as an estimate of $p_{\mathrm{te}}/p_{\mathrm{tr}}$ is both inefficient — since density estimation in high-dimensional spaces often requires exponentially many samples — and, more importantly, insufficient (see, for example, Section 2.1 in Huang et al. (2006), or Section 2.2 in Yu and Szepesvári (2012)).

This intuition is natural, as even a small error in estimating the denominator p_{tr}, can lead to a significant error in the density ratio. Consequently, it remains unclear whether, even in the situation where we could obtain estimates for p_{tr} and p_{te} within ε *total variation* (TV) distance in time polynomial in $1/\varepsilon$, this would be helpful in solving the covariate shift problem.

Surprisingly, we show that with only a polynomial increase in sample complexity, it is indeed possible to solve the covariate-shifted mean estimation problem for any pair of distributions within an efficiently learnable class – provided that $p_{\mathrm{te}}(\boldsymbol{x})/p_{\mathrm{tr}}(\boldsymbol{x})$ remains reasonably bounded for most $\boldsymbol{x} \sim p_{\mathrm{tr}}$. There is a rich body of work in distribution learning theory demonstrating that many distribution families can be efficiently learned, such as mixtures of Gaussians Moitra and Valiant (2010); Liu and Li (2022), low-dimensional log-concave distributions Diakonikolas et al. (2017), or graphical models with bounded treewidth Narasimhan and Bilmes (2004).

RHKS based density ratio estimations A recent work on the density ratio estimation essentially assumes that the density ratio $p_{\mathrm{te}}/p_{\mathrm{tr}}$ is bounded in some fixed RKHS (i.e. concrete strong regularity

condition on the density ratio) and provides an algorithm with a very strong recovery guarantee in terms of the RKHS norm of the error $\|\hat{r} - p_{\mathrm{tr}}/p_{\mathrm{te}}\|_{\mathcal{H}}$ Nguyen et al. (2024). For bounded kernels, this implies a pointwise bound, and it is not difficult to see that such a strong guarantee for density ratio estimation can be translated into guarantees for covariate-shifted mean estimation with arbitrary bounded $\boldsymbol{f}$. Those bounds are similar in spirit to our Theorem 16; but we instead assume that log-density ratio belongs to RKHS, and provide directly a guarantee for covariate-shift problem. For example, their assumptions, in contrast to ours, do not easily capture the case where p_{tr} and p_{te} are slightly shifted spherical Gaussians.

1.2. Our Results

In this paper, we introduce several algorithms for the covariate-shifted mean estimation problem, providing theoretical guarantees on approximation bounds and sample complexity. Our first result is an algorithm with small sample complexity when both p_{tr} and p_{te} are multivariate d-dimensional Gaussian distributions. Surprisingly, this is the first work to establish formal bounds for covariate shift, just in the case of Gaussian distributions

Theorem 1 (Informal Statement of Theorem 5) *Let p_{tr} and p_{te} be d-dimensional Gaussian distributions which are close to each other. There is an algorithm (Algorithm 1), such that for any $\boldsymbol{f}$ satisfying $\sup_{\boldsymbol{x}\in\mathbb{R}^d}|\boldsymbol{f}(\boldsymbol{x})|\leq 1$, requires at most $O\left(d^2/\varepsilon^2 \log\frac{1}{\delta}\right)$ samples of $(\boldsymbol{x}_i, \boldsymbol{f}(\boldsymbol{x}_i)), \boldsymbol{x}_i \sim p_{\mathrm{tr}}$ and $\widetilde{\boldsymbol{x}}_i \sim p_{\mathrm{te}}$, and returns Z such that with probability at least $1-\delta$,*

$$|Z - \mathbb{E}_{\boldsymbol{x}\in p_{\mathrm{te}}} \boldsymbol{f}(\boldsymbol{x})|\leq \varepsilon.$$

In the case when $p_{\mathrm{te}}, p_{\mathrm{tr}}$ are isotropic Gaussians, $O\left(d/\varepsilon^2 \log(1/\delta)\right)$ suffice (Algorithm 2).

Our algorithms indicate that it is sufficient to learn the training and test Gaussian distributions to within an ε *total variation distance*. Given such estimates, $\widehat{p_{\mathrm{tr}}}$ and $\widehat{p_{\mathrm{te}}}$, the estimator $\frac{\widehat{p_{\mathrm{te}}}}{\widehat{p_{\mathrm{te}}}}\boldsymbol{f}$ is good enough for our purposes.

Next, we extend our approach to a broader class of probability distributions. Specifically, we show that if p_{tr} and p_{te} belong to any efficiently learnable class of distributions, the covariate-shifted mean estimation problem can be solved with polynomially many samples by learning both distributions individually.

Theorem 2 (Informal Statement of Theorem 10) *Let $\boldsymbol{f}$ be a bounded function, $p_{\mathrm{tr}}, p_{\mathrm{te}}$ be a pair of distributions, and let $B > 0$, be such that $\Pr_{\boldsymbol{x}\sim p_{\mathrm{te}}}(p_{\mathrm{te}}(\boldsymbol{x})/p_{\mathrm{tr}}(\boldsymbol{x}) > B/4) \leq \varepsilon$. Let $\widehat{r}$ be s.t. $\widehat{r} = \widehat{p_{\mathrm{te}}}/\widehat{p_{\mathrm{tr}}}$ for some $\widehat{p_{\mathrm{te}}}$, and $\widehat{p_{\mathrm{tr}}}$ satisfying $d_{TV}(p_{\mathrm{te}}, \widehat{p_{\mathrm{te}}}) \leq \varepsilon$ and $d_{TV}(p_{\mathrm{tr}}, \widehat{p_{\mathrm{tr}}}) \leq \varepsilon/B$.*

Then given access to $\hat{r}$ and using $O(B^2/\varepsilon^2 \log(1/\delta))$ samples samples $(\boldsymbol{x}_i, \boldsymbol{f}(\boldsymbol{x}_i))$, $\boldsymbol{x}_i \sim p_{\mathrm{tr}}$ we can return Z (Algorithm 3) such that, with probability at least $1-\delta$,

$$|Z - \mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{te}}} \boldsymbol{f}(\boldsymbol{x})|\leq O(\varepsilon).$$

Our algorithm again employs the estimator $\frac{\widehat{p_{\mathrm{te}}}}{\widehat{p_{\mathrm{tr}}}}\boldsymbol{f}$ where $\widehat{p_{\mathrm{te}}}, \widehat{p_{\mathrm{tr}}}$, are approximations of p_{te} and p_{tr} respectively — except of dismissing all samples for which $\frac{\widehat{p_{\mathrm{te}}}}{\widehat{p_{\mathrm{tr}}}}$ exceeds a threshold B. This challenges the common belief that separately estimating both densities is insufficient for accurately estimating the density ratio. We prove that when the tail of the density ratio is bounded, independently estimating the training and test distributions to a small error is sufficient to solve the covariate-shifted

mean estimation problem. While Algorithm 3 applies to a broad class of distributions, for Gaussian distributions, Algorithm 1 still achieves better sample complexity (Theorem 1).

Our remaining generalizations build on Algorithm 3 and Theorem 2, focusing on efficiently computing the approximate ratio $\widehat{r}$. As previously mentioned, logistic regression is a popular method for estimating such quantities, yet no theoretical guarantees exist for its performance. We prove that when p_{te} and p_{tr} belong to the same exponential family, logistic regression successfully computes the required $\widehat{r}$. Thus, we provide the first theoretical analysis of logistic regression-based density estimation methods. Our analysis, combined with Theorem 2, leads to the following result.

Theorem 3 (Informal Statement of Theorem 13) *Assume that p_{tr} and p_{te} belong to the same exponential family parameterized by $\theta \in \mathbb{R}^D$. If the distance between p_{te} and p_{tr} is at most B, we can solve the covariate-shifted mean estimation problem using at most $O(B^4 D\varepsilon^{-8})$ samples.*

Finally, we present our most general result for the covariate-shifted mean estimation problem. We prove that if $\ln(p_{\text{tr}}/p_{\text{te}})$ belongs to a *Reproducing Kernel Hilbert Space*, the problem can be solved with polynomial sample complexity using an efficient algorithm. Specifically, we show that *kernel logistic regression* can be used to obtain the required estimate $\widehat{r}$, which can then be used in Algorithm 3 to obtain the required solution. Notably, our result only assumes regularity of the density ratio $p_{\text{tr}}/p_{\text{te}}$ with either of these densities can potentially be pathological by itself.

Theorem 4 (Informal Statement of Theorem 16) *Let $\mathcal{H}$ be a RKHS of functions over U, and $p_{\text{tr}}, p_{\text{te}}$ a pair of distributions over U, such that $\ln(p_{\text{tr}}(\boldsymbol{x})/p_{\text{te}}(\boldsymbol{x})) \in \mathcal{H}$. Assume that the kernel K associated with $\mathcal{H}$ is bounded by D and the distance between p_{tr} and p_{te} is at most B. We can then solve the covariate-shifted mean estimation problem using at most $O(B^4 D\varepsilon^{-8})$ samples.*

Notation. We use *high-probability bounds* to refer to sample complexity bounds that are true with probability at least $1-\delta$ with the dependence of δ on the sample complexity is $\log \frac{1}{\delta}$. We use $d_{TV}(p,q)$ to denote the total variation distance between distributions p and q. The notation $\text{supp}(p)$ is used to denote the support of the distribution p. We define for $p \geq 1$, the discrepancy metric between two probability distributions as

$$R_p(p_1||p_2) := \mathbb{E}_{\boldsymbol{x}\sim p_2}\left(\frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})}\right)^p.$$

Note that this is closely related with the Renyi divergence between p_1 and p_2, that is $D_p(p_1||p_2) = \frac{1}{p-1}\exp(R_p(p_1||p_2)$. The quantity R_p will be more convenient for us.

2. Covariate Shift for Gaussian Distributions

In this section, we present our results for Gaussian distributions. We prove that it is sufficient to learn the means and variances of the training and test distributions to within an ε error in order to solve the covariate-shifted mean estimation problem with ε accuracy. In particular, we prove,

Theorem 5 *Let $p_{\text{tr}} = \mathcal{N}(\boldsymbol{\mu}_{\text{tr}}, \boldsymbol{\Sigma}_{\text{tr}})$ and $p_{\text{te}} = \mathcal{N}(\boldsymbol{\mu}_{\text{te}}, \boldsymbol{\Sigma}_{\text{te}})$, where $\|\boldsymbol{\mu}_{\text{tr}} - \boldsymbol{\mu}_{\text{te}}\|_2 \leq O(1)$ and $\|\boldsymbol{\Sigma}_{\text{tr}}^{-1} - \boldsymbol{\Sigma}_{\text{te}}^{-1}\|_F \leq O(1)$. Also, let $\|\boldsymbol{\mu}_{\text{tr}}\|_2 \leq O(1)$, $\|\boldsymbol{\Sigma}_{\text{tr}}^{-1}\|_{op} \leq O(1)$. For any $\boldsymbol{f}$ such that $\sup_{\boldsymbol{x}\in\mathbb{R}^d}|\boldsymbol{f}(\boldsymbol{x})| \leq 1$, Algorithm 1 returns Z such that with probability at least $1-\delta$,*

$$|Z - \mathbb{E}_{\boldsymbol{x}\in p_{\text{te}}}\boldsymbol{f}(\boldsymbol{x})| \leq \varepsilon. \tag{1}$$

Furthermore, Algorithm 1 requires sampling at most $O\left(\frac{d^2}{\varepsilon^2}\log\frac{1}{\delta}\right)$ *samples of* $(\boldsymbol{x}_i, \boldsymbol{f}(\boldsymbol{x}_i)), \boldsymbol{x}_i \sim p_{\mathrm{tr}}$ *and* $\widetilde{\boldsymbol{x}}_i \sim p_{\mathrm{te}}$.

For isotropic Gaussians, we can prove a tighter bound on the sample complexity.

Theorem 6 *Let* $p_{\mathrm{tr}} = \mathcal{N}(\boldsymbol{\mu}_{\mathrm{tr}}, \boldsymbol{I})$ *and* $p_{\mathrm{te}} = \mathcal{N}(\boldsymbol{\mu}_{\mathrm{te}}, \boldsymbol{I})$, *where* $\|\boldsymbol{\mu}_{\mathrm{tr}} - \boldsymbol{\mu}_{\mathrm{te}}\|_2 \leq O(1)$. *For any* $\boldsymbol{f}$ *such that* $\sup_{\boldsymbol{x}\in\mathbb{R}^d}|\boldsymbol{f}(\boldsymbol{x})| \leq 1$, *Algorithm 2 returns* Z *such that with probability at least* $1-\delta$,

$$|Z - \mathbb{E}_{\boldsymbol{x}\in p_{\mathrm{te}}} \boldsymbol{f}(\boldsymbol{x})| \leq \varepsilon. \tag{2}$$

Furthermore, Algorithm 2 requires sampling at most $O\left(\frac{d}{\varepsilon^2}\log\frac{1}{\delta}\right)$ *samples of* $(\boldsymbol{x}_i, \boldsymbol{f}(\boldsymbol{x}_i)), \boldsymbol{x}_i \sim p_{\mathrm{tr}}$ *and* $\widetilde{\boldsymbol{x}}_i \sim p_{\mathrm{te}}$.

Algorithm 1: Covariate shift for General Gaussian p_{te} and p_{tr}

Input: d, ε, δ
$K = O\left(\frac{d^2}{\varepsilon^2}\log\frac{1}{\delta}\right)$
Generate K samples $\boldsymbol{x}_i$ from p_{tr} and K samples $\widetilde{\boldsymbol{x}}_i$ from p_{te}.
Compute $\widehat{\boldsymbol{\mu}_{\mathrm{tr}}} = \frac{2}{K}\sum_{i\in[K/2]} \boldsymbol{x}_i$
Compute $\widehat{\boldsymbol{\mu}_{\mathrm{te}}} = \frac{2}{K}\sum_{i\in[K/2]} \widetilde{\boldsymbol{x}}_i$
Compute $\widehat{\boldsymbol{\Sigma}_{\mathrm{tr}}} = \frac{1}{K}\sum_{i\in[K/2]}(\boldsymbol{x}_{2i} - \boldsymbol{x}_{2i-1})(\boldsymbol{x}_{2i} - \boldsymbol{x}_{2i-1})^\top$
Compute $\widehat{\boldsymbol{\Sigma}_{\mathrm{te}}} = \frac{1}{K}\sum_{i\in[K/2]}(\widetilde{\boldsymbol{x}}_{2i} - \widetilde{\boldsymbol{x}}_{2i-1})(\widetilde{\boldsymbol{x}}_{2i} - \widetilde{\boldsymbol{x}}_{2i-1})^\top$
$\widehat{p_{\mathrm{tr}}} = \mathcal{N}(\widehat{\boldsymbol{\mu}_{\mathrm{tr}}}, \widehat{\boldsymbol{\Sigma}_{\mathrm{tr}}}), \widehat{p_{\mathrm{te}}} = \mathcal{N}(\widehat{\boldsymbol{\mu}_{\mathrm{te}}}, \widehat{\boldsymbol{\Sigma}_{\mathrm{te}}})$
for $t = 1, \cdots, O(\log(1/\delta))$ **do**
  Generate $m = O\left(\varepsilon^{-2}\right)$ samples $(\boldsymbol{y}_i, \boldsymbol{f}(\boldsymbol{y}_i))$, with $\boldsymbol{y}_i \sim p_{\mathrm{tr}}$.
  for $i \in [m]$ **do**
    $Z_i = \frac{\widehat{p_{\mathrm{te}}}(\boldsymbol{y}_i)}{\widehat{p_{\mathrm{tr}}}(\boldsymbol{y}_i)}\boldsymbol{f}(\boldsymbol{y}_i)$
  end for
  $\overline{Z}_t = \frac{1}{m}\sum_{i=1}^m Z_i$
end for
Return: Median of $\overline{Z}_t$'s

Our analysis proves that the estimator $X = \frac{\widehat{p_{\mathrm{te}}}}{\widehat{p_{\mathrm{tr}}}}\boldsymbol{f}$ is a good estimator. Note that this estimator is biased. We divide the analysis into two parts: in the first part, we show that our estimator has a small bias, and in the second part, we provide a bound on its variance. Finally, we apply concentration bounds to establish the total sample complexity.

For the first part, we prove the following result. The proofs of these lemmas are deferred to Appendix B.

Lemma 7 (Bound for Isotropic Gaussian Distributions) *Let* $p_{\mathrm{tr}} = \mathcal{N}(\boldsymbol{\mu}_{\mathrm{tr}}, \boldsymbol{I})$ *and* $p_{\mathrm{te}} = \mathcal{N}(\boldsymbol{\mu}_{\mathrm{te}}, \boldsymbol{I})$ *such that* $\|\boldsymbol{\mu}_{\mathrm{tr}}\|_2 \leq O(1)$ *and* $\|\boldsymbol{\mu}_{\mathrm{tr}} - \boldsymbol{\mu}_{\mathrm{te}}\|_2 \leq B \leq O(1)$. *Further, let* $\widehat{p_{\mathrm{tr}}}$ *and* $\widehat{p_{\mathrm{te}}}$ *be as defined in Algorithm 1. Then for any function* $\boldsymbol{f}$ *such that* $\sum_{\boldsymbol{x}\in\mathbb{R}^d}|\boldsymbol{f}(\boldsymbol{x})| \leq 1$,

$$\left|\mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{tr}}} \frac{\widehat{p_{\mathrm{te}}}(\boldsymbol{x})}{\widehat{p_{\mathrm{tr}}}(\boldsymbol{x})}\boldsymbol{f}(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{te}}} \boldsymbol{f}(\boldsymbol{x})\right| \leq O(\varepsilon).$$

Algorithm 2: Covariate shift for Isotropic Gaussian p_{te} and p_{tr}

Input: d, ε, δ
$K = O\left(\frac{d}{\varepsilon^2} \log \frac{1}{\delta}\right)$
Generate K samples $\boldsymbol{x}_i$ from p_{tr} and K samples $\widetilde{\boldsymbol{x}}_i$ from p_{te}.
Compute $\widehat{\boldsymbol{\mu}_{\text{tr}}} = \frac{2}{K} \sum_{i \in [K/2]} \boldsymbol{x}_i$
Compute $\widehat{\boldsymbol{\mu}_{\text{te}}} = \frac{2}{K} \sum_{i \in [K/2]} \widetilde{\boldsymbol{x}}_i$
$\widehat{p_{\text{tr}}} = \mathcal{N}(\widehat{\boldsymbol{\mu}_{\text{tr}}}, \boldsymbol{I}), \widehat{p_{\text{te}}} = \mathcal{N}(\widehat{\boldsymbol{\mu}_{\text{te}}}, \boldsymbol{I})$
for $t = 1, \cdots, O(\log(1/\delta))$ **do**
  Generate $m = O\left(\varepsilon^{-2}\right)$ samples $(\boldsymbol{y}_i, \boldsymbol{f}(\boldsymbol{y}_i))$, with $\boldsymbol{y}_i \sim p_{\text{tr}}$.
  for $i \in [m]$ **do**
    $Z_i = \frac{\widehat{p_{\text{te}}}(\boldsymbol{y}_i)}{\widehat{p_{\text{tr}}}(\boldsymbol{y}_i)} \boldsymbol{f}(\boldsymbol{y}_i)$
  end for
  $\overline{Z}_t = \frac{1}{m} \sum_{i=1}^{m} Z_i$
end for
Return: Median of $\overline{Z}_t$'s

Lemma 8 (Bound for General Gaussian Distributions) *Let* $p_{\text{te}} := \mathcal{N}(\boldsymbol{\mu}_{\text{te}}, \boldsymbol{\Sigma}_{\text{te}}), p_{\text{tr}} := \mathcal{N}(\boldsymbol{\mu}_{\text{tr}}, \boldsymbol{\Sigma}_{\text{tr}})$ *and* $\widehat{p_{\text{tr}}} = \mathcal{N}(\widehat{\boldsymbol{\mu}_{\text{tr}}}, \widehat{\boldsymbol{\Sigma}_{\text{tr}}}), \widehat{p_{\text{te}}} = \mathcal{N}(\widehat{\boldsymbol{\mu}_{\text{te}}}, \widehat{\boldsymbol{\Sigma}_{\text{te}}})$ *be as defined in Algorithm 1. If* $\|\boldsymbol{\mu}_{\text{tr}} - \boldsymbol{\mu}_{\text{te}}\|_2 \leq B \leq O(1)$, $\|\boldsymbol{\Sigma}_{\text{tr}}^{-1}\|_{op} \leq O(1)$ *then, for any function* $\boldsymbol{f}$*, such that* $\sup_{\boldsymbol{x} \in \mathbb{R}^d} |\boldsymbol{f}(\boldsymbol{x})| \leq 1$,

$$\left| \mathbb{E}_{\boldsymbol{x} \sim p_{\text{tr}}} \frac{\widehat{p_{\text{te}}}(\boldsymbol{x})}{\widehat{p_{\text{tr}}}(\boldsymbol{x})} \boldsymbol{f}(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{x} \sim p_{\text{te}}} \boldsymbol{f}(\boldsymbol{x}) \right| \leq O(\varepsilon).$$

In the next part, we show that the variance of our estimator is small. The proof is again deferred to Appendix B.

