Submitted 6/24; Revised 11/24; Published 1/25

Nearest Neighbor Sampling for Covariate Shift Adaptation

François Portier FRANCOIS.PORTIER@ENSAI.FR
Department of Statistics,
Univ Rennes, Ensai, CNRS, CREST—UMR 9194, F-35000 Rennes, France

Lionel Truquet LIONEL.TRUQUET@ENSAI.FR
Department of Statistics,
Univ Rennes, Ensai, CNRS, CREST—UMR 9194, F-35000 Rennes, France

Ikko Yamane IKKO.YAMANE@ENSAI.FR
Department of Computer Science,
Univ Rennes, Ensai, CNRS, CREST—UMR 9194, F-35000 Rennes, France

Editor: Aryeh Kontorovich

Abstract

Many existing covariate shift adaptation methods estimate sample weights to mitigate the gap between the source and the target distribution. However, estimating the optimal weights typically involves computationally expensive matrix inversion and hyper-parameter tuning. In this paper, we propose a new covariate shift adaptation method which avoids estimating the weights. The basic idea is to directly work on unlabeled target data, labeled according to the k-nearest neighbors in the source dataset. Our analysis reveals that setting $k = 1$ is an optimal choice. This property eliminates the necessity of tuning the hyper-parameter k and leads to a running time quasi-linear in the sample size. Our results include sharp rates of convergence for our estimator, with a tight control of the mean square error and explicit constants. In particular, the variance of our estimator has the same rate of convergence as for standard parametric estimation despite their non-parametric nature. The proposed estimator shares similarities with some matching-based treatment effect estimators used, e.g., in biostatistics, econometrics, and epidemiology. Our experiments show that it achieves drastic reduction in the running time with remarkable accuracy.

Keywords: nearest neighbor, resampling, covariate shift, scalability, matching

1. Introduction

Traditional machine learning methods assume that the source data distribution P and the target data distribution Q are identical. However, this assumption can be violated in practice when there is a *distribution shift* (Chen et al., 2022) between them. Various types of shift have been studied in the literature, and one of the most common scenarios is *covariate shift* (Shimodaira, 2000) in which there is a shift in the input distribution: $P_X \neq Q_X$ while the conditional distributions of the output variable given the input variable are the same: $P_{Y\,|\,X} = Q_{Y\,|\,X}$, where X is the input and Y is the output variable. The goal of *covariate shift adaptation* is to adapt a supervised learning algorithm to the target distribution using labeled source data and unlabeled target data.

A standard approach to covariate shift is weighting source examples (Shimodaira, 2000), and