Lemma 9 (Variance of our Estimator) *If* $\|\boldsymbol{\Sigma}_{\text{tr}}^{-1}\|_{op}, \|\boldsymbol{\Sigma}_{\text{te}}^{-1}\|_{op} \leq O(1)$, $\|\boldsymbol{\Sigma}_{\text{tr}}^{-1} - \boldsymbol{\Sigma}_{\text{te}}^{-1}\|_F \leq O(1)$, *and* $\|\boldsymbol{\mu}_{\text{tr}}\|_2, \|\boldsymbol{\mu}_{\text{tr}} - \boldsymbol{\mu}_{\text{te}}\|_2 \leq O(1)$*, then, the variance of our estimator can be bounded as,*

$$\mathbb{E}_{\boldsymbol{x} \sim p_{\text{tr}}} \left[\left(\frac{\widehat{p_{\text{te}}}(\boldsymbol{x})}{\widehat{p_{\text{tr}}}(\boldsymbol{x})} \boldsymbol{f}(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{x} \sim p_{\text{te}}} \boldsymbol{f}(\boldsymbol{x}) \right)^2 \right] \leq O(1).$$

Having established that our estimator has a small bias, i.e., $O(\epsilon)$, and a small variance, $O(1)$, the result follows directly from standard Chebyshev's inequality.

For the high-probability statement, we apply the standard technique: given an estimator Z that lies within the desired interval I with probability $2/3$, we can take the median of $O(\log(1/\delta))$ independent realizations of this estimator to reduce the failure probability to δ, using the Chernoff bound.

3. TV Learnability implies Covariate Shift

In this section, we extend our algorithms to more general distributions. We show that, under mild conditions on the training and test distributions, it is sufficient to learn both distributions to a small total variation distance. In particular, we prove the following:

Theorem 10 *Let $\boldsymbol{f}$ be a bounded function, $p_{\mathrm{tr}}, p_{\mathrm{te}}$ be a pair of distributions, such that for $B > 0$, $\Pr_{\boldsymbol{x}\sim p_{\mathrm{te}}}(p_{\mathrm{te}}(\boldsymbol{x})/p_{\mathrm{tr}}(\boldsymbol{x}) > B/4) \le \varepsilon$. Let $\widehat{r}$, be s.t. $\widehat{r} = \widehat{p_{\mathrm{te}}}/\widehat{p_{\mathrm{tr}}}$ for some $\widehat{p_{\mathrm{te}}}$, and $\widehat{p_{\mathrm{tr}}}$ satisfying $d_{TV}(p_{\mathrm{te}}, \widehat{p_{\mathrm{te}}}) \le \varepsilon$ and $d_{TV}(p_{\mathrm{tr}}, \widehat{p_{\mathrm{tr}}}) \le \varepsilon/B$. Then Algorithm 3 returns Z such that, with probability at least 0.9, we get*

$$|Z - \mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{te}}} \boldsymbol{f}(\boldsymbol{x})| \le O(\varepsilon).$$

Furthermore, Algorithm 3 uses only $O(B^2/\varepsilon^2)$ samples $(\boldsymbol{x}_i, \boldsymbol{f}(\boldsymbol{x}_i))$, $\boldsymbol{x}_i \sim p_{\mathrm{tr}}$.

Algorithm 3: Covariate shift for any p_{te} and p_{tr} with bounded tails

Input: $\widehat{r} = \widehat{p_{\mathrm{te}}}/\widehat{p_{\mathrm{tr}}}, B, \varepsilon$
$K = O(B^2/\varepsilon^2)$
Generate K samples $(\boldsymbol{x}_i, \boldsymbol{f}(\boldsymbol{x}_i))$ from p_{tr}.
if $\widehat{r}(\boldsymbol{x}_i) \le B$ **then**
  $Z_i = \widehat{r}(\boldsymbol{x}_i)\boldsymbol{f}(\boldsymbol{x}_i)$
else
  $Z_i = 0$
end if
Return: $Z = \frac{1}{K}\sum_{i=1}^{K} Z_i$

In the next section, we will show how to compute the density ratio $\widehat{r}$ used in Algorithm 3, satisfying the conditions of the above theorem, for certain classes of probability distributions, without explicitly estimating p_{tr} and p_{te}.

Remark 11 *This sample complexity upper bound should be contrasted with the following simple lower bound: let B be such that $\Pr_{\boldsymbol{x}\sim p_{\mathrm{te}}}(p_{\mathrm{te}}(\boldsymbol{x})/p_{\mathrm{tr}}(\boldsymbol{x}) > B) = 2\varepsilon$. Even with exact knowledge of p_{te} and p_{tr}, we require $\Omega(B/\varepsilon)$ samples to solve the covariate-shifted mean estimation problem up to error ε. A simple proof of this lower bound can be found in Appendix C.*

A relatively common assumption in the covariate shift literature, which is stronger than the one we use, is that the density ratio $p_{\mathrm{te}}(\boldsymbol{x})/p_{\mathrm{tr}}(\boldsymbol{x})$ is bounded everywhere by B. We chose to relax this assumption because it is not even satisfied for two non-equal univariate Gaussian distributions.

To put the sample complexity in Theorem 10 into context, note that we could apply it when p_{tr} and p_{te} are isotropic Gaussian distributions, as in Theorem 6, with $|\mu_1 - \mu_2| \le O(1)$. In this case, for the desired accuracy ε, we would need to set $B := (1/\varepsilon)^{1+o(1)}$ and learn an isotropic Gaussian p_{tr} up to an error of $(1/\varepsilon)^{2+o(1)}$ in total variation distance. This can be done using $O(d/\varepsilon^{4+o(1)})$ samples. While the dependency on the dimension d matches that of Theorem 5, the dependency on the accuracy ε is polynomially worse.

The following technical lemma will be crucial in our proof.

Lemma 12 *Let $p_{\mathrm{tr}}, p_{\mathrm{te}}$ be a pair of distributions, and $S \subseteq U$ be such that $\Pr_{\boldsymbol{x}\sim p_{\mathrm{te}}}(S) \ge 1 - \varepsilon$ and $\Pr_{\boldsymbol{x}\sim p_{\mathrm{tr}}}(S) \ge 1 - \varepsilon$. Let $\widehat{p_{\mathrm{tr}}}$ and $\widehat{p_{\mathrm{te}}}$ be a pair of estimations of $p_{\mathrm{tr}}, p_{\mathrm{te}}$ respectively with $\mathrm{supp}(\widehat{p_{\mathrm{te}}}) \subseteq \mathrm{supp}(\widehat{p_{\mathrm{tr}}}) \subseteq S$. Moreover, assume that $\forall \boldsymbol{x} \in S$, $\widehat{p_{\mathrm{te}}}(\boldsymbol{x})/\widehat{p_{\mathrm{tr}}}(\boldsymbol{x}) \le B$. Then*

$$\left|\mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{tr}}} \frac{\widehat{p_{\mathrm{te}}}(\boldsymbol{x})}{\widehat{p_{\mathrm{tr}}}(\boldsymbol{x})}\boldsymbol{f}(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{te}}} \boldsymbol{f}(\boldsymbol{x})\right| \le B d_{TV}(p_{\mathrm{tr}}, \widehat{p_{\mathrm{tr}}}) + d_{TV}(p_{\mathrm{te}}, \widehat{p_{\mathrm{te}}}).$$

The proof of the above is in Appendix C. We now prove the main result of the section.

4. Learning Density Ratio by Classification

In this section, we present an algorithm based on logistic regression that approximates the density ratio $\widehat{r}$ such that $\widehat{r} = \widehat{p_{\text{te}}}/\widehat{p_{\text{tr}}}$ for some pair of distributions, with the guarantees that $d_{TV}(\widehat{p_{\text{tr}}}, p_{\text{tr}}) \leq \varepsilon$ and $d_{TV}(\widehat{p_{\text{te}}}, p_{\text{te}}) \leq \varepsilon$. This holds when both p_{te} and p_{tr} belong to the same exponential family (Definition 27), i.e., have the same values of the functions $\boldsymbol{h}$ and $\boldsymbol{T}$. Moreover, the probabilities $\widehat{p_{\text{tr}}}$ and $\widehat{p_{\text{te}}}$ provide guarantees for the density ratio, i.e., $\widehat{r} = \frac{\widehat{p_{\text{te}}}}{\widehat{p_{\text{tr}}}}$ is a good approximation to $r = \frac{p_{\text{te}}}{p_{\text{tr}}}$, making it suitable for use as importance weights in the covariate shift problem (after appropriate truncation). Although logistic regression has been applied in practice for estimating density ratios in several previous works, such theoretical guarantees remained unknown. We prove the following:

Theorem 13 *Assume that p_{tr} and p_{te} belong to the same exponential family, with $p_{\text{te}}(\boldsymbol{x}) = p(\boldsymbol{x}|\boldsymbol{\theta}_{\text{te}})$ and $p_{\text{tr}}(\boldsymbol{x}) = p(\boldsymbol{x}|\boldsymbol{\theta}_{\text{tr}})$. Let $\boldsymbol{\Sigma}_{\text{tr}} = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{tr}}} \boldsymbol{T}(\boldsymbol{x})\boldsymbol{T}(\boldsymbol{x})^T$ and similarly for $\boldsymbol{\Sigma}_{\text{te}}$. Define $B_1 := \text{Tr}(\boldsymbol{\Sigma}_{\text{tr}} + \boldsymbol{\Sigma}_{\text{te}})$ and assume, in addition, that $\|\boldsymbol{\theta}_{\text{tr}} - \boldsymbol{\theta}_{\text{te}}\|_2 + |A(\boldsymbol{\theta}_{\text{tr}}) - A(\boldsymbol{\theta}_{\text{te}})| \leq B_2$.*

As long as $R_2(p_{\text{tr}}||p_{\text{te}}) \leq B_3$, we can solve the covariate-shifted mean eastimation problem using $O(B_1 B_2^2 B_3^4 \varepsilon^{-8})$ samples.

We now present the high-level idea of our approach. The details and proof of Theorem 13 is in Appendix D.

General scheme. Consider a distribution $\mathcal{D}_0$ over pairs $(\boldsymbol{x}, y)$ where $\boldsymbol{x} \in U$ and $y \in \{\pm 1\}$, s.t. $\Pr(y = 1) = 1/2$ and $\Pr(\boldsymbol{x}|y = 1) = p_{\text{te}}(\boldsymbol{x})$ and $\Pr(\boldsymbol{x}|y = -1) = p_{\text{tr}}(\boldsymbol{x})$. Then by Bayes rule

$$\Pr(y = -1|\boldsymbol{x}) = \frac{\Pr(\boldsymbol{x}|y = -1)}{\Pr(\boldsymbol{x}|y = 1) + \Pr(\boldsymbol{x}|y = -1)} = \frac{1}{1 + p_{\text{te}}(\boldsymbol{x})/p_{\text{tr}}(\boldsymbol{x})}.$$

Hence, if we can find a Bayes optimal classifier for the distribution $(\boldsymbol{x}, y)$ calculating $\Pr(y = -1|\boldsymbol{x})$, we could invert it to get a density ratio:

$$\frac{p_{\text{te}}(\boldsymbol{x})}{p_{\text{tr}}(\boldsymbol{x})} = \frac{1}{\Pr(y = -1|\boldsymbol{x})} - 1. \tag{3}$$

A known method (see for instance Bickel and Scheffer (2006a)) of estimating the ratio $p_{\text{te}}(\boldsymbol{x})/p_{\text{tr}}(\boldsymbol{x})$ is then training a classifier for the binary classification problem (to predict y given features $\boldsymbol{x}$ from distribution $\mathcal{D}_0$), and using its output in place of $\Pr(y = -1|\boldsymbol{x})$ in Equation (3).

We show that if we use the negative log-likelihood as a loss function for classification training (also known as the cross-entropy loss), to any classifier $\hat{r}(\boldsymbol{x})$ we have found, we can associate a pair of distributions $\widehat{p_{\text{te}}}$ and $\widehat{p_{\text{tr}}}$, s.t. $\widehat{p_{\text{te}}}/\widehat{p_{\text{tr}}} = \frac{1}{\hat{r}(x)} - 1$, and the regret of the classifier $\hat{r}$ (difference between the population loss of the classifier, and the population loss of the Bayes optimal classifier $\Pr_{\mathcal{D}_0}[y = -1|x]$) is an upper bound for the sum of KL-divergences $D_{KL}(p_{\text{te}}||\widehat{p_{\text{te}}}) + D_{KL}(p_{\text{tr}}||\widehat{p_{\text{tr}}})$; see Theorem 40.

In particular, whenever we can guarantee that this regret is low, we know that the estimated density ratio corresponds to the density ratio of a pair of distributions $\widehat{p_{\text{tr}}}, \widehat{p_{\text{te}}}$ that are both KL-close to the actual distributions $p_{\text{tr}}, p_{\text{te}}$. Using the Pinskers inequality, this can be translated into TV-distance guarantees, and via Theorem 10, after appropriate truncation, such a density ratio can be used as importance weight for the covariate shift problem.

Interestingly, this method never explicitly recovers the densities $\widehat{p_{\text{te}}}$ and $\widehat{p_{\text{tr}}}$ which are shown to exist — it only recovers the density ratio.

Using logistic regression. When the distributions p_{te} and p_{tr} are from the same exponential family, we can use logistic regression to find the classifier that provides the right density ratio estimate. As mentioned above, this is a commonly used method in practice, in Appendix D.1 we provide full details of how to prove the appropriate recovery guarantees for this method. As it turns out, when the p_{te} and p_{tr} are from the same exponential family, the $\Pr(y = -1|\mathbf{x})$ follows exactly the logistic model, and we can use standard regret bounds for logistic regression in a realizable setting. The rest of the argument follows as described in the previous two paragraphs.

Moreover, in Appendix D.2, we discuss that in a special case where both p_{tr} and p_{te} are spherical Gaussian distributions (with unknown means), using new results of Bach (2021); Ostrovskii and Bach (2018), we can actually obtain bounds on the *parameter recovery* in logistic regression, which translates into a proof that logistic regression can be used for covariate shifted mean estimation with better dependency on ε (i.e. $1/\varepsilon^2$ instead of $1/\varepsilon^6$) by avoiding use of Theorem 10.

5. Reproducing Kernel Hilbert Spaces — Kernel Trick for Density Estimation

We now present a generalization of logistic regression beyond exponential families.

Specifically, we consider the case where $\ln(p_{\mathrm{tr}}/p_{\mathrm{te}})$ lies in a *Reproducing Kernel Hilbert Space* $\mathcal{H}$. We show that using *kernel logistic regression* Zhu and Hastie (2005), we can find an approximation to the density ratio $p_{\mathrm{tr}}/p_{\mathrm{tr}}$, which is good enough to solve the covariate shift problem.

Before stating the main theorem, let us briefly review the theory behind RKHS. For a more detailed exposition, refer to Ghojogh et al. (2021).

5.1. Quick RKHS Preliminary

Consider a Hilbert space $\mathcal{H}$ of functions over U (not necessarily finitely dimensional), together with an inner product $\langle f, g\rangle_{\mathcal{H}}$. If for each $x \in U$ the evaluation functional $\mathrm{ev}_x(f) := f(x)$ is bounded (i.e. for each $x \in U$, there is a bound T_x, such that for each $f \in \mathcal{H}$ we have $f(x) \le T_x\|f\|_{\mathcal{H}}$), then we say that $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ is a *Reproducing Kernel Hilbert Space (RKHS)*.

Using standard functional-analytic tools, this implies the existence of a *feature map* $\Phi : U \to \mathcal{H}$ — i.e., with any point $x \in U$, we can associate a function $\Phi_x \in \mathcal{H}$ satisfying, for every $f \in \mathcal{H}$, $\langle f, \Phi_x\rangle_{\mathcal{H}} = f(x)$. We can further associate with this RKHS a kernel $K : U \times U \to \mathbb{R}$, given as $K(x_1, x_2) := \langle \Phi_{x_1}, \Phi_{x_2}\rangle_{\mathcal{H}}$. As it turns out, every such kernel is positive semidefinite, and in fact, this is a complete characterization of RKHS — every positive semidefinite kernel $K : U \times U \to \mathbb{R}$ corresponds to some RKHS on U. For a given kernel K, the corresponding feature map is $\Phi_x := K(x, \cdot)$, and the subspace of finite sums $\sum_i \lambda_i K(x_i, \cdot)$ is dense in $\mathcal{H}$.

Why are RKHS good — the Kernel trick The key tool related to RKHS is the so-called 'kernel trick' and, more specifically, the relatively simple yet powerful Representer Theorem. This theorem states that the solution to a general optimization problem over the (potentially infinitely dimensional) $\mathcal{H}$ can be found by reducing it to a finite-dimensional problem. For example, consider a sample $(x_1, y_1), (x_2, y_2), \dots (x_n, y_n)$ where $x_i \in \mathcal{U}$ and $y_i \in \mathbb{R}$.

Theorem 14 (Representer Theorem Ghojogh et al. (2021)) *Let $\mathcal{H}$ be an RKHS of functions over U, together with the feature map $\Phi : U \to \mathcal{H}$. Let $E : (U \times \mathbb{R} \times \mathbb{R})^n \to \mathbb{R}$ be an arbitrary error map, $(x_1, y_1), \dots (x_n, y_n) \in U \times \mathbb{R}$ be a sample, and $B \ge 0$ a non-negative bound. Then the minimizer $\hat{\theta}$ of*

$$\min_{\boldsymbol{\theta}\in\mathcal{H}\,:\,\|\boldsymbol{\theta}\|_{\mathcal{H}}\le B} E((x_1, y_1, \boldsymbol{\theta}(x_1)), \dots (x_n, y_n, \boldsymbol{\theta}(y_n)) \tag{4}$$

is a linear combination

$$\hat{\boldsymbol{\theta}} = \sum_i \gamma_i \Phi_{x_i}. \tag{5}$$

Kernel Logistic Regression Most of the content in this section is well-known. However, for completeness, we provide a detailed analysis of kernel logistic regression. Kernel logistic regression assumes that we observe samples $(\boldsymbol{x}_i, y_i)$ where $\boldsymbol{x}_i \in U$ is drawn from some distribution and $y_i|x_i \in \{\pm 1\}$ follows the logistic distribution,

$$\Pr[y|\boldsymbol{x}] = (1 + \exp(y\boldsymbol{\theta}^*(\boldsymbol{x})))^{-1},$$

where $\boldsymbol{\theta}^* \in \mathcal{H}$ for some RKHS $\mathcal{H}$. Let us call the joint distribution of $(\boldsymbol{x}, y)$ drawn this way as $\mathcal{D}_{\boldsymbol{\theta}^*}$. Note that as long as $\|\boldsymbol{\theta}^*\|_{\mathcal{H}} \leq B$, $\boldsymbol{\theta}^*$ is a unique minimizer of the population loss function given by negative log-likelihood,

$$\mathcal{L}_{\boldsymbol{\theta}}(\mathcal{D}_{\boldsymbol{\theta}^*}) := \mathbb{E}_{(x,y)\sim\mathcal{D}_{\boldsymbol{\theta}^*}} \log(1 + \exp(y\boldsymbol{\theta}(\boldsymbol{x})).$$

That is $\boldsymbol{\theta}_* = \arg\min_{\|\boldsymbol{\theta}\|_{\mathcal{H}} \leq B} \mathcal{L}_{\boldsymbol{\theta}}(\mathcal{D}_{\boldsymbol{\theta}^*})$. Using a standard argument based on Rademacher complexity, we bound the error between the population loss and the empirical loss for all $\|\boldsymbol{\theta}\|_{\mathcal{H}} \leq B$. Combined with the fact that empirical loss minimization for kernel logistic regression can be efficiently solved (after applying the representer theorem, it reduces to an n-dimensional convex problem). We show the following.

Lemma 15 *Using $n \approx \mathbb{E}_{\boldsymbol{x}}[K(\boldsymbol{x}, \boldsymbol{x})]B^2/\varepsilon^2$ samples from $\mathcal{D}_{\boldsymbol{\theta}^*}$ we can efficiently find $\hat{\boldsymbol{\theta}}$ satisfying regret bound*

$$\mathcal{L}_{\hat{\boldsymbol{\theta}}}(\mathcal{D}_{\boldsymbol{\theta}^*}) - \mathcal{L}_{\boldsymbol{\theta}^*}(\mathcal{D}_{\boldsymbol{\theta}^*}) \leq \varepsilon.$$

See Appendix E.1 for a full proof.

5.2. Covariate Shift via Kernel Logistic Regression

We now give the main result of this section. If the log of the density ratio of the training and test distribution lies in some RKHS $\mathcal{H}$, we can efficiently solve the covariate shift problem for these distributions. This result is obtained by combining the regret bounds from Theorem 15, which imply the KL-divergence bounds by Theorem 40, with the Pinsker inequality and Theorem 10.

Theorem 16 *Let $\mathcal{H}$ be a RKHS of functions over U, and $p_{\mathrm{tr}}, p_{\mathrm{te}}$ a pair of distributions over U, such that $\boldsymbol{\theta}^*(x) := \ln(p_{\mathrm{tr}}(x)/p_{\mathrm{te}}) \in \mathcal{H}$.*

Assume that the kernel K associated with $\mathcal{H}$ satisfies $\sqrt{\mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{tr}}} K(x,x)} + \sqrt{\mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{te}}} K(\boldsymbol{x},\boldsymbol{x})} \leq D$ for all $x \in U$, $\|\boldsymbol{\theta}^*\| \leq B_1$, and $R_2(p_{\mathrm{te}}||p_{\mathrm{tr}}) \leq B_2$.

Then, we can solve the covariate-shifted mean estimation problem from p_{tr} to p_{te} using at most $O(DB_1^2B_2^4/\varepsilon^8)$ samples.

The full proof can be found in Appendix E.2.

Again, similar to the argument that using Theorem 16 directly in the special case of univariate Gaussian distributions gives a suboptimal sample complexity, i.e., when $p_{\mathrm{tr}} = \mathcal{N}(\mu_{tr}, 1), p_{\mathrm{te}} = \mathcal{N}(\mu_{te}, 1)$ s.t. $\|\mu_1 - \mu_2\| \leq O(1)$, this result also gives a suboptimal sample complexity of $O(\varepsilon^{-8})$ instead of $O(\varepsilon^{-2})$.

Acknowledgments

DA is supported by Dr. Max Rössler, the Walter Haefner Foundation and the ETH Zürich Foundation.

References

Hassan Ashtiani, Shai Ben-David, Nicholas JA Harvey, Christopher Liaw, Abbas Mehrabian, and Yaniv Plan. Near-optimal sample complexity bounds for robust learning of gaussian mixtures via compression schemes. *Journal of the ACM (JACM)*, 67(6):1–42, 2020.

Francis Bach. Going beyond least-squares – ii : Self-concordant analysis for logistic regression, 2021.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Steffen Bickel and Tobias Scheffer. Dirichlet-enhanced spam filtering based on biased samples. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006a. URL https://proceedings.neurips.cc/paper_files/paper/2006/file/55d99a37b2e1badba7c8df4ccd506a88-Paper.pdf.

Steffen Bickel and Tobias Scheffer. Dirichlet-enhanced spam filtering based on biased samples. *Advances in neural information processing systems*, 19, 2006b.

Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10(9), 2009.

Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper_files/paper/2010/file/59c33016884a62116be975a9bb8257e3-Paper.pdf.

Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Learning multivariate log-concave distributions. In *Conference on Learning Theory*, pages 711–727. PMLR, 2017.

Benyamin Ghojogh, Ali Ghodsi, Fakhri Karray, and Mark Crowley. Reproducing kernel hilbert space, mercer's theorem, eigenfunctions, nystr\" om method, and use of kernels in machine learning: Tutorial and survey. *arXiv preprint arXiv:2106.08443*, 2021.

Parikshit Gopalan, Omer Reingold, Vatsal Sharan, and Udi Wieder. Multicalibrated partitions for importance weights. In *International Conference on Algorithmic Learning Theory*, pages 408–435. PMLR, 2022.

Hirotaka Hachiya, Masashi Sugiyama, and Naonori Ueda. Importance-weighted least-squares probabilistic classifier for covariate shift adaptation with application to human activity recognition. *Neurocomputing*, 80:93–101, 2012.

Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex Smola. Correcting sample selection bias by unlabeled data. *Advances in neural information processing systems*, 19, 2006.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

Yan Li, Hiroyuki Kambara, Yasuharu Koike, and Masashi Sugiyama. Application of covariate shift adaptation techniques in brain–computer interfaces. *IEEE Transactions on Biomedical Engineering*, 57(6):1318–1324, 2010.

Allen Liu and Jerry Li. Clustering mixtures with almost optimal separation in polynomial time. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1248–1261, 2022.

Ron Meir and Tong Zhang. Generalization error bounds for bayesian mixture algorithms. *Journal of Machine Learning Research*, 4(Oct):839–860, 2003.

Ankur Moitra and Gregory Valiant. Settling the polynomial learnability of mixtures of gaussians. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 93–102. IEEE, 2010.

Mukund Narasimhan and Jeff Bilmes. Pac-learning bounded tree-width graphical models. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, UAI '04, page 410–417, Arlington, Virginia, USA, 2004. AUAI Press. ISBN 0974903906.

Duc Hoan Nguyen, Werner Zellinger, and Sergei Pereverzyev. On regularized radon-nikodym differentiation. *Journal of Machine Learning Research*, 25(266):1–24, 2024.

Dmitrii Ostrovskii and Francis Bach. Finite-sample analysis of m-estimators using self-concordance, 2018.

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.

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*, 64:1009–1044, 2012b.

Junfeng Wen, Russell Greiner, and Dale Schuurmans. Correcting covariate shift with the frank-wolfe algorithm. In *Twenty-fourth International Joint Conference on Artificial Intelligence*, 2015.

Makoto Yamada, Masashi Sugiyama, and Tomoko Matsui. Semi-supervised speaker identification under covariate shift. *Signal Processing*, 90(8):2353–2361, 2010.

Makoto Yamada, Leonid Sigal, and Michalis Raptis. No bias left behind: Covariate shift adaptation for discriminative 3d pose estimation. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part IV 12*, pages 674–687. Springer, 2012.

Yao-Liang Yu and Csaba Szepesvári. Analysis of kernel mean matching under covariate shift. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, page 1147–1154, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851.

Ji Zhu and Trevor Hastie. Kernel logistic regression and the import vector machine. *Journal of Computational and Graphical Statistics*, 14(1):185–205, 2005. ISSN 10618600. URL http://www.jstor.org/stable/27594104.

Appendix A. Extended preliminaries

Definition 17 (Distance Measure) *For any probability distributions $p, q, \widehat{p}$, define*

$$Q_1(\hat{p}\,|\,p;\ q) := \mathbb{E}_{\boldsymbol{x}\sim q}\left|\frac{p(\boldsymbol{x})}{\widehat{p}(\boldsymbol{x})} - 1\right|,$$

and more generally

$$Q_s(\widehat{p}\,|\,p;\ q) := \left(\mathbb{E}_{\boldsymbol{x}\sim q}\left|\frac{p(\boldsymbol{x})}{\widehat{p}(\boldsymbol{x})} - 1\right|^s\right)^{1/s}.$$

By standard norm comparison for any $1 \le t_1 \le t_2$, $Q_{t_1} \le Q_{t_2}$.

Definition 18 (KL Divergence) *For any two probability distributions p and q, the KL-divergence, $D_{KL}(p\,||\,q)$, between p and q is defined as*

$$D_{KL}(p\,||\,q) := \mathbb{E}_{\boldsymbol{x}\sim p} \ln \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}.$$

Definition 19 (TV Distance) *For a pair p, q of probability distributions that are absolutely continuous with respect to some common measure, define the total variation distance between them as,*

$$d_{TV}(p, q) := \int |p(\boldsymbol{x}) - q(\boldsymbol{x})|\,\mathrm{d}\boldsymbol{x}$$

Theorem 20 (Pinsker's Inequality) *The famous Pinsker inequality sates that for any pair p, q of probability distributions that are absolutely continuous with respect to some common measure, $d_{TV}(p, q) \le \sqrt{2D_{KL}(p\,||\,q)}$.*

Gaussian Random Variables

The following are well-known results about learning d-dimensional Gaussian random variables. Let $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{2m} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and define,

$$\widehat{\boldsymbol{\mu}} = \frac{1}{m}\sum_{i\in[m]} \boldsymbol{x}_i, \quad \widehat{\boldsymbol{\Sigma}} = \frac{1}{2m}\sum_{i\in[m]} (\boldsymbol{x}_{2i} - \boldsymbol{x}_{2i-1})(\boldsymbol{x}_{2i} - \boldsymbol{x}_{2i-1})^\top. \tag{6}$$

Lemma 21 (TV Learning Gaussians, Theorem C.1, Ashtiani et al. (2020)) *There exists an absolute constant C such that if we take $2m = 2C(d^2 + d\log(1/\delta))/\epsilon^2$ samples $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{2m} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then for $\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}$ as defined in Eq. (6), with probability $1-\delta$,*

$$d_{TV}\left(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \mathcal{N}(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})\right) \leq \epsilon/2.$$

Lemma 22 (Estimating Mean of Gaussians, Lemma C.2 Ashtiani et al. (2020)) *If $m \geq (2d + 6\sqrt{d\log(2/\delta)})/\epsilon^2$, then we have for $\widehat{\boldsymbol{\mu}}$ as defined in Eq. (6),*

$$\Pr\left[(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \geq \frac{\epsilon^2}{2}\right] \leq \frac{\delta}{2}.$$

Lemma 23 (Estimating the Covariance of Gaussians, Lemma C.3 Ashtiani et al. (2020)) *There exists an absolute constant C such that if $m \geq C(d^2 + d\log(1/\delta))/\epsilon^2$, then for $\widehat{\boldsymbol{\Sigma}}$ as defined in Eq. (6) with probability at least $1 - \delta/2$,*

$$\left\|\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1/2} - \boldsymbol{I}\right\|_{op} \leq \frac{\epsilon}{\sqrt{2d}}.$$

Note that since for any positive semi-definite matrix $\boldsymbol{A}$, it holds that $\|\boldsymbol{A}\|_F \leq \|\boldsymbol{A}\|_{op} \cdot \sqrt{d}$, the above guarantee is stronger than the standard on the Frobenius norms of the covariance matrix.

Subexponential Random Variables

Definition 24 (Subesponential Random Variables) *We say that a random variable Z over $\mathbb{R}$ is (σ^2, B)-subexponential, if and only if for every $\lambda \leq 1/B$, $\mathbb{E}\exp(\lambda(Z - \mathbb{E}\,Z)) \leq \exp(\lambda^2\sigma^2/2)$.*

The following fact is easy to derive directly from the definition of subexponential random variables.

Fact 25 *If $Z_1, \ldots Z_n$ are (σ_i, B_i)-subexponential independent random variables, then $Z = \sum_i Z_i$ is (σ, B)-subexponential, where $\sigma = \sqrt{\sum \sigma_i^2}$, and $B = \max_i(B_i)$.*

The following are well known facts about Gaussian random variables.

Fact 26 *If $Z = \mathcal{N}(\mu, 1)$, then Z^2 is $(16, 1/4)$-subexponential. Further, if Z is a Gaussian vector $Z \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\|\boldsymbol{\Sigma}\|_{op} \leq O(1)$, and A is a symmetric matrix, then $Z^T A Z$ is $(O(\|A\|_F), O(\|A\|_{op}))$-subgaussian, where $\|A\|_{op}$ is the largest eigenvalue of A.*

Exponential Families

Definition 27 (Probability Density of Exp. Distributions) *We define an exponential family of probability distributions as those distributions whose density (relative to parameter $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^D$) have the following general form:*

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \boldsymbol{h}(\boldsymbol{x}) \cdot \exp\left(\boldsymbol{\theta}^\top \boldsymbol{T}(\boldsymbol{x}) - A(\boldsymbol{\theta})\right),$$

for a parameter $\boldsymbol{\theta}$, and functions $\boldsymbol{h}$ and $\boldsymbol{T}$. The function A is then automatically determined as the normalizing constant.

This family captures several widely used probability distributions such as Gaussian, Gamma, Binomial, Bernoulli, and Poisson.

In particular the log density ratio for any two distributions in the exponential family, is an affine function of the reparmetrization $\boldsymbol{T}$

$$\ln(p(\boldsymbol{x}|\boldsymbol{\theta}_1)/p(\boldsymbol{x}|\boldsymbol{\theta}_2)) = \langle \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2, \boldsymbol{T}(x)\rangle - A(\boldsymbol{\theta}_1) + A(\boldsymbol{\theta}_2).$$

Rademacher complexity

Definition 28 *Rademacher complexity of a family $\mathcal{F}$ of functions $\Theta \to \mathbb{R}$, with respect to the distribution $\mathcal{D}$ is defined as*

$$\mathcal{R}_{n,\mathcal{D}}(\mathcal{F}) := \mathbb{E}_{\boldsymbol{x}_1,\dots\boldsymbol{x}_n\sim\mathcal{D}}\, \mathbb{E}_{\sigma_1,\dots\sigma_n} \frac{1}{n} \sup_{f\in\mathcal{F}} \sum_i \sigma_i f(\boldsymbol{x}_i),$$

where $\sigma_1, \dots \sigma_n$ are independent $\{\pm 1\}$ Rademacher random variables.

We will repeatedly use that left-composition of the entire function class with a single univariate Lipschitz function does not increase the Rademacher complexity. The following is a corollary of Lemma 5 in Meir and Zhang (2003), see also Theorem 4.12 in Ledoux and Talagrand (2013).

Theorem 29 *If $\ell : \mathbb{R} \to \mathbb{R}$ is a 1-Lipschitz function, $\mathcal{F}'$ is arbitrary family of functions from Θ to $\mathbb{R}$, $\mathcal{D}$ is a distribution on Θ and $\mathcal{F} := \{\ell \circ f : f \in \mathcal{F}'\}$, then*

$$\mathcal{R}_{n,\mathcal{D}}(\mathcal{F}) \le \mathcal{R}_{n,\mathcal{D}}(\mathcal{F}').$$

The main reason Rademacher complexity is useful is to provide generalization bounds: using a finite sample, we can estimate simultanously expectation of all functions in the class with bounded Rademacher complexity.

Theorem 30 *Let $\boldsymbol{x}_1, \dots \boldsymbol{x}_n \sim \mathcal{D}$ be a sequence of i.i.d. random variables, and let $\mathcal{F}$ be a family of functions. Then*

$$\mathbb{E}_{\boldsymbol{x}} \sup_{f\in\mathcal{F}} \left| \frac{1}{n} \sum_{i\le n} f(\boldsymbol{x}_i) - \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}} f(\boldsymbol{x}) \right| \le 2\mathcal{R}_{n,\mathcal{D}}(\mathcal{F}). \tag{7}$$

Note that if $\mathcal{F}$ is in addition a family of functions bounded by c, a stronger concentration guarantees are known for (7). That is, with probability $1-\delta$, we have

$$\sup_{f\in\mathcal{F}} \left| \frac{1}{n} \sum_{i\le n} f(\boldsymbol{x}_i) - \mathbb{E}_{\boldsymbol{x}} f(\boldsymbol{x}) \right| \lesssim \mathcal{R}_{n,\mathcal{D}}(\mathcal{F}) + \frac{c\sqrt{\log(1/\delta)}}{n},$$

but we will not be using this bound explicitly.

Concentration of Random Variables

Lemma 31 (Hoeffding's Inequality) *Let $X_1, \cdots, X_n$ be independent random variables such that $a \le X_i \le b$, and $\mathbb{E}[X_i] = \mu$ for all i. Consider the average of these random variables, $\overline{X} = \frac{1}{n}\sum_i X_i$. Then, for all $t > 0$,*

$$\Pr\left[|\overline{X} - \mu| \ge t\right] \le 2\exp\left\{-\frac{2nt^2}{(b-a)^2}\right\}.$$

Appendix B. Proofs from Section 2

For the analysis of the algorithm, we let $\widehat{p_{\text{te}}} = \mathcal{N}(\widehat{\boldsymbol{\mu}_{\text{te}}}, \widehat{\boldsymbol{\Sigma}_{\text{te}}}), \widehat{p_{\text{tr}}} = \mathcal{N}(\widehat{\boldsymbol{\mu}_{\text{tr}}}, \widehat{\boldsymbol{\Sigma}_{\text{tr}}})$ to denote approximations to the $p_{\text{te}}, p_{\text{tr}}$ respectively. We now begin with the first part of the analysis.

Bound on Bias of the Estimator: Proof of Lemmas 7 and 8

Lemma 32 *For any $p_{\text{tr}}, p_{\text{te}}$ with corresponding approximations $\widehat{p_{\text{tr}}}, \widehat{p_{\text{te}}}$, and bounded $\boldsymbol{f}(\boldsymbol{x})$ such that $\sup_{\boldsymbol{x}} |\boldsymbol{f}(\boldsymbol{x})| \leq 1$ we have*

$$\left| \mathbb{E}_{\boldsymbol{x}\sim p_{\text{tr}}} \frac{\widehat{p_{\text{te}}}(\boldsymbol{x})}{\widehat{p_{\text{tr}}}(\boldsymbol{x})} \boldsymbol{f}(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{x}\sim p_{\text{te}}} \boldsymbol{f}(\boldsymbol{x}) \right| \leq Q_1(\widehat{p_{\text{tr}}} \,|\, p_{\text{tr}};\ \widehat{p_{\text{te}}}) + d_{TV}(\widehat{p_{\text{te}}}, p_{\text{te}}).$$

Proof First by triangle inequality we can decompose

$$\begin{aligned} \left| \mathbb{E}_{\boldsymbol{x}\sim p_{\text{tr}}} \frac{\widehat{p_{\text{te}}}(\boldsymbol{x})}{\widehat{p_{\text{tr}}}(\boldsymbol{x})} \boldsymbol{f}(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{x}\sim p_{\text{te}}} \boldsymbol{f}(\boldsymbol{x}) \right| &\leq \left| \mathbb{E}_{\boldsymbol{x}\sim p_{\text{tr}}} \frac{\widehat{p_{\text{te}}}(\boldsymbol{x})}{\widehat{p_{\text{tr}}}(\boldsymbol{x})} \boldsymbol{f}(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{x}\sim \widehat{p_{\text{te}}}} \boldsymbol{f}(\boldsymbol{x}) \right| \\ &\quad + |\mathbb{E}_{\boldsymbol{x}\sim \widehat{p_{\text{te}}}} \boldsymbol{f}(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{x}\sim p_{\text{te}}} \boldsymbol{f}(\boldsymbol{x})| \end{aligned} \tag{8}$$

We now consider the two terms separately. Using the fact that $|\boldsymbol{f}(\boldsymbol{x})| \leq 1$, the second term gives,

$$\begin{aligned} |\mathbb{E}_{\boldsymbol{x}\sim \widehat{p_{\text{te}}}} \boldsymbol{f}(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{x}\sim p_{\text{te}}} \boldsymbol{f}(\boldsymbol{x})| &= \left| \int_{\boldsymbol{x}} \boldsymbol{f}(\boldsymbol{x})(\widehat{p_{\text{te}}}(\boldsymbol{x}) - p_{\text{te}}(\boldsymbol{x}))d\boldsymbol{x} \right| \\ &\leq \sup_{\boldsymbol{x}} |\boldsymbol{f}(\boldsymbol{x})| \int_{\boldsymbol{x}} |\widehat{p_{\text{te}}} - p_{\text{te}}| d\boldsymbol{x} \\ &\leq d_{TV}(\widehat{p_{\text{te}}}, p_{\text{te}}). \end{aligned}$$

We now consider the first term.

$$\begin{aligned} \left| \mathbb{E}_{\boldsymbol{x}\sim p_{\text{tr}}} \frac{\widehat{p_{\text{te}}}(\boldsymbol{x})}{\widehat{p_{\text{tr}}}(\boldsymbol{x})} \boldsymbol{f}(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{x}\sim \widehat{p_{\text{te}}}} \boldsymbol{f}(\boldsymbol{x}) \right| &= \left| \mathbb{E}_{\boldsymbol{x}\sim p_{\text{tr}}} \frac{\widehat{p_{\text{te}}}(\boldsymbol{x}) p_{\text{tr}}(\boldsymbol{x})}{p_{\text{tr}}(\boldsymbol{x}) \widehat{p_{\text{tr}}}(\boldsymbol{x})} \boldsymbol{f}(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{x}\sim \widehat{p_{\text{te}}}} \boldsymbol{f}(\boldsymbol{x}) \right| \\ &= \left| \mathbb{E}_{\boldsymbol{x}\sim \widehat{p_{\text{te}}}} \frac{p_{\text{tr}}(\boldsymbol{x})}{\widehat{p_{\text{tr}}}(\boldsymbol{x})} \boldsymbol{f}(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{x}\sim \widehat{p_{\text{te}}}} \boldsymbol{f}(\boldsymbol{x}) \right| \\ &\leq \mathbb{E}_{\boldsymbol{x}\sim \widehat{p_{\text{te}}}} \left| \frac{p_{\text{tr}}(\boldsymbol{x})}{\widehat{p_{\text{tr}}}(\boldsymbol{x})} - 1 \right|. \end{aligned}$$

Using these bounds in Eq. (8) gives the result.

■

From the above lemma, we see that we need to first give a bound on $Q_1(\widehat{p_{\text{tr}}} \,|\, p_{\text{tr}};\ \widehat{p_{\text{te}}})$. Since from Definition 17, we know that $Q_1(\cdot \,|\, \cdot;\ \cdot) \leq Q_2(\cdot \,|\, \cdot;\ \cdot)$, in the remaining part we will show how to bound $Q_2(\widehat{p_{\text{tr}}} \,|\, p_{\text{tr}};\ \widehat{p_{\text{te}}})$.

Lemma 33 *Let $p, \widehat{p}$, and q denote probability distributions and let Z be a random variable defined as $Z = \ln(p(\boldsymbol{x})/\widehat{p}(\boldsymbol{x}))$ for $\boldsymbol{x} \sim q$. If $\mu := \mathbb{E}\, Z$ and Z is $(\sigma, 1/2)$-subexponential for $\sigma \leq O(1)$, then $Q_2(\widehat{p} \,|\, p;\ q) \leq O(\mu + \sigma)$.*

Proof We have

$$Q_2(\widehat{p}\,|\,p;\ q)^2 = \mathbb{E}_{\boldsymbol{x}\sim q}\left(\frac{p(\boldsymbol{x})}{\widehat{p}(\boldsymbol{x})} - 1\right)^2$$
$$= \mathbb{E}_{\boldsymbol{x}\sim q}(\exp(Z)-1)^2 = \mathbb{E}_{\boldsymbol{x}\sim q}[\exp(2Z) - 2\exp(Z) + 1].$$

Since Z is subexponential, we have $\mathbb{E}_{\boldsymbol{x}\sim q}\exp(Z-\mu) \le \exp(\sigma)$, and therefore, $\mathbb{E}_{\boldsymbol{x}\sim q}\exp(2Z) \le \exp(2\mu + 2\sigma)$. Now, we will use that for any $y \ge 0$, $e^y \le 1 + y + O(1)y^2$ for $y \le O(1)$ and $e^y \ge 1 + y$ which yields

$$\mathbb{E}_{\boldsymbol{x}\in q}[\exp(2Z) - 2\exp(Z) + 1] = 1 + 2\mu + O(1)\cdot(\mu+\sigma)^2 - 2(1+\mu) - 2$$
$$= O(1)\cdot(\mu+\sigma)^2.$$

■

We begin with the easy case and first give a bound on the bias of our estimator in the case when p_{tr} and p_{te} are isotropic Gaussian distributions.

Lemma 34 (Bound for Isotropic Gaussian Distributions) *Let $p_{\mathrm{tr}} = \mathcal{N}(\boldsymbol{\mu}_{\mathrm{tr}}, \boldsymbol{I})$ and $p_{\mathrm{te}} = \mathcal{N}(\boldsymbol{\mu}_{\mathrm{te}}, \boldsymbol{I})$ such that $\|\boldsymbol{\mu}_{\mathrm{tr}}\|_2 \le O(1)$ and $\|\boldsymbol{\mu}_{\mathrm{tr}} - \boldsymbol{\mu}_{\mathrm{te}}\|_2 \le B \le O(1)$. Further, let $\widehat{p_{\mathrm{tr}}}$ and $\widehat{p_{\mathrm{te}}}$ be as defined in Algorithm 1. Then for any function $\boldsymbol{f}$ such that $\sum_{\boldsymbol{x}\in\mathbb{R}^d}|\boldsymbol{f}(\boldsymbol{x})| \le 1$,*

$$\left|\mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{tr}}}\frac{\widehat{p_{\mathrm{te}}}(\boldsymbol{x})}{\widehat{p_{\mathrm{tr}}}(\boldsymbol{x})}\boldsymbol{f}(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{te}}}\boldsymbol{f}(\boldsymbol{x})\right| \le O(\varepsilon).$$

Proof We will use theorem 33 to show $Q_2(\widehat{p_{\mathrm{tr}}}\,|\,p_{\mathrm{tr}};\ \widehat{p_{\mathrm{te}}}) \le O(\varepsilon)$ in this case. The remaining follows from theorem 32.

Let $p = p_{\mathrm{tr}}$, $\widehat{p} = \widehat{p_{\mathrm{tr}}}$, and $q = \widehat{p_{\mathrm{te}}}$. We need to show that Z, as defined in theorem 33 is $(\varepsilon, 1/2)$ subexponential and $\mathbb{E}_{\boldsymbol{x}\sim q} Z \le O(\varepsilon)$.

Now, $Z = \ln p_{\mathrm{tr}} - \ln\widehat{p_{\mathrm{tr}}} = \langle \boldsymbol{\mu}_{\mathrm{tr}} - \widehat{\boldsymbol{\mu}_{\mathrm{tr}}}, \boldsymbol{x}\rangle + \frac{1}{2}\langle \widehat{\boldsymbol{\mu}_{\mathrm{tr}}} - \boldsymbol{\mu}_{\mathrm{tr}}, \boldsymbol{\mu}_{\mathrm{tr}} + \widehat{\boldsymbol{\mu}_{\mathrm{tr}}}\rangle$ is a Gaussian with mean $\langle \widehat{\boldsymbol{\mu}_{\mathrm{tr}}} - \boldsymbol{\mu}_{\mathrm{tr}}, \widehat{\boldsymbol{\mu}_{\mathrm{te}}}\rangle + \frac{1}{2}\langle \widehat{\boldsymbol{\mu}_{\mathrm{tr}}} - \boldsymbol{\mu}_{\mathrm{tr}}, \boldsymbol{\mu}_{\mathrm{tr}} + \widehat{\boldsymbol{\mu}_{\mathrm{tr}}}\rangle$ and variance $\|\boldsymbol{\mu}_{\mathrm{tr}} - \widehat{\boldsymbol{\mu}_{\mathrm{tr}}}\|_2^2$. From Lemma 22, we know that that $\|\boldsymbol{\mu}_{\mathrm{tr}} - \widehat{\boldsymbol{\mu}_{\mathrm{tr}}}\|_2^2 \le \varepsilon^2$, and therefore, $Var(Z) \le \varepsilon^2$. Further, the mean satisfies,

$$\mathbb{E}\,Z = \langle \widehat{\boldsymbol{\mu}_{\mathrm{tr}}} - \boldsymbol{\mu}_{\mathrm{tr}}, 2\widehat{\boldsymbol{\mu}_{\mathrm{te}}} + \boldsymbol{\mu}_{\mathrm{tr}} + \widehat{\boldsymbol{\mu}_{\mathrm{tr}}}\rangle \le \|\widehat{\boldsymbol{\mu}_{\mathrm{tr}}} - \boldsymbol{\mu}_{\mathrm{tr}}\|_2\|2\widehat{\boldsymbol{\mu}_{\mathrm{te}}} + \boldsymbol{\mu}_{\mathrm{tr}} + \widehat{\boldsymbol{\mu}_{\mathrm{tr}}}\|_2.$$

In the above, since $\|\boldsymbol{\mu}_{\mathrm{tr}} - \boldsymbol{\mu}_{\mathrm{te}}\|_2 \le O(1)$, and $\|\boldsymbol{\mu}_{\mathrm{tr}}\|_2 \le O(1)$, we must have $\|\widehat{\boldsymbol{\mu}_{\mathrm{te}}}\|_2 \le O(1)$. Therefore,

$$\mathbb{E}\,Z \le \|\widehat{\boldsymbol{\mu}_{\mathrm{tr}}} - \boldsymbol{\mu}_{\mathrm{tr}}\|_2(2\|\widehat{\boldsymbol{\mu}_{\mathrm{te}}}\|_2 + \|\boldsymbol{\mu}_{\mathrm{tr}}\|_2 + \|\widehat{\boldsymbol{\mu}_{\mathrm{tr}}}\|_2) \le O(\varepsilon).$$

We now claim that Z is $(\varepsilon, 1/2)$ subexponential with mean bounded by $O(\varepsilon)$. This follows since Z is gaussian with std deviation ε, the MGF of Z is bouneded by $e^{\varepsilon^2\lambda^2}$ for all $\lambda > 0$, which implies Z is $(\varepsilon, 1/2)$ subexponential. This concludes the proof for isotropic Gaussian distributions. ■

Lemma 35 (Bound for General Gaussian Distributions) *Let $p_{\mathrm{te}} := \mathcal{N}(\boldsymbol{\mu}_{\mathrm{te}}, \boldsymbol{\Sigma}_{\mathrm{te}}), p_{\mathrm{tr}} := \mathcal{N}(\boldsymbol{\mu}_{\mathrm{tr}}, \boldsymbol{\Sigma}_{\mathrm{tr}})$ and $\widehat{p_{\mathrm{tr}}} = \mathcal{N}(\widehat{\boldsymbol{\mu}_{\mathrm{tr}}}, \widehat{\boldsymbol{\Sigma}_{\mathrm{tr}}}), \widehat{p_{\mathrm{te}}} = \mathcal{N}(\widehat{\boldsymbol{\mu}_{\mathrm{te}}}, \widehat{\boldsymbol{\Sigma}_{\mathrm{te}}})$ be as defined in Algorithm 1. If $\|\boldsymbol{\mu}_{\mathrm{tr}} - \boldsymbol{\mu}_{\mathrm{te}}\|_2 \le B \le O(1)$, $\|\boldsymbol{\Sigma}_{\mathrm{tr}}^{-1}\|_{op} \le O(1)$ then, for any function $\boldsymbol{f}$, such that $\sup_{\boldsymbol{x}\in\mathbb{R}^d}|\boldsymbol{f}(\boldsymbol{x})| \le 1$,*

$$\left|\mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{tr}}}\frac{\widehat{p_{\mathrm{te}}}(\boldsymbol{x})}{\widehat{p_{\mathrm{tr}}}(\boldsymbol{x})}\boldsymbol{f}(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{te}}}\boldsymbol{f}(\boldsymbol{x})\right| \le O(\varepsilon).$$

Proof We will deduce this as a corollary from Theorem 33. To this end, let $Z := \ln p_{\mathrm{tr}}(\boldsymbol{x}) - \ln \widehat{p_{\mathrm{tr}}}(\boldsymbol{x})$ for $\boldsymbol{x} \sim \widehat{p_{\mathrm{te}}}$. Note that $Z = \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^T x + c$, where

$$\begin{aligned}
\boldsymbol{A} &= (\boldsymbol{\Sigma}_{\mathrm{tr}}^{-1} - \widehat{\boldsymbol{\Sigma}_{\mathrm{tr}}}^{-1})/2 \\
\boldsymbol{b} &= \boldsymbol{\Sigma}_{\mathrm{tr}}^{-1}\boldsymbol{\mu}_{\mathrm{tr}} - \widehat{\boldsymbol{\Sigma}_{\mathrm{tr}}}^{-1}\widehat{\boldsymbol{\mu}_{\mathrm{tr}}} \\
c &= (\boldsymbol{\mu}_{\mathrm{tr}}^\top \boldsymbol{\Sigma}_{\mathrm{tr}}^{-1}\boldsymbol{\mu}_{\mathrm{tr}} - \widehat{\boldsymbol{\mu}_{\mathrm{tr}}}^\top \widehat{\boldsymbol{\Sigma}_{\mathrm{tr}}}^{-1}\widehat{\boldsymbol{\mu}_{\mathrm{tr}}})/2.
\end{aligned}$$

Now, from Fact 26, $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}$ is $(\|\boldsymbol{A}\|_F, \|\boldsymbol{A}\|_{op})$-subexponential, and we know that $\|\boldsymbol{A}\|_{op} \leq \|\boldsymbol{A}\|_F \leq \sqrt{d}\|\boldsymbol{A}\|_{op}$.

$$\|\widehat{\boldsymbol{\Sigma}_{\mathrm{tr}}}^{-1} - \boldsymbol{\Sigma}_{\mathrm{tr}}^{-1}\|_{op} \leq \|\boldsymbol{\Sigma}_{\mathrm{tr}}^{-1/2}\|_{op}\|\boldsymbol{\Sigma}_{\mathrm{tr}}^{1/2}\widehat{\boldsymbol{\Sigma}_{\mathrm{tr}}}^{-1}\boldsymbol{\Sigma}_{\mathrm{tr}}^{1/2} - \boldsymbol{I}\|_{op}\|\boldsymbol{\Sigma}_{\mathrm{tr}}^{-1/2}\|_{op}.$$

From Lemma 23, we know that $\|\boldsymbol{\Sigma}_{\mathrm{tr}}^{-1/2}\widehat{\boldsymbol{\Sigma}_{\mathrm{tr}}}\boldsymbol{\Sigma}_{\mathrm{tr}}^{-1/2} - \boldsymbol{I}\|_{op} \leq \frac{\varepsilon}{2\sqrt{d}}$. This implies that all eigenvalues of $\boldsymbol{\Sigma}_{\mathrm{tr}}^{-1/2}\widehat{\boldsymbol{\Sigma}_{\mathrm{tr}}}\boldsymbol{\Sigma}_{\mathrm{tr}}^{-1/2}$ are between $1 - \varepsilon/2\sqrt{d}$ and $1 + \varepsilon/2\sqrt{d}$ and as a result, $\|\boldsymbol{\Sigma}_{\mathrm{tr}}^{1/2}\widehat{\boldsymbol{\Sigma}_{\mathrm{tr}}}^{-1}\boldsymbol{\Sigma}_{\mathrm{tr}}^{1/2} - \boldsymbol{I}\|_{op} \leq \varepsilon/\sqrt{d}$. Now, from assumption, $\|\boldsymbol{\Sigma}_{\mathrm{tr}}^{-1}\|_{op} \leq O(1)$, and we get that,

$$\|\widehat{\boldsymbol{\Sigma}_{\mathrm{tr}}}^{-1} - \boldsymbol{\Sigma}_{\mathrm{tr}}^{-1}\|_F \leq \sqrt{d} \cdot \|\widehat{\boldsymbol{\Sigma}_{\mathrm{tr}}}^{-1} - \boldsymbol{\Sigma}_{\mathrm{tr}}^{-1}\|_{op} \leq O(\varepsilon).$$

Therefore, Z is $(O(\varepsilon), O(\varepsilon))$-subexponential. The random variable $\boldsymbol{b}^\top \boldsymbol{x}$ is $\|\boldsymbol{b}\|_2$-subgaussian, and we can bound,

$$\|\boldsymbol{b}\|_2 = \|\boldsymbol{\Sigma}_{\mathrm{tr}}^{-1}\boldsymbol{\mu}_{\mathrm{tr}} - \widehat{\boldsymbol{\Sigma}_{\mathrm{tr}}}^{-1}\widehat{\boldsymbol{\mu}_{\mathrm{tr}}}\|_2 \leq \|\boldsymbol{\Sigma}_{\mathrm{tr}}^{-1}(\boldsymbol{\mu}_{\mathrm{tr}} - \widehat{\boldsymbol{\mu}_{\mathrm{tr}}})\|_2 + \|(\widehat{\boldsymbol{\Sigma}_{\mathrm{tr}}}^{-1} - \boldsymbol{\Sigma}_{\mathrm{tr}}^{-1})\widehat{\boldsymbol{\mu}_{\mathrm{tr}}}\|_2 \leq O(\varepsilon).$$

For the first term we use Lemma 22 and for the second term we use the previous calculation along with the assumption that $\|\boldsymbol{\mu}_{\mathrm{tr}}\|_2 \leq O(1)$. The sum of ε-subgaussian and $(\varepsilon, \varepsilon)$-subexponential random variable is $(O(\varepsilon), O(\varepsilon))$-subexponential again.

Finally $\mathbb{E}\, Z = \operatorname{Tr} \boldsymbol{A} + \boldsymbol{b}^\top \boldsymbol{\mu}_{\mathrm{te}} + c$, and since $\|\boldsymbol{\mu}_{\mathrm{tr}} - \boldsymbol{\mu}_{\mathrm{te}}\|_2 \leq O(1)$, and $\|\boldsymbol{\mu}_{\mathrm{tr}}\|_2 \leq O(1)$ imply that $\|\boldsymbol{\mu}_{\mathrm{te}}\|_2 \leq O(1)$, $\operatorname{Tr} \boldsymbol{A} \leq \|\boldsymbol{A}\|_F$, all we need to do is bound c.

$$(\boldsymbol{\mu}_{\mathrm{tr}}^\top \boldsymbol{\Sigma}_{\mathrm{tr}}^{-1}\boldsymbol{\mu}_{\mathrm{tr}} - \widehat{\boldsymbol{\mu}_{\mathrm{tr}}}^\top \widehat{\boldsymbol{\Sigma}_{\mathrm{tr}}}^{-1}\widehat{\boldsymbol{\mu}_{\mathrm{tr}}}) \leq \|\boldsymbol{\mu}_{\mathrm{tr}}^\top(\boldsymbol{\Sigma}_{\mathrm{tr}}^{-1} - \widehat{\boldsymbol{\Sigma}_{\mathrm{tr}}}^{-1})\boldsymbol{\mu}_{\mathrm{tr}}\|_2 + \langle \boldsymbol{\Sigma}_{\mathrm{tr}}^{-1}(\boldsymbol{\mu}_{\mathrm{tr}} - \widehat{\boldsymbol{\mu}_{\mathrm{tr}}}), (\boldsymbol{\mu}_{\mathrm{tr}} + \widehat{\boldsymbol{\mu}_{\mathrm{tr}}})\rangle \leq O(\varepsilon).$$

The final statement of the lemma now follows from Lemma 32. ■

Bound on Variance of the Estimator: Proof of Lemma 9

In this section, we give a bound on the variance of our estimator $\frac{\widehat{p_{\mathrm{te}}}(\boldsymbol{x})}{\widehat{p_{\mathrm{tr}}}(\boldsymbol{x})}\boldsymbol{f}(\boldsymbol{x})$ for $\boldsymbol{x} \sim p_{\mathrm{tr}}$.

Lemma 36 *Let X denote the estimator $\frac{\widehat{p_{\mathrm{te}}}(\boldsymbol{x})}{\widehat{p_{\mathrm{tr}}}(\boldsymbol{x})}\boldsymbol{f}(\boldsymbol{x})$for $\boldsymbol{x} \sim p_{\mathrm{tr}}$. If $\|\boldsymbol{\Sigma}_{\mathrm{tr}}^{-1}\|_{op} \leq O(1)$, $\|\boldsymbol{\Sigma}_{\mathrm{tr}}^{-1} - \boldsymbol{\Sigma}_{\mathrm{te}}^{-1}\|_F \leq O(1)$, and $\|\boldsymbol{\mu}_{\mathrm{tr}}\|_2, \|\boldsymbol{\mu}_{\mathrm{tr}} - \boldsymbol{\mu}_{\mathrm{te}}\|_2 \leq O(1)$, then*

$$\mathbb{E}_{\boldsymbol{x} \sim p_{\mathrm{tr}}}[X^2] \leq 2 + 2Q_2(\widehat{p_{\mathrm{tr}}} \,|\, \widehat{p_{\mathrm{te}}};\ p_{\mathrm{tr}}) \leq O(1).$$

Proof

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{tr}}}[X^2] &= \mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{tr}}}\left[\left(\frac{\widehat{p_{\mathrm{te}}}(\boldsymbol{x})}{\widehat{p_{\mathrm{tr}}}(\boldsymbol{x})}\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{f}(\boldsymbol{x})\right)^2\right] \\ &\le 2\,\mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{tr}}}\left[\left(\frac{\widehat{p_{\mathrm{te}}}(\boldsymbol{x})}{\widehat{p_{\mathrm{tr}}}(\boldsymbol{x})} - 1\right)^2 \boldsymbol{f}(\boldsymbol{x})^2\right] + 2\,\mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{tr}}}[\boldsymbol{f}(\boldsymbol{x})^2] \\ &\le 2\cdot Q_2(\widehat{p_{\mathrm{tr}}}\,|\,\widehat{p_{\mathrm{te}}};\ p_{\mathrm{tr}}) + 2.\end{aligned}$$

In the last line we used $|\boldsymbol{f}(\boldsymbol{x})|\le 1$.

Now, we consider the distribution of $Z := \ln\widehat{p_{\mathrm{te}}}(\boldsymbol{x}) - \ln\widehat{p_{\mathrm{tr}}}(\boldsymbol{x})$ for $\boldsymbol{x}\sim p_{\mathrm{tr}}$. Note that $Z = \boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^T x + c$, where

$$\begin{aligned}\boldsymbol{A} &= (\widehat{\boldsymbol{\Sigma}_{\mathrm{te}}}^{-1} - \widehat{\boldsymbol{\Sigma}_{\mathrm{tr}}}^{-1})/2 \\ \boldsymbol{b} &= \widehat{\boldsymbol{\Sigma}_{\mathrm{te}}}^{-1}\widehat{\boldsymbol{\mu}_{\mathrm{te}}} - \widehat{\boldsymbol{\Sigma}_{\mathrm{tr}}}^{-1}\widehat{\boldsymbol{\mu}_{\mathrm{tr}}} \\ c &= (\widehat{\boldsymbol{\mu}_{\mathrm{te}}}^\top\widehat{\boldsymbol{\Sigma}_{\mathrm{te}}}^{-1}\widehat{\boldsymbol{\mu}_{\mathrm{tr}}} - \widehat{\boldsymbol{\mu}_{\mathrm{tr}}}^\top\widehat{\boldsymbol{\Sigma}_{\mathrm{tr}}}^{-1}\widehat{\boldsymbol{\mu}_{\mathrm{tr}}})/2.\end{aligned}$$

Now, from Fact 26, $\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x}$ is $(\|\boldsymbol{A}\|_F, \|\boldsymbol{A}\|_{op})$-subexponential, and since $\|\boldsymbol{A}\|_{op}\le \|\boldsymbol{A}\|_F\le \sqrt{d}\|\boldsymbol{A}\|_{op}$ we will give a bound on $\|\boldsymbol{A}\|_F$.

From triangle inequality,

$$\|\boldsymbol{A}\|_F\le \|\widehat{\boldsymbol{\Sigma}_{\mathrm{te}}}^{-1} - \boldsymbol{\Sigma}_{\mathrm{te}}^{-1}\|_F + \|\boldsymbol{\Sigma}_{\mathrm{te}}^{-1} - \boldsymbol{\Sigma}_{\mathrm{tr}}^{-1}\|_F + \|\widehat{\boldsymbol{\Sigma}_{\mathrm{tr}}}^{-1} - \boldsymbol{\Sigma}_{\mathrm{tr}}^{-1}\|_F.$$

Now,

$$\|\widehat{\boldsymbol{\Sigma}_{\mathrm{te}}}^{-1} - \boldsymbol{\Sigma}_{\mathrm{te}}^{-1}\|_{op}\le \|\boldsymbol{\Sigma}_{\mathrm{te}}^{-1/2}\|_{op}\|\boldsymbol{\Sigma}_{\mathrm{te}}^{1/2}\widehat{\boldsymbol{\Sigma}_{\mathrm{te}}}^{-1}\boldsymbol{\Sigma}_{\mathrm{te}}^{1/2} - \boldsymbol{I}\|_{op}\|\boldsymbol{\Sigma}_{\mathrm{te}}^{-1/2}\|_{op}.$$

From Lemma 23, we know that $\|\boldsymbol{\Sigma}_{\mathrm{te}}^{-1/2}\widehat{\boldsymbol{\Sigma}_{\mathrm{te}}}\boldsymbol{\Sigma}_{\mathrm{te}}^{-1/2} - \boldsymbol{I}\|_{op}\le \frac{\varepsilon}{2\sqrt{d}}$. This implies that all eigenvalues of $\boldsymbol{\Sigma}_{\mathrm{te}}^{-1/2}\widehat{\boldsymbol{\Sigma}_{\mathrm{te}}}\boldsymbol{\Sigma}_{\mathrm{te}}^{-1/2}$ are between $1-\varepsilon/2\sqrt{d}$ and $1+\varepsilon/2\sqrt{d}$ and as a result, $\|\boldsymbol{\Sigma}_{\mathrm{te}}^{1/2}\widehat{\boldsymbol{\Sigma}_{\mathrm{te}}}^{-1}\boldsymbol{\Sigma}_{\mathrm{te}}^{1/2} - \boldsymbol{I}\|_{op}\le \varepsilon/\sqrt{d}$. Now, from assumption, $\|\boldsymbol{\Sigma}_{\mathrm{te}}^{-1}\|_{op}\le O(1)$, and we get that,

$$\|\widehat{\boldsymbol{\Sigma}_{\mathrm{te}}}^{-1} - \boldsymbol{\Sigma}_{\mathrm{te}}^{-1}\|_F\le \sqrt{d}\cdot\|\widehat{\boldsymbol{\Sigma}_{\mathrm{te}}}^{-1} - \boldsymbol{\Sigma}_{\mathrm{te}}^{-1}\|_{op}\le O\left(\varepsilon\right).$$

Similarly, we can show that

$$\|\widehat{\boldsymbol{\Sigma}_{\mathrm{tr}}}^{-1} - \boldsymbol{\Sigma}_{\mathrm{tr}}^{-1}\|_F\le O(\varepsilon).$$

From assumption, we know that $\|\boldsymbol{\Sigma}_{\mathrm{te}}^{-1} - \boldsymbol{\Sigma}_{\mathrm{tr}}^{-1}\|_F\le O(1)$, and as a result we get that $\|\boldsymbol{A}\|_{op}\le \|\boldsymbol{A}\|_F\le O(1)$.

Therefore, Z is $(O(1), O(1))$-subexponential. The random variable $\boldsymbol{b}^\top\boldsymbol{x}$ is $\|\boldsymbol{b}\|_2$-subgaussian, and we can bound,

$$\begin{aligned}\|\boldsymbol{b}\|_2 &= \|\widehat{\boldsymbol{\Sigma}_{\mathrm{te}}}^{-1}\widehat{\boldsymbol{\mu}_{\mathrm{te}}} - \widehat{\boldsymbol{\Sigma}_{\mathrm{tr}}}^{-1}\widehat{\boldsymbol{\mu}_{\mathrm{tr}}}\|_2 \\ &\le \|\widehat{\boldsymbol{\Sigma}_{\mathrm{te}}}^{-1}\widehat{\boldsymbol{\mu}_{\mathrm{te}}} - \widehat{\boldsymbol{\Sigma}_{\mathrm{te}}}^{-1}\widehat{\boldsymbol{\mu}_{\mathrm{tr}}}\|_2 + \|\widehat{\boldsymbol{\Sigma}_{\mathrm{te}}}^{-1}\widehat{\boldsymbol{\mu}_{\mathrm{tr}}} - \widehat{\boldsymbol{\Sigma}_{\mathrm{tr}}}^{-1}\widehat{\boldsymbol{\mu}_{\mathrm{tr}}}\|_2 \\ &\le \|\widehat{\boldsymbol{\Sigma}_{\mathrm{te}}}^{-1}\|_{op}\|\widehat{\boldsymbol{\mu}_{\mathrm{tr}}} - \widehat{\boldsymbol{\mu}_{\mathrm{te}}}\|_2 + \|\widehat{\boldsymbol{\mu}_{\mathrm{tr}}}\|_2\|\widehat{\boldsymbol{\Sigma}_{\mathrm{tr}}}^{-1} - \widehat{\boldsymbol{\Sigma}_{\mathrm{te}}}^{-1}\|_{op}\le O(1).\end{aligned}$$

The sum of $O(1)$-subgaussian and $(O(1), O(1))$-subexponential random variable is $(O(1), O(1))$-subexponential again.

Finally $\mathbb{E}\, Z = \operatorname{Tr} \boldsymbol{A} + c$, and since $\operatorname{Tr} \boldsymbol{A} \leq \|\boldsymbol{A}\|_F$, all we need to do is bound c.

$$(\widehat{\boldsymbol{\mu}_{\text{te}}}^\top \widehat{\boldsymbol{\Sigma}_{\text{te}}}^{-1} \widehat{\boldsymbol{\mu}_{\text{te}}} - \widehat{\boldsymbol{\mu}_{\text{tr}}}^\top \widehat{\boldsymbol{\Sigma}_{\text{tr}}}^{-1} \widehat{\boldsymbol{\mu}_{\text{tr}}}) \leq \|\widehat{\boldsymbol{\mu}_{\text{tr}}}^\top (\widehat{\boldsymbol{\Sigma}_{\text{te}}}^{-1} - \widehat{\boldsymbol{\Sigma}_{\text{tr}}}^{-1}) \widehat{\boldsymbol{\mu}_{\text{tr}}}\|_2 + \langle \widehat{\boldsymbol{\Sigma}_{\text{te}}}^{-1} (\widehat{\boldsymbol{\mu}_{\text{te}}} - \widehat{\boldsymbol{\mu}_{\text{tr}}}), (\widehat{\boldsymbol{\mu}_{\text{te}}} + \widehat{\boldsymbol{\mu}_{\text{tr}}}) \rangle \leq O(1).$$

From Lemma 33, we can now bound $Q_2(\widehat{p_{\text{tr}}} \,|\, \widehat{p_{\text{te}}};\ p_{\text{tr}}) \leq O(1)$. ■

Lemma 37 (Variance of our Estimator) *If $\|\boldsymbol{\Sigma}_{\text{tr}}^{-1}\|_{op}, \|\boldsymbol{\Sigma}_{\text{te}}^{-1}\|_{op} \leq O(1)$, $\|\boldsymbol{\Sigma}_{\text{tr}}^{-1} - \boldsymbol{\Sigma}_{\text{te}}^{-1}\|_F \leq O(1)$, and $\|\boldsymbol{\mu}_{\text{tr}}\|_2, \|\boldsymbol{\mu}_{\text{tr}} - \boldsymbol{\mu}_{\text{te}}\|_2 \leq O(1)$, then, the variance of our estimator can be bounded as,*

$$\mathbb{E}_{\boldsymbol{x} \sim p_{\text{tr}}} \left[\left(\frac{\widehat{p_{\text{te}}}(\boldsymbol{x})}{\widehat{p_{\text{tr}}}(\boldsymbol{x})} \boldsymbol{f}(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{x} \sim p_{\text{te}}} \boldsymbol{f}(\boldsymbol{x}) \right)^2 \right] \leq O(1).$$

Proof The proof follows from the following equation: for any random variable X and μ,

$$\mathbb{E}[(X - \mu)^2] = \mathbb{E}[X^2] - 2\mu\, \mathbb{E}[X] + \mu^2 \leq \mathbb{E}[X^2] + 2\mu |\mathbb{E}[X] - \mu|.$$

Now, letting $X = \frac{\widehat{p_{\text{te}}}(\boldsymbol{x})}{\widehat{p_{\text{tr}}}(\boldsymbol{x})} \boldsymbol{f}(\boldsymbol{x})$ and $\mu = \mathbb{E}_{\boldsymbol{x} \sim p_{\text{te}}} \boldsymbol{f}(\boldsymbol{x})$, from Theorem 36, $\mathbb{E}[X^2] \leq O(1)$ and from Theorem 8 $|\mathbb{E}[X] - \mu| \leq \varepsilon$. Furthermore, since $|\boldsymbol{f}(\boldsymbol{x})| \leq 1$, we can further bound $\mathbb{E}_{\boldsymbol{x} \sim p_{\text{te}}} \boldsymbol{f}(\boldsymbol{x}) \leq 1$. Therefore $\mathbb{E}[(X - \mu)^2] \leq O(1)$. ■

Proof of Theorems 5 and 6

Proof [Proof of Theorem 5]

From Lemmas 22 and 23, it is sufficient to use $O(d^2/\varepsilon^2 \log 1/\delta)$ samples to obtain $\widehat{p_{\text{tr}}}$ and $\widehat{p_{\text{te}}}$ to the required accuracy.

Now, from Lemma 8 our estimator $X = \frac{\widehat{p_{\text{te}}}(\boldsymbol{x})}{\widehat{p_{\text{tr}}}(\boldsymbol{x})} \boldsymbol{f}(\boldsymbol{x})$ for $\boldsymbol{x} \sim p_{\text{tr}}$ satisfies

$$|\mathbb{E}_{\boldsymbol{x} \sim p_{\text{tr}}} X - \mathbb{E}_{\boldsymbol{x} \sim p_{\text{te}}} \boldsymbol{f}(\boldsymbol{x})| \leq \varepsilon. \tag{9}$$

Further, from Lemma 9,

$$Var(X) = \mathbb{E}_{\boldsymbol{x} \sim p_{\text{tr}}} \left[\left(\frac{\widehat{p_{\text{te}}}(\boldsymbol{x})}{\widehat{p_{\text{tr}}}(\boldsymbol{x})} \boldsymbol{f}(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{x} \sim p_{\text{te}}} \boldsymbol{f}(\boldsymbol{x}) \right)^2 \right] \leq O(1).$$

We now use Chebyshev's inequality to obtain our sample complexity bounds. Since the variance of the estimator is at most $O(1)$ the variance of $Z = \frac{1}{m} \sum_i \frac{\widehat{p_{\text{te}}}}{\widehat{p_{\text{tr}}}} \boldsymbol{f}(\boldsymbol{x}_i)$ is at most $O(1/m)$. Now from Chebyshev's inequality and Eq. (9), for $m \geq O(1/\epsilon^2)$,

$$\Pr \left[\left| \frac{1}{m} \sum_i \frac{\widehat{p_{\text{te}}}}{\widehat{p_{\text{tr}}}} \boldsymbol{f}(\boldsymbol{x}_i) - \mathbb{E}_{\boldsymbol{x} \sim p_{\text{te}}} \boldsymbol{f}(\boldsymbol{x}) \right| \geq \epsilon \right] \leq \frac{Var(Z)}{\epsilon^2} \leq 0.1.$$

We can now use the standard median trick to boost the probability, i.e., repeat the sampling process of sampling m points $O(\log(1/\delta))$ times. Let $\overline{X}$ denote the average of the m samples above. Now, the median trick says generate $\overline{X}_i$ for $i = 1, \cdots, O(\log(1/\delta))$. The median Y of these $\overline{X}_i$'s satisfies with probability at least $1 - \delta$,

$$|Y - \mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{te}}} \boldsymbol{f}(\boldsymbol{x})| \leq \epsilon$$

■

Proof [Proof of Theorem 6] From Lemma 22 it is sufficient to use $O(d/\epsilon^2 \log 1/\delta)$ samples to obtain $\widehat{p_{\mathrm{tr}}}$ and $\widehat{p_{\mathrm{te}}}$ to the required accuracy.

Now, from Lemma 7 our estimator $X = \frac{\widehat{p_{\mathrm{te}}}(\boldsymbol{x})}{\widehat{p_{\mathrm{tr}}}(\boldsymbol{x})} \boldsymbol{f}(\boldsymbol{x})$ for $\boldsymbol{x} \sim p_{\mathrm{tr}}$ satisfies

$$|\mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{tr}}} X - \mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{te}}} \boldsymbol{f}(\boldsymbol{x})| \leq \varepsilon.$$

The remaining proof follows the same as in the proof of Theorem 5. ■

Appendix C. Proofs from Section 3

We will elaborate on Theorem 11, showing that the dependence on B in Theorem 10 is necessary.

Lemma 38 *Let p_{tr} and p_{te} be any pair of distributions s.t. $\Pr_{\boldsymbol{x}\sim p_{\mathrm{te}}}(p_{\mathrm{te}}(\boldsymbol{x})/p_{\mathrm{tr}}(\boldsymbol{x}) \geq B) = 2\varepsilon$. Then, even if the distributions p_{te} and p_{tr} are known, we need $\Omega(B/\varepsilon)$ samples to solve the covariate shifted mean estimation problem for p_{tr} and p_{te}.*

Proof Let $S = \{\boldsymbol{x} : p_{\mathrm{te}}(\boldsymbol{x})/p_{\mathrm{tr}}(\boldsymbol{x}) > B\}$. By assumption, we have $p_{\mathrm{te}}(S) = 2\varepsilon$, and moreover $p_{\mathrm{tr}}(S) \leq 2\varepsilon/B$ by the construction of the set B. The expectations with respect to p_{te} of functions $\mathbf{1}_S$ and the constant 0 function, differ by at least 2ε, so in order to solve the covariate-shifted mean estimation problem, we need to be able to distinguish those two functions. But unless we sample $\Omega(B/\varepsilon)$ elements $\boldsymbol{x}$ from p_{tr} it is unlikely that any of those will be in S, and the algorithm will only have observed value $\boldsymbol{f}(\boldsymbol{x}) = 0$ in either of those two cases. ■

Lemma 39 *Let $p_{\mathrm{tr}}, p_{\mathrm{te}}$ be a pair of distributions, and $S \subseteq U$ be such that $\Pr_{\boldsymbol{x}\sim p_{\mathrm{te}}}(S) \geq 1 - \varepsilon$ and $\Pr_{\boldsymbol{x}\sim p_{\mathrm{tr}}}(S) \geq 1 - \varepsilon$. Let $\widehat{p_{\mathrm{tr}}}$ and $\widehat{p_{\mathrm{te}}}$ be a pair of estimations of $p_{\mathrm{tr}}, p_{\mathrm{te}}$ respectively with $\mathrm{supp}(\widehat{p_{\mathrm{te}}}) \subseteq \mathrm{supp}(\widehat{p_{\mathrm{tr}}}) \subseteq S$. Moreover, assume that $\forall \boldsymbol{x} \in S$, $\widehat{p_{\mathrm{te}}}(\boldsymbol{x})/\widehat{p_{\mathrm{tr}}}(\boldsymbol{x}) \leq B$. Then*

$$\left| \mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{tr}}} \frac{\widehat{p_{\mathrm{te}}}(\boldsymbol{x})}{\widehat{p_{\mathrm{tr}}}(\boldsymbol{x})} \boldsymbol{f}(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{te}}} \boldsymbol{f}(\boldsymbol{x}) \right| \leq B d_{TV}(p_{\mathrm{tr}}, \widehat{p_{\mathrm{tr}}}) + d_{TV}(p_{\mathrm{te}}, \widehat{p_{\mathrm{te}}}).$$

Proof From Theorem 32,

$$\left| \mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{tr}}} \frac{\widehat{p_{\mathrm{te}}}(\boldsymbol{x})}{\widehat{p_{\mathrm{tr}}}(\boldsymbol{x})} \boldsymbol{f}(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{te}}} \boldsymbol{f}(\boldsymbol{x}) \right| \leq Q_1(\widehat{p_{\mathrm{tr}}} \,|\, p_{\mathrm{tr}};\ \widehat{p_{\mathrm{te}}}) + d_{TV}(p_{\mathrm{te}}, \widehat{p_{\mathrm{te}}}).$$

Therefore, it is sufficient to bound $Q_1(\widehat{p_{\text{tr}}} \,|\, p_{\text{tr}};\ \widehat{p_{\text{te}}}) \leq B d_{TV}(\widehat{p_{\text{tr}}}, p_{\text{tr}})$. Since $\text{supp}(\widehat{p_{\text{te}}}) \subseteq S$ and $\forall x \in S, \widehat{p_{\text{te}}}(\boldsymbol{x})/\widehat{p_{\text{tr}}}(\boldsymbol{x}) \leq B$, we have

$$\begin{aligned} Q_1(\widehat{p_{\text{tr}}} \,|\, p_{\text{tr}};\ \widehat{p_{\text{te}}}) &= \mathbb{E}_{\boldsymbol{x}\sim\widehat{p_{\text{te}}}} \left| \frac{p_{\text{tr}}(\boldsymbol{x})}{\widehat{p_{\text{tr}}}(\boldsymbol{x})} - 1 \right| \\ &\leq B\, \mathbb{E}_{\boldsymbol{x}\sim\widehat{p_{\text{tr}}}} \left| \frac{p_{\text{tr}}(\boldsymbol{x})}{\widehat{p_{\text{tr}}}(\boldsymbol{x})} - 1 \right| \\ &= B d_{TV}(p_{\text{tr}}, \widehat{p_{\text{tr}}}). \end{aligned}$$

■

Proof [Proof of Theorem 10] Let $S := \{\boldsymbol{x} : \widehat{p_{\text{te}}}(\boldsymbol{x})/\widehat{p_{\text{tr}}}(\boldsymbol{x}) \leq B\}$, and $S' := \{\boldsymbol{x} : p_{\text{te}}(\boldsymbol{x})/p_{\text{tr}}(\boldsymbol{x}) \leq B/4\}$. Moreover, define sets $A_{\text{te}} := \{\boldsymbol{x} : \widehat{p_{\text{te}}}(\boldsymbol{x}) \leq 2p_{\text{te}}(\boldsymbol{x})\}$, and $A_{\text{tr}} := \{\boldsymbol{x} : p_{\text{tr}}(\boldsymbol{x}) \leq 2p_{\text{tr}}(\boldsymbol{x})\}$.

We would first prove that $p_{\text{te}}(S) \geq 1 - O(\varepsilon)$ and $p_{\text{tr}}(S) \geq 1 - O(\varepsilon)$. Note that since $\frac{\widehat{p_{\text{te}}}}{\widehat{p_{\text{tr}}}} = \frac{p_{\text{te}}}{p_{\text{tr}}} \cdot \frac{\widehat{p_{\text{te}}}}{p_{\text{te}}} \cdot \frac{p_{\text{tr}}}{\widehat{p_{\text{tr}}}}$, $A_{\text{te}} \cap A_{\text{tr}} \cap S' \subseteq S$. Let U denote the entire space. We have

- $p_{\text{te}}(U \setminus A_{\text{te}}) \leq \widehat{p_{\text{te}}}(U \setminus A_{\text{te}}) + \varepsilon \leq O(\varepsilon)$,
- $p_{\text{te}}(S' \setminus A_{\text{tr}}) \leq \frac{B}{4} \cdot p_{\text{tr}}(S' \setminus A_{\text{tr}}) \leq \frac{B}{4} \cdot p_{\text{tr}}(U \setminus A_{\text{tr}}) \leq O(\varepsilon)$,
- $p_{\text{te}}(S') \leq \varepsilon$ by assumption.

This yields that $p_{\text{te}}(S) \geq p_{\text{te}}(A_{\text{te}} \cap A_{\text{tr}} \cap S') \geq 1 - O(\varepsilon)$ and therefore $\widehat{p_{\text{te}}}(S) \geq 1 - O(\varepsilon)$. Furthermore, from the definition of S we also have $\widehat{p_{\text{tr}}}(S^c) < \widehat{p_{\text{te}}}(S^c)/B \leq (p_{\text{te}}(S^c) + \varepsilon)/B \leq O(\varepsilon/B)$. Therefore, we also have, $p_{\text{tr}}(S) \geq 1 - O(\varepsilon)$.

Now, consider the conditional distributions $\widehat{p_{\text{te}}}|_S$ and $\widehat{p_{\text{tr}}}|_S$, which are defined as, $p|_S(\boldsymbol{x}) = p(\boldsymbol{x})$, if $\boldsymbol{x} \in S$ and 0 if $\boldsymbol{x} \notin S$. We get the desired bound on the density ratio on the set S.

$$\frac{\widehat{p_{\text{te}}}|_S(\boldsymbol{x})}{\widehat{p_{\text{tr}}}|_S(\boldsymbol{x})} \leq B$$

Note that conditioning on S doesn't introduce too much error, i.e., if $d_{TV}(p_{\text{te}}, \widehat{p_{\text{te}}}) \leq \varepsilon$ and $p_{\text{te}}(S) \geq 1 - \varepsilon$ then $d_{TV}(p_{\text{te}}, \widehat{p_{\text{te}}}|_S) \leq O(\varepsilon)$ by triangle inequality. Similarly, $d_{TV}(\widehat{p_{\text{tr}}}|_S, p_{\text{tr}}) \leq O(\varepsilon/B)$ and we can apply the lemma Theorem 12 to deduce that, the bias is at most ε, i.e.,

$$\left| \mathbb{E}_{\boldsymbol{x}\sim p_{\text{tr}}} \frac{\widehat{p_{\text{te}}}(\boldsymbol{x})}{\widehat{p_{\text{tr}}}(\boldsymbol{x})} \boldsymbol{f}(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{x}\sim p_{\text{te}}} \boldsymbol{f}(\boldsymbol{x}) \right| \leq O(\varepsilon).$$

Finally, the random variable $Z := \boldsymbol{f}(\boldsymbol{x}) \mathbf{1}_S(\boldsymbol{x}) \frac{\widehat{p_{\text{te}}}|_S(\boldsymbol{x})}{\widehat{p_{\text{tr}}}|_S(\boldsymbol{x})}$ for $\boldsymbol{x} \sim p_{\text{te}}$ is always bounded by B, hence $Var(Z) \leq B$. The final bound now follows directly from the application of the Chebyshev inequality. Further, we can boost the probability using the standard median trick.

This concludes the proof of Theorem 10. ■

Appendix D. Proofs from Section 4

As discussed in Section 4, we can connect the population regret of a classifier with respect to the cross-entropy loss, with KL-divergence between some pair of distributions with estimated density ratio, and actual test and training distributions.

Lemma 40 *Let $\mathcal{D}_0$ be a distribution described above, $r^*(\boldsymbol{x}) := \Pr_{\mathcal{D}_0}[y = -1|\boldsymbol{x}]$ and consider a classifier $\hat{r} : U \to (0, 1)$, intended to estimate r^*.*

Let $\mathcal{L}_{\mathcal{D}_0,\hat{r}}$ be an expected cross-entropy loss of classifier $\hat{r}$, namely

$$\mathcal{L}_{\mathcal{D},\hat{r}} := -\mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\mathbf{1}_{y=-1}\log\hat{r}(\boldsymbol{x}) + \mathbf{1}_{y=1}\log(1-\hat{r}(\boldsymbol{x}))].$$

Then there exist a pair of distributions $\widehat{p_{\mathrm{tr}}}, \widehat{p_{\mathrm{te}}}$ such that the population regret for a classifier $\hat{r}$ is equal

$$\mathcal{R}_{\mathcal{D},\hat{r}} := \mathcal{L}_{\mathcal{D},\hat{r}} - \mathcal{L}_{\mathcal{D},r^*} = \frac{1}{2}(D_{KL}(p_{\mathrm{te}}||\widehat{p_{\mathrm{te}}}) + D_{KL}(p_{\mathrm{tr}}||\widehat{p_{\mathrm{tr}}})) + (1 - H(y')),$$

where $H(y') \le 1$ is a Shannon entropy of some binary random variable. In particular

$$D_{KL}(p_{\mathrm{te}}||\widehat{p_{\mathrm{te}}}) + D_{KL}(p_{\mathrm{tr}}||\widehat{p_{\mathrm{tr}}}) \le 2\mathcal{R}_{\hat{r}}.$$

Proof Consider a joint distribution $\mathcal{D}_1$ over $(\boldsymbol{x}, y')$, s.t. the marginal distribution of $\boldsymbol{x}$ is the same as in $\mathcal{D}$, i.e. $\boldsymbol{x} \sim \frac{1}{2}p_{\mathrm{tr}} + \frac{1}{2}p_{\mathrm{te}}$, and $\Pr[y' = -1|\boldsymbol{x}] := \hat{r}(\boldsymbol{x})$.

We have

$$\begin{aligned}\mathcal{L}_{\hat{r}} - \mathcal{L}_{r^*} &= \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}_0}\,\mathbf{1}_{y=-1}(\log(\hat{r}(\boldsymbol{x})) - \log(r^*(\boldsymbol{x}))) + \mathbf{1}_{y=1}(\log(1-\hat{r}(\boldsymbol{x})) - \log(1-r^*(\boldsymbol{x})))\\ &= \frac{1}{2}\,\mathbb{E}_{x\sim p_{\mathrm{tr}}}\log\frac{\hat{r}(\boldsymbol{x})}{r^*(\boldsymbol{x})} + \frac{1}{2}\,\mathbb{E}_{x\sim p_{\mathrm{te}}}\log\frac{1-\hat{r}(\boldsymbol{x})}{1-r^*(\boldsymbol{x})}. \qquad (10)\end{aligned}$$

Now $\hat{r}(\boldsymbol{x}) = \Pr[y' = -1|\boldsymbol{x}] = \Pr[y' = -1, \boldsymbol{x}]/Pr[\boldsymbol{x}]$, and similarly $r(\boldsymbol{x}) = \Pr[y' = -1, \boldsymbol{x}]/\Pr[\boldsymbol{x}]$. Hence

$$\frac{\hat{r}(\boldsymbol{x})}{r^*(\boldsymbol{x})} = \frac{\Pr[y' = -1, \boldsymbol{x}]}{\Pr[y = -1, \boldsymbol{x}]} = \frac{\Pr[y' = -1]}{\Pr[y = -1]}\frac{p_{\mathrm{tr}}(\boldsymbol{x})}{\widehat{p_{\mathrm{tr}}}(\boldsymbol{x})} = 2\Pr[y' = -1]\frac{p_{\mathrm{tr}}(\boldsymbol{x})}{\widehat{p_{\mathrm{tr}}}(\boldsymbol{x})},$$

and analogously

$$\frac{1-\hat{r}(\boldsymbol{x})}{1-r^*(\boldsymbol{x})} = 2\Pr[y' = 1]\frac{p_{\mathrm{te}}(\boldsymbol{x})}{\widehat{p_{\mathrm{te}}}(\boldsymbol{x})}.$$

Plugging this back to (10), we get

$$\begin{aligned}\mathcal{L}_{\hat{r}} - \mathcal{L}_{r^*} &= \frac{1}{2}\,\mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{tr}}}\log\frac{p_{\mathrm{tr}}(\boldsymbol{x})}{\widehat{p_{\mathrm{tr}}}(\boldsymbol{x})} + \mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{te}}}\log\frac{p_{\mathrm{te}}(\boldsymbol{x})}{\widehat{p_{\mathrm{te}}}(\boldsymbol{x})} + \frac{1}{2}\log\Pr[y' = -1] + \frac{1}{2}\log\Pr[y' = -1] + 1\\ &= \frac{1}{2}D_{KL}(p_{\mathrm{tr}}||\widehat{p_{\mathrm{tr}}}) + D_{KL}(p_{\mathrm{te}}||\widehat{p_{\mathrm{te}}}) + D_{KL}(y||y').\end{aligned}$$

■

D.1. Using logistic regression — exponential families

When the distributions p_{te} and p_{tr} are from the same exponential family, we can use logistic regression to find the classifier provides the right density ratio estimate, for which the Theorem 40 applies. In particular, for parameters $\boldsymbol{\theta}_{\mathrm{te}}$ and $\boldsymbol{\theta}_{\mathrm{tr}}$, and functions $\boldsymbol{h}, \boldsymbol{T}$, let $p_{\mathrm{te}} = \boldsymbol{h}(\boldsymbol{x}) \cdot \exp\left(\boldsymbol{\theta}_{\mathrm{te}}^\top \boldsymbol{T}(\boldsymbol{x}) - A(\boldsymbol{\theta}_{\mathrm{te}})\right)$, and $p_{\mathrm{tr}} = \boldsymbol{h}(\boldsymbol{x}) \cdot \exp\left(\boldsymbol{\theta}_{\mathrm{tr}}^\top \boldsymbol{T}(\boldsymbol{x}) - A(\boldsymbol{\theta}_{\mathrm{tr}})\right)$.

For these distributions,

$$\Pr_{(\boldsymbol{x},y)\sim\mathcal{D}_0}[y=-1|\boldsymbol{x}] = \frac{1}{1+p_{\mathrm{te}}(\boldsymbol{x})/p_{\mathrm{tr}}(\boldsymbol{x})} = \frac{1}{1+\exp\left(\boldsymbol{T}(\boldsymbol{x})^\top(\boldsymbol{\theta}_{\mathrm{te}}-\boldsymbol{\theta}_{\mathrm{tr}}) + A(\boldsymbol{\theta}_{\mathrm{te}}) - A(\boldsymbol{\theta}_{\mathrm{tr}})\right)}$$

The logistic model is given by a joint distribution $(\boldsymbol{x}, y)$, parameterized by $\boldsymbol{\theta}^* \in \mathbb{R}^d$ and $s^* \in \mathbb{R}$, where

$$\Pr[y=-1|\boldsymbol{x}] = \frac{1}{1+\exp(\langle\boldsymbol{\theta}^*, \boldsymbol{x}\rangle + s^*)}. \tag{11}$$

Hence, the distribution $\mathcal{D}_1$ of $(\boldsymbol{T}(\boldsymbol{x}), y)$ for $(\boldsymbol{x}, y) \sim \mathcal{D}_0$ indeed follows the logistic model, Eq.(11) with parameters $\boldsymbol{\theta}^* = \boldsymbol{\theta}_{\mathrm{te}} - \boldsymbol{\theta}_{\mathrm{tr}}$ and $s^* = A(\boldsymbol{\theta}_{\mathrm{te}}) - A(\boldsymbol{\theta}_{\mathrm{tr}})$. Note that for the distribution following the logistic model we have

$$\Pr[y=1|\boldsymbol{x}] = \frac{1}{1+\exp(-\langle\boldsymbol{\theta}, \boldsymbol{x}\rangle - s)},$$

and more succinctly

$$\Pr[y|\boldsymbol{x}] = (1+\exp(-y(\langle\boldsymbol{\theta}, \boldsymbol{x}\rangle + s)))^{-1}.$$

Since the empirical minimization of negative log-likelihood for logistic regression is a convex problem, we can efficiently find a minimizer on a finite sample. Using a Rademacher complexity bound for the associated class of loss functions, the minimizer on a finite sample will be an approximate minimizer with respect to the population loss. Concretely we can find $(\hat{\boldsymbol{\theta}}, \hat{s})$ with population loss ε close to optimal, i.e., regret bounded by ε, where the loss is the negative log-likelihood. This regret bound, by Theorem 40, lets us control the total KL-divergence $D_{KL}(p_{\mathrm{te}}||\widehat{p_{\mathrm{te}}}) + D_{KL}(p_{\mathrm{tr}}||\widehat{p_{\mathrm{tr}}})$. By Pinsker inequality, this implies TV-closeness between the respective distributions and our approximations. We can then invoke Theorem 10, to conclude that after truncation, the density ratio $\widehat{p_{\mathrm{te}}}/\widehat{p_{\mathrm{tr}}}$ can be used as importance weights for the covariate-shifted mean estimation problem.

Detailed proof of Theorem 13 As discussed above, in order to prove Theorem 13, we will use Theorem 40, Pinsker inequality and Theorem 10. In order to implemenent this strategy, we need to first show the regret bounds for logistic regression.

As is standard in this context, we will bound the Rademacher complexity of the class $\{\ell_{\boldsymbol{\theta}} : \|\boldsymbol{\theta}\| \leq B_2, s \leq B_2\}$, where

$$\ell_{\boldsymbol{\theta},s}(\boldsymbol{x}, y) = -\log(1+\exp(-y(\langle\boldsymbol{\theta}, \boldsymbol{x}\rangle + s))). \tag{12}$$

We show the following lemma using methods that are now standard, but many of the proofs of similar statements in the literature use stronger assumptions on the distribution of $\boldsymbol{x}$ (for example, assuming that almost surely $\|\boldsymbol{x}\| \leq C\sqrt{\|\boldsymbol{\Sigma}\|_{op}}$), in order to provide better concentration results.

Lemma 41 *The Radeamcher complexity of the class $\mathcal{F} := \{\boldsymbol{x} \mapsto \langle\boldsymbol{\theta}, \boldsymbol{x}\rangle : \|\boldsymbol{\theta}\| \leq B\}$ is at most $B\sqrt{\operatorname{Tr}\boldsymbol{\Sigma}}/\sqrt{n}$ where $\boldsymbol{\Sigma} := \mathbb{E}_{\boldsymbol{x}\sim p}\, \boldsymbol{x}\boldsymbol{x}^T$ is the covariance matrix.*

Proof We can bound the Rademacher complexity of $\mathcal{F}$ (taking σ_i to be independent Rademacher valued random variables) as follows

$$\begin{aligned} n\mathcal{R}_{n,\mathcal{D}}(\mathcal{F}) &= \mathbb{E}_{\boldsymbol{x},\sigma} \sup_{\|\boldsymbol{\theta}\|\le B} \sum_i \sigma_i \langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle \\ &= \mathbb{E}_{\boldsymbol{x},\sigma} \sup_{\|\theta\|\le B} \left\langle \sum_i \sigma_i \boldsymbol{x}_i, \theta \right\rangle \\ &= \mathbb{E}_{\boldsymbol{x},\sigma}\, B \left\| \sum_i \sigma_i \boldsymbol{x}_i \right\| \le B\sqrt{\mathbb{E}\left\|\sum \sigma_i x_i\right\|^2} \end{aligned}$$

Now, since $\mathbb{E}\,\sigma_i\sigma_j = 0$ for $i \neq j$, we get

$$\mathbb{E}\left\|\sum_i \sigma_i x_i\right\|^2 = \sum_i \mathbb{E}\|\boldsymbol{x}_i\|^2 = n \operatorname{Tr} \boldsymbol{\Sigma}.$$

Combining those two, we get

$$\mathcal{R}_{n,\mathcal{D}}(\mathcal{F}) \le B\frac{\sqrt{\operatorname{Tr}\boldsymbol{\Sigma}}}{n},$$

■

Since composition with a Lipschitz function does not increase the Rademacher complexity (Theorem 29) we get for any Lipschitz function γ, the class $\{\boldsymbol{x} \mapsto \gamma(\langle \boldsymbol{\theta}, \boldsymbol{x}\rangle) : \|\boldsymbol{\theta}\| \le B\}$ is also bounded by $B\sqrt{\operatorname{Tr}\boldsymbol{\Sigma}}/\sqrt{n}$.

We can now lift this bound to a bound on the class of relevant loss functions.

Lemma 42 *Let $\mathcal{F}'$ be any family of functions from $U \to \mathbb{R}$. Consider a family of functions $\mathcal{F} \subset U \times \{0,1\} \to \mathbb{R}$ s.t. for every $f \in \mathcal{F}$ and every $y \in \{0,1\}$ we have $f(\cdot, y) \in \mathcal{F}'$. Let $\mathcal{D}$ be a distribution over $U \times \{0,1\}$, s.t. $\mathcal{D}_{\boldsymbol{x}}$ is the marginal of $\mathcal{D}$ on U. Then*

$$\mathcal{R}_{n,\mathcal{D}}(\mathcal{F}) \lesssim \mathcal{R}_{n,\mathcal{D}_{\boldsymbol{x}}}(\mathcal{F}').$$

Proof We can write arbitrary function $f \in \mathcal{F}$ as $f(\boldsymbol{x}, y) = yf_1(\boldsymbol{x}) + (1-y)f_0(\boldsymbol{x})$, where $f_0, f_1 \in \mathcal{F}'$. Then

$$\begin{aligned} \mathcal{R}_{n,\mathcal{D}}(\mathcal{F}) &= \mathbb{E}_{(\boldsymbol{x},y),\sigma} \sup_{f\in\mathcal{F}} \sum_i \sigma_i f(\boldsymbol{x}_i, y_i) \\ &\le \mathbb{E}_{(\boldsymbol{x},y),\sigma} \sup_{f_1\in\mathcal{F}'} \sum_i \sigma_i y_i f_0(x_i) + \mathbb{E}_{(\boldsymbol{x},y),\sigma} \sup_{f_0\in\mathcal{F}'} \sum_i \sigma_i (1-y_i) f_1(x_i). \end{aligned}$$

Now, all we need to show is that for a fixed sample $(\boldsymbol{x}_1, y_1), \ldots (\boldsymbol{x}_n, y_n)$ we have

$$\mathbb{E}_\sigma \sup_{f\in\mathcal{F}'} \sigma_i y_i f(\boldsymbol{x}_i) \le \mathbb{E}_\sigma \sup_{f\in\mathcal{F}'} \sum \sigma_i f(\boldsymbol{x}_i).$$

That is a special case of Talagrand's contraction principle (Theorem 4.12 in Ledoux and Talagrand (2013)), since for each i we have $|y_i f(\boldsymbol{x}_i)| \le |f(\boldsymbol{x}_i)|$. ■

Lemma 43 *Let $\mathcal{D}$ be a distribution over pairs $(\boldsymbol{x}, y)$ with $\boldsymbol{x} \in U, y \in \{\pm 1\}$, and let $\boldsymbol{\Sigma} = \mathbb{E}\, \boldsymbol{x}\boldsymbol{x}^T$ be a covariance matrix of the marginal $\boldsymbol{x}$. Consider a sample $S = ((\boldsymbol{x}_1, y_1), \dots (\boldsymbol{x}_n, y_n))$ drawn i.i.d. from $\mathcal{D}$. Then*

$$\mathbb{E}_{S\sim\mathcal{D}^n} \sup_{\|\boldsymbol{\theta}\|\le B, s\le B} |\widehat{\mathcal{L}}_{\boldsymbol{\theta},s}(S) - \mathcal{L}_{\boldsymbol{\theta},s}(\mathcal{D})| \le \frac{B\sqrt{\operatorname{Tr}\boldsymbol{\Sigma}}}{\sqrt{n}}.$$

where $\mathcal{L}_{\boldsymbol{\theta},s}(\mathcal{D}) := \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}\, \ell_{\boldsymbol{\theta},s}$ and $\widehat{\mathcal{L}}_{\boldsymbol{\theta},s}(S) := \frac{1}{n}\sum_i \ell_{\boldsymbol{\theta},s}(\boldsymbol{x}_i, y_i)$.

Proof Combining Theorem 41 and Theorem 42, together with the observation that $t \mapsto \log(1 + \exp(t + s))$ is a 1-Lipschitz function, we obtain the following corollary, where as a reminder, we consider a family of loss functions $\ell_{\theta,s}(\boldsymbol{x}, y) := \log(1 + \exp(y\langle\theta, \boldsymbol{x}\rangle + s))$.

Corollary 44 *Family of functions $\mathcal{F}_\ell := \{\ell_{\theta,s}(\boldsymbol{x}, y) : \|\boldsymbol{\theta}\| \le B_2, |s| \le B_2\}$, with respect to the distribution $\mathcal{D}_0$ as in Section 4, has Rademacher complexity bounded as*

$$\mathcal{R}_{n,\mathcal{D}_0}(\mathcal{F}_\ell) \lesssim B_2 \frac{\sqrt{\operatorname{Tr}(\boldsymbol{\Sigma}_{\mathrm{tr}} + \boldsymbol{\Sigma}_{\mathrm{te}})}}{\sqrt{n}}.$$

This, together with Theorem 30 and Markov inequality completes the proof of Theorem 43. ■

Finally, since for a given sample S, a function $(\boldsymbol{\theta}, s) \mapsto \widehat{\mathcal{L}}_{\boldsymbol{\theta},s}(S)$ is convex, we can efficiently find an approximate minimum of this function. That is,

Fact 45 *Given a sample $S = (\boldsymbol{x}_1, y_1), \dots (\boldsymbol{x}_n, y_n)$ and ε' we can find in polynomial time $\hat{\boldsymbol{\theta}}, \hat{s}$ such that*

$$\widehat{\mathcal{L}}_{\hat{\boldsymbol{\theta}},\hat{s}}(S) \le \min_{\boldsymbol{\theta},s} \widehat{\mathcal{L}}_{\boldsymbol{\theta},S} + \varepsilon'.$$

We are now ready to prove Theorem 13.

Proof [Proof of Theorem 13] Take ε', depending on $\varepsilon, B_1, B_2, B_3$, which we will fix later. Using Fact 45, we can find $\hat{\boldsymbol{\theta}}, \hat{s}$ s.t. $\widehat{\mathcal{L}}_{\hat{\boldsymbol{\theta}},\hat{s}}(S) \le \varepsilon' + \min_{\boldsymbol{\theta},s} \widehat{\mathcal{L}}_{\boldsymbol{\theta},s}(S)$. Using Markov inequality for $n \gtrsim \varepsilon'^{-2} B_2 \operatorname{Tr}(\boldsymbol{\Sigma}_{\mathrm{tr}} + \boldsymbol{\Sigma}_{\mathrm{te}})$, Theorem 43 implies that with probability $9/10$ simultaneously for all $\boldsymbol{\theta}, s$ we have $\mathcal{L}_{\boldsymbol{\theta},s}(\mathcal{D}) = \widehat{\mathcal{L}}_{\boldsymbol{\theta},s}(S) \pm \varepsilon'$. Combining those inequalities, we get

$$\mathcal{L}_{\hat{\boldsymbol{\theta}},\hat{s}}(\mathcal{D}) \le \varepsilon' + \widehat{\mathcal{L}}_{\hat{\boldsymbol{\theta}},\hat{s}}(S) \le 2\varepsilon' + \widehat{\mathcal{L}}_{\boldsymbol{\theta}^*,s^*}(S) \le 3\varepsilon' + \mathcal{L}_{\boldsymbol{\theta}^*,s^*}(\mathcal{D}).$$

This implies bound on population regret

$$\mathcal{L}_{\hat{\boldsymbol{\theta}},\hat{s}}(\mathcal{D}) - \mathcal{L}_{\boldsymbol{\theta}^*,s^*}(\mathcal{D}) \le 3\varepsilon',$$

and by Theorem 40, this implies

$$\max(D_{KL}(p_{\mathrm{te}}||\widehat{p_{\mathrm{te}}}), D_{KL}(p_{\mathrm{tr}}||\widehat{p_{\mathrm{tr}}})) \lesssim \varepsilon'.$$

By Pinsker inequality (Theorem 20), we can deduce $d_{TV}(p_{\mathrm{te}}, \widehat{p_{\mathrm{te}}}) \lesssim \sqrt{\varepsilon'}$ and similarly $d_{TV}(p_{\mathrm{te}}, \widehat{p_{\mathrm{tr}}}) \lesssim \sqrt{\varepsilon'}$. If $R_2(p_{\mathrm{tr}}||p_{\mathrm{te}}) \le B_3$, then $\mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{te}}}\, p_{\mathrm{te}}(\boldsymbol{x})/p_{\mathrm{tr}}(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{tr}}}(p_{\mathrm{te}}(\boldsymbol{x})/p_{\mathrm{tr}}(\boldsymbol{x}))^2 \le B_3$ and by Markov inequality $\Pr_{\boldsymbol{x}\sim p_{\mathrm{te}}}(p_{\mathrm{te}}(\boldsymbol{x})/p_{\mathrm{tr}}(\boldsymbol{x}) > B_3/\varepsilon) \le \varepsilon$. We can now apply Theorem 10, with

$B = B_3/\varepsilon$ to deduce the desired statement. To this end, we need to choose $\varepsilon' \approx (\varepsilon/B)^2 \lesssim \varepsilon^4/B_3^2$, such that the assumption of Theorem 10 are satisfied.

This choice of ε', yields sample complexity used in the logistic regression phase of the argument as $n \approx B_2^2 \operatorname{Tr}(\Sigma_{tr} + \Sigma_{te})(\varepsilon')^{-2} = B_1 B_2^2 B_3^4 \varepsilon^{-8}$.

We need additional $O(B^2/\varepsilon^2) = O(B_3^2\varepsilon^3)$ samples to apply estimation in Algorithm 3, which is negligible. The total sample complexity is then $O(B_1 B_2^2 B_3^4 \varepsilon^{-6})$ as desired. ■

D.2. Improved Bounds for Gaussians

Error bounds on the estimated parameter $\hat{\theta}$ in the Gaussian case. In the case where both distributions are d-dimensional Gaussian vectors with the same known covariance, the sample complexity bound obtained by Theorem 13 is easily seen to be suboptimal. For instance, when $p_{\mathrm{tr}} = \mathcal{N}(\boldsymbol{\mu}_{\mathrm{tr}}, \boldsymbol{I})$, $p_{\mathrm{te}} = \mathcal{N}(\boldsymbol{\mu}_{\mathrm{te}}, \boldsymbol{I})$, s.t. $\|\boldsymbol{\mu}_{\mathrm{tr}} - \boldsymbol{\mu}_{\mathrm{te}}\| \leq O(1)$ according to Theorem 13, we can solve the covariate-shifted mean estimation problem using $O(d/\varepsilon^8)$ samples (in this case we would have $B_1 \approx d, B_2 \leq O(1)$, and $B_3 \leq O(1)$). In contrast Theorem 6 shows that by learning the means of both Gaussian distributions separately, we can solve the covariate shift problem using $O(d/\varepsilon^2)$ samples.

As it turns out, specifically in the case of Gaussians, we can improve the sample complexity in Theorem 13, and we can recover the importance weights $\hat{r}$, using only $O(\max\{d \log d, d/\varepsilon^2\})$ samples using logistic regression, nearly matching the guarantees of Theorem 6.

This is a consequence of new results for the error bounds on the *estimated parameter* $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|$ in logistic regression, as opposed to classical bounds showing only that the (population) loss at $\hat{\boldsymbol{\theta}}$ is not much larger than the minimum value attained at $\boldsymbol{\theta}^*$. Bounds of this form are only known when covariate $\boldsymbol{x}$ itself is a sub-Gaussian random variable, so this statement is much less general then Theorem 13, but provides much stronger results in this scenario. We present these improvements in Appendix D.2

Theorem 46 (Bach (2021); Ostrovskii and Bach (2018)) *Let $(\boldsymbol{x}, y)$ follow the logistic model with parameter $\boldsymbol{\theta}^*$, s.t. $\|\boldsymbol{\theta}^*\| \leq O(1)$. Then using logistic regression by running an empirical minimization of the negative log-likelihood in parameter space $\|\boldsymbol{\theta}\| \leq C$, after observing $n \gtrsim d \log d$ samples, the estimated parameter $\hat{\boldsymbol{\theta}}$ satisfies*

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2 \leq \frac{d}{n}.$$

In our case, the covariate $\boldsymbol{x}$ follows a mixture of two Gaussian distributions. We will show that such a random variable is indeed subgaussian, and therefore we can apply Theorem 46

Fact 47 *Let $\boldsymbol{x}$ be distributed according to the mixture of two Gaussian distributions $\boldsymbol{x} \sim \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{I}) + \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{I})$, where $\|\boldsymbol{\mu}_1\|, \|\boldsymbol{\mu}_2\| \leq O(1)$. Then $\boldsymbol{x}$ is $O(1)$-subgaussian.*

Proof Taking $Z = \langle \boldsymbol{x}, v \rangle$ for a unit vector v, we reduce to one-dimensional case: Z is a mixture of two univariate gaussians.

Then

$$\begin{aligned}\mathbb{E}\exp(\lambda(Z-\mathbb{E}\,Z)) &= \exp(-\lambda(\mu_1+\mu_2)/2)\frac{\mathbb{E}\exp(\lambda Z_1)+\mathbb{E}\exp(\lambda Z_2)}{2}\\ &= \exp(-\lambda(\mu_1+\mu_2)/2)\frac{\exp(\lambda\mu_1+O(\lambda^2))+\exp(\lambda\mu_2+O(\lambda^2)}{2}\\ &= \exp(O(\lambda^2))\frac{\exp(\lambda\Delta)+\exp(-\lambda\Delta)}{2},\end{aligned}$$

where $\Delta=(\mu_1-\mu_2)/2$. If σ is a Rademacher random variable independent of Z, then

$$\frac{\exp(\lambda\Delta)+\exp(-\lambda\Delta)}{2}=\mathbb{E}\exp(\lambda\Delta\sigma)=\exp(O(\lambda^2\Delta^2),$$

using a well-known fact that Rademacher random variables are $O(1)$-subgaussian. This implies

$$\mathbb{E}\exp(\lambda(Z-\mathbb{E}\,Z))\le\exp(O(\lambda^2))\exp(O(\lambda^2\Delta^2)=\exp(O(\lambda^2)),$$

when $\Delta=O(1)$. ■

We can now show that a bound on covariate shift, using $\hat{\boldsymbol{w}}(x):=\exp(\langle\hat{\boldsymbol{\theta}},\boldsymbol{x}\rangle)$ as an estimate for the true density ratio $p_{\rm te}(\boldsymbol{x})/p_{\rm tr}(\boldsymbol{x})=\exp(\langle\boldsymbol{\theta}^*,\boldsymbol{x}\rangle)$.

Lemma 48 *Taking $\hat{\boldsymbol{w}}(\boldsymbol{x})$ as above, if $\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|\le\varepsilon$, then for any bounded function $\boldsymbol{f}$, we have*

$$|\mathbb{E}_{\boldsymbol{x}\sim p_{\rm tr}}\,\hat{\boldsymbol{w}}(\boldsymbol{x})\boldsymbol{f}(\boldsymbol{x})-\mathbb{E}_{\boldsymbol{x}\sim p_{\rm te}}\,\boldsymbol{f}(\boldsymbol{x})|\lesssim\varepsilon.$$

Proof Taking $w(x):=p_1(x)/p_2(x)$, we can bound

$$\begin{aligned}|\mathbb{E}_{x\sim p_2}\,\hat{w}(x)f(x)-\mathbb{E}_{x\sim p_1}\,f(x)| &\le \mathbb{E}_{x\sim p_2}|\hat{w}(x)-w(x)|\\ &= \mathbb{E}_{x\sim p_2}\,w(x)\left|1-\frac{\hat{w}(x)}{w(x)}\right|\\ &= \mathbb{E}_{x\sim p_1}\left|1-\frac{\hat{w}(x)}{w(x)}\right|.\end{aligned}$$

Now, just by definition $\hat{w}(x)/w(x)=\exp(\langle\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*,x\rangle)$. Taking $Z:=\langle\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*,x\rangle$, since $x\sim p_1$ is a multivariate Gaussian with mean $\boldsymbol{\mu}_1$ and covariance $\boldsymbol{I}$, Z is a univariate gaussian with mean $\mathbb{E}\,Z=\langle\boldsymbol{\theta}^*-\hat{\boldsymbol{\theta}},\boldsymbol{\mu}_1\rangle\le O(\|\boldsymbol{\theta}^*-\hat{\boldsymbol{\theta}}\|_2)\le O(\varepsilon)$ and variance $Var(Z)=\|\boldsymbol{\theta}^*-\hat{\boldsymbol{\theta}}\|_2^2\le\varepsilon^2$.

The bound on $\mathbb{E}|1-\exp(Z)|\le O(\varepsilon)$ is a simple calculation proven after this proof. ■

Lemma 49 *Let Z be a univariate gaussian with $|\mathbb{E}\,Z|\le\varepsilon$ and $Var(Z)\le\varepsilon^2$. Then*

$$\mathbb{E}|1-\exp(Z)|\lesssim\varepsilon.$$

Proof Let $\mu:=\mathbb{E}\,Z$, $s:=\exp(\mu)-1$, $\tilde{Z}=Z-\mathbb{E}\,Z$ and $\sigma:=\sqrt{\mathbb{E}\,\tilde{Z}^2}$. We have

$$\begin{aligned}\mathbb{E}|1-\exp(Z)| &= \mathbb{E}|1-\exp(\tilde{Z})(1+s)|\\ &= \mathbb{E}|1-\exp(\tilde{Z})|+s\,\mathbb{E}|\exp(\tilde{Z})|\\ &\le \mathbb{E}|1-\exp(\tilde{Z}|)+s\sqrt{\mathbb{E}\exp(2\tilde{Z})}\\ &\le \mathbb{E}|1-\exp(\tilde{Z})|+O(\varepsilon).\end{aligned}$$

We can now bound

$$\begin{aligned}\mathbb{E}(1-\exp(\tilde{Z}))^2 &= 1+\exp(2\tilde{Z})-2\exp(\tilde{Z})\\ &= 1+\exp(4\sigma^2)-2\exp(\sigma^2)\\ &= 1+(1+O(\sigma^2))-2(1+O(\sigma^2))\\ &\le O(\sigma^2).\end{aligned}$$

Which gives

$$\mathbb{E}|1-\exp(\tilde{Z})| \le \|1-\exp(\tilde{Z})\|_2 \le O(\sigma). \tag{13}$$

Hence $\mathbb{E}|1-\exp(Z)| \le O(\sigma)+O(\varepsilon) = O(\varepsilon)$

■

Lemma 50 *Let f be a bounded function, and $Z := \hat{w}(\boldsymbol{x})\boldsymbol{f}(\boldsymbol{x})$ for $\boldsymbol{x} \sim p_{\mathrm{tr}}$ and $\hat{w}$ as above. Then $Var(Z) \le O(1)$.*

Proof Indeed,

$$Var(Z) \le \mathbb{E}\, Z^2 \le \mathbb{E}_{x\sim p_1}\, \hat{w}(x) = \mathbb{E}_{x\sim p_1} \exp(\langle 2\hat{\theta}, x\rangle) = \exp(2\hat{\theta}\mu_1 + 4\|\hat{\theta}\|^2) = O(1).$$

■

Appendix E. Proofs from Section 5

E.1. Proof of Theorem 15

As usual, instead of minimizing the population loss directly, we minimize the empirical loss. For a sample $S = ((\boldsymbol{x}_1, y_1), \dots (\boldsymbol{x}_n, y_n))$ drawn i.i.d. from $\mathcal{D}_{\boldsymbol{\theta}^*}$ the empirical loss is defined as

$$\widehat{\mathcal{L}}_{\boldsymbol{\theta},S} := \frac{1}{n}\sum_i \log(1+\exp(y_i\boldsymbol{\theta}(\boldsymbol{x}_i))),$$

Lemma 51
Given sample $S = ((\boldsymbol{x}_1, y_1), \dots (\boldsymbol{x}_n, y_n))$ and desired accuracy ε' we can efficiently find $\hat{\boldsymbol{\theta}}$, s.t.

$$\widehat{\mathcal{L}}_{\hat{\boldsymbol{\theta}}}(S) - \min_{\boldsymbol{\theta}} \widehat{\mathcal{L}}_{\boldsymbol{\theta}}(S) \le \varepsilon'.$$

Proof By the Representer Theorem, the optimal θ can be written as a linear combination

$$\hat{\theta}^* := \sum_i \gamma_i \Phi_{x_i}. \tag{14}$$

Finding the minimizer of this form of the empirical loss function (14) is just a finite-dimensional convex problem

$$\min_{\gamma\in\mathbb{R}^n} \sum_i \log(1+\exp(y_i\langle\gamma, K_i\rangle)),$$

where K_i is the i-th row of the kernel matrix $K_{i,j} := K(x_i, x_j)$, and as such it can be solved efficiently. ■

When the sample size is large enough, the standard Rademacher-complexity based considerations leads to a generalization bound. The proof is almost identical to the one of Theorem 43. Instead of direct calculation for the Rademacher complexity of the class of all bounded linear forms, we need to show the corresponding bound for Rademacher complexity of the kernel classes. Variants of this lemma appear for example in Bartlett and Mendelson (2002).

Lemma 52 *If $\mathcal{H}$ is an RKHS of functions over U with kernel $K : U \times U \to \mathbb{R}$, and $\mathcal{F} := \{\boldsymbol{\theta} \in \mathcal{H} : \|\boldsymbol{\theta}\|_{\mathcal{H}} \leq B\}$, then*

$$\mathcal{R}_{n,\mathcal{D}}(\mathcal{F}) \leq \sqrt{\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}} K(\boldsymbol{x}, \boldsymbol{x})} B/\sqrt{n}.$$

Proof For given $\boldsymbol{x}$, let $\Phi_{\boldsymbol{x}} \in \mathcal{H}$ be such that $\langle \Phi_{\boldsymbol{x}}, \theta\rangle_{\mathcal{H}} = \boldsymbol{\theta}(\boldsymbol{x})\rangle_{\mathcal{H}}$. Sampling $\boldsymbol{x}_1, \ldots \boldsymbol{x}_n$ i.i.d. from $\mathcal{D}$ and, $\sigma_1, \ldots \sigma_n$ independent Rademacher random variables, we have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x},\sigma} \sup_{\boldsymbol{\theta}} \sum_i \sigma_i \boldsymbol{\theta}(x_i) &= \mathbb{E}_{x,\sigma} \sup_{\boldsymbol{\theta}} \langle \sum_i \sigma_i \Phi_{x_i}, \boldsymbol{\theta}\rangle \\ &\leq B\, \mathbb{E}_{x,\sigma} \|\sum_i \sigma_i \Phi_{x_i}\|_{\mathcal{H}}\end{aligned}$$

By Jensen inequality

$$\mathbb{E}_{\boldsymbol{x},\sigma} \|\sum_i \sigma_i \Phi_{x_i}\|_{\mathcal{H}} \leq \sqrt{\mathbb{E}_{\boldsymbol{x},\sigma} \|\sum_i \sigma_i \Phi_{\boldsymbol{x}_i}\|^2} = \sqrt{\mathbb{E}_{\boldsymbol{x}} \sum_i \langle \Phi_{\boldsymbol{x}_i}, \Phi_{\boldsymbol{x}_i}\rangle} = \sqrt{n}\sqrt{\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}} K(\boldsymbol{x}, \boldsymbol{x})}.$$

Combining those two we get

$$\frac{1}{n} \mathbb{E}_{\boldsymbol{x},\sigma} \sup_{\theta} \sum_i \sigma_i \theta(x_i) \leq \frac{B}{\sqrt{n}} \sqrt{\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}} K(\boldsymbol{x}, \boldsymbol{x})}.$$

■

The Rademacher complexity bound above, as usual, implies the following uniform bound for the error between population loss an empirical loss of the kernel logistic regression classifier.

Lemma 53 *Let $\mathcal{H}$ be a RKHS with kernel K, and assume that $\mathbb{E}_{\boldsymbol{x}} K(\boldsymbol{x}, \boldsymbol{x}) \leq R^2$. Then when $n \gtrsim R/\varepsilon^2$, with probability at least $9/10$ with respect to random sample S we have*

$$\sup_{\|\boldsymbol{\theta}\|_{\mathcal{H}} \leq B} |\mathcal{L}_{\boldsymbol{\theta}}(S) - \widehat{\mathcal{L}}_{\boldsymbol{\theta}}(S)| \leq \varepsilon.$$

Proof Combining Theorem 52 and Theorem 42, together with the observation that $t \mapsto \log(1 + \exp(t+s))$ is a 1-Lipschitz function, we get that for an RKHS $\mathcal{H}$, the family $\{(\boldsymbol{x}, y) \mapsto \ell_{\boldsymbol{\theta}}(\boldsymbol{x}, y) : \theta \in \mathcal{H}, \|\boldsymbol{\theta}\|_{\mathcal{H}} \leq B\}$ has Rademacher complexity at most $O(B\sqrt{\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}} K(\boldsymbol{x}, \boldsymbol{x})}/\sqrt{n})$, where $\ell_{\boldsymbol{\theta}}(\boldsymbol{x}, y) := \log(1 + \exp(\boldsymbol{\theta}(\boldsymbol{x})))$. This, together with Theorem 30 and Markov inequality completes the proof of Theorem 43. ■

As usual, those two lemmas together give the concrete regret bound for the kernel logistic regression

Proof [Proof of Theorem 15] We will invoke Theorem 51 and Theorem 53 with $\varepsilon' := \varepsilon/3$

We can decompose

$$\mathcal{L}_{\hat\theta}(\mathcal{D}_{\theta^*}) - \mathcal{L}_{\theta^*}(\mathcal{D}_{\theta^*}) = (\mathcal{L}_{\hat\theta}(\mathcal{D}_{\theta^*}) - \widehat{\mathcal{L}}_{\hat\theta}(S) + (\widehat{\mathcal{L}}_{\hat\theta}(S) - \widehat{\mathcal{L}}_{\theta^*}(S)) + (\widehat{\mathcal{L}}_{\hat\theta}(S) - \mathcal{L}_{\theta^*}(\mathcal{D}_{\theta^*})).$$

By Theorem 53, the first and the last summand here are each at most ε'. On the other hand, by Theorem 51, we have $\widehat{\mathcal{L}}_{\hat\theta}(S) - \widehat{\mathcal{L}}_{\theta^*}(S) \le \varepsilon'$, leading to

$$\mathcal{L}_{\hat\theta}(\mathcal{D}_{\theta^*}) - \mathcal{L}_{\theta^*}(\mathcal{D}_{\theta^*}) \le 3\varepsilon' \le \varepsilon.$$

■

E.2. Proof of Theorem 16

Proof

With Theorem 15, this proof is identical to the proof of Theorem 13.

Consider a distribution $\mathcal{D}$ given by sampling $y \in \{\pm 1\}$, and then x from p_{tr} if $y = 1$ and x from p_{te} if $y = -1$.

We can apply the kernel logistic regression to a random sample from $\mathcal{D}$ — according to Theorem 16, we will be able to find a function $\widehat{\theta} \in \mathcal{H}$ with bounded regret with respect to distribution $\mathcal{D}$. According to Theorem 40, this $\widehat{\theta}$ is of form $\ln(\widehat{p_{\mathrm{te}}}/\widehat{p_{\mathrm{tr}}})$ for some $\widehat{p_{\mathrm{tr}}}$ and $\widehat{p_{\mathrm{te}}}$ both $(\varepsilon')^2$-close in KL-divergence to p_{tr} and p_{te} respectively. Using now Pinsker we get the TV-distance bounds on $d_{TV}(\widehat{p_{\mathrm{tr}}}, p_{\mathrm{tr}}) \le \varepsilon'$ and $d_{TV}(\widehat{p_{\mathrm{te}}}, p_{\mathrm{te}}) \le \varepsilon'$. The bound $R_2(p_{\mathrm{tr}}||p_{\mathrm{te}}) \le B_2$, together with Markov inequality implies that $\Pr_{\boldsymbol{x}\sim p_{\mathrm{te}}}[p_{\mathrm{te}}(\boldsymbol{x})/p_{\mathrm{tr}}(\boldsymbol{x}) > B] \le \varepsilon$ for $B = B_2/\varepsilon$.

We can now apply Theorem 10 with this B to deduce that we can use the truncated version of $\exp(\theta(x))$ as importance weights to solve the covariate shift problem; to this end we need $\varepsilon' = \varepsilon/B = \varepsilon^2/B_2$; so the desired regret error in Theorem 15 is of order ε^4/B_2^2, and according to this lemma we can get it with $O(DB_1^2B_2^4/\varepsilon^8)$ samples. ■

Appendix F. Discussing KMM guarantees

Description of the Kernel Mean Matching algorithm The KMM algorithm in the realizable scenario assumes that the unknown function $\boldsymbol{f}$ is bounded in some RKHS $\mathcal{H}$, $\|\boldsymbol{f}\|_{\mathcal{H}} \le M$. If the kernel corresponding to $\mathcal{H}$ is also bounded, say $K(\boldsymbol{x}, \boldsymbol{x}) \le 1$ for all $\boldsymbol{x}$.

In this case, with enough samples we can guarantee that $\|n_{tr}^{-1}\sum_i \Phi_{\boldsymbol{x}_i} - \mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{tr}}} \Phi_{\boldsymbol{x}}\|_{\mathcal{H}} \le \varepsilon/M$ (and similarly for $\tilde{\boldsymbol{x}} \sim p_{\mathrm{te}}$). Then by solving a linear program, we can find some weights $\hat\beta_i$ s.t. the empirical averages of the reproducing kernel map are close after reweighting, i.e., $\|\sum_i \hat\beta_i \Phi_{\boldsymbol{x}_i}/n_{tr} - \sum_i \Phi_{\tilde{\boldsymbol{x}}_i}/n_{te}\|_{\mathcal{H}} \le \varepsilon/M$ (these weights exist, because in particular $\beta_i = p_{\mathrm{te}}(\boldsymbol{x}_i)/p_{\mathrm{tr}}(\boldsymbol{x}_i)$ is a feasible

solution). Now it is easy to show that $\hat{\beta}$ can be used as importance weights to solve the covariate-shifted mean estimation:

$$\begin{aligned}\sum_i \frac{1}{n}\hat{\beta}_i \boldsymbol{f}(\boldsymbol{x}_i) &= \langle \sum_i \frac{1}{n}\hat{\beta}_i \Phi_{\boldsymbol{x}_i}, \boldsymbol{f}\rangle_{\mathcal{H}} \\ &= \langle \mathbb{E}_{\tilde{\boldsymbol{x}}\sim p_{\mathrm{te}}} \Phi_{\tilde{\boldsymbol{x}}}, \boldsymbol{f}\rangle_{\mathcal{H}} + \langle \Delta, \boldsymbol{f}\rangle_{\mathcal{H}} \\ &= \mathbb{E}_{\tilde{\boldsymbol{x}}\sim p_{\mathrm{te}}} \boldsymbol{f}(\boldsymbol{x}) \pm \|\Delta\|_{\mathcal{H}}\|\boldsymbol{f}\|_{\mathcal{H}},\end{aligned}$$

where $\Delta := \sum_i n^{-1}\hat{\beta}_i \Phi_{\boldsymbol{x}_i} - \mathbb{E}_{\tilde{\boldsymbol{x}}\sim p_{\mathrm{te}}} \Phi_{\tilde{\boldsymbol{x}}}$ satisfies $\|\Delta\|_{\mathcal{H}} \leq \varepsilon/M$.

As discussed above, collecting $n_{tr} \gtrsim (MB/\varepsilon)^2$ samples from p_{tr} and $n_{te} \gtrsim (M/\varepsilon)^2$ samples from p_{te} is enough to guarantee the corresponding closeness of three relevant empirical averages to their population averages.

Analysis of the plug-in method Alternative, naive approach, to solve the covariate shift mean estimation problem in the same setup as above, is to find $\hat{\boldsymbol{f}}$ with a sufficiently small distance L_1 from $\boldsymbol{f}$ with respect to p_{tr}, and use it as a proxy for $\boldsymbol{f}$ when computing the averages with respect to p_{te}.

Lemma 54 *When $\boldsymbol{f} \in \mathcal{H}$ is bounded by $\|\boldsymbol{f}\| \leq M$, using $O((M/\varepsilon')^2)$ samples $(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x}))$ for $\boldsymbol{x} \sim p_{\mathrm{tr}}$ we can find a function $\hat{\boldsymbol{f}}$ s.t.*

$$\mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{tr}}} |\boldsymbol{f}(\boldsymbol{x}) - \hat{\boldsymbol{f}}(\boldsymbol{x})| \leq \varepsilon'.$$

Proof Consider the family of function $\mathcal{F}' := \{\boldsymbol{x} \mapsto |\boldsymbol{h}(x)|\colon \boldsymbol{h} \in \mathcal{H}, \|\boldsymbol{h}\| \leq 2M\}$. Since $\boldsymbol{x} \mapsto |\boldsymbol{x}|$ is 1-Lipschitz, and we have a Rademacher complexity bound for a radius M-ball in the kernel space (Theorem 52), the Rademacher complexity of $\mathcal{R}_n(\mathcal{F}') \leq O(M/\sqrt{n})$.

Using Theorem 30 we obtain a generalization bound: after collecting $n = O(M/\varepsilon')^2$ samples, with probability $9/10$ we get a uniform bound on the generalization error for every $\boldsymbol{h} \in \mathcal{H}$ with $\|\boldsymbol{h}\| \leq 2$.

$$\sup_{\|\boldsymbol{h}\|\leq 2M} \frac{1}{n}\sum_i |\boldsymbol{h}(\boldsymbol{x}_i)| = \mathbb{E}_{\boldsymbol{x}}|\boldsymbol{h}(\boldsymbol{x})| \pm \varepsilon'/2.$$

In particular, taking $\boldsymbol{h}$ of form $\boldsymbol{f} - \boldsymbol{f}'$ for some $\|\boldsymbol{f}'\|_{\mathcal{H}} \leq M$, we get

$$\sup_{\|\boldsymbol{f}'\|\leq M} \frac{1}{n}\sum_i |\boldsymbol{f}(\boldsymbol{x}_i) - \boldsymbol{f}'(\boldsymbol{x}_i)| = \mathbb{E}_{\boldsymbol{x}}|\boldsymbol{f}'(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x})| \pm \varepsilon'/2.$$

It is enough therefore to approximately minimize the empirical loss: given samples $(\boldsymbol{x}_i, \boldsymbol{f}(\boldsymbol{x}_i))$, we need to find

$$\min_{\|\boldsymbol{f}'\|_{\mathcal{H}}\leq M} \frac{1}{n}\sum_i |\boldsymbol{f}'(\boldsymbol{x}_i) - \boldsymbol{f}(\boldsymbol{x}_i)|.$$

By the Representer Theorem (Theorem 14), this minimum is realized by some $\boldsymbol{f}'$ of form $\sum_i \Phi_{\boldsymbol{x}_i}\lambda_i$, and the problem becomes just a minimization

$$\min_{\lambda\in\mathbb{R}^n} \|K\lambda - y\|_1,$$

where K is a kernel matrix $K_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$, and y is a vector of observations $y_i = \boldsymbol{f}(\boldsymbol{x}_i)$. This minimization is just a linear program and can be solved efficiently. ■

Note that if we now find $\hat{\boldsymbol{f}}$ with L_1 error $\varepsilon' := \varepsilon/B$ with respect to p_{tr}, since the density ratio $p_{\mathrm{te}}/p_{\mathrm{tr}}$ is bounded by B we can also bound the L_1 error with respect to p_{te}. This clearly implies closeness of the expectations

$$\mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{te}}} \boldsymbol{f}(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{te}}} \hat{\boldsymbol{f}}(\boldsymbol{x}) \leq \mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{te}}} |\boldsymbol{f}(\boldsymbol{x}) - \hat{\boldsymbol{f}}(\boldsymbol{x})| \leq B\, \mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{tr}}} |\boldsymbol{f}(\boldsymbol{x}) - \hat{\boldsymbol{f}}(\boldsymbol{x})| \leq \varepsilon.$$

We can now use $(M/\varepsilon)^2$ samples from p_{te} to estimate $\mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{te}}} \hat{\boldsymbol{f}}(\boldsymbol{x})$ by a sample average. This leads to the final sample complexity $n_{tr} \gtrsim (MB/\varepsilon)^2$ samples from p_{tr} and $n_{te} \gtrsim (M/\varepsilon)^2$ samples from p_{te}, matching the guarantees of the KMM method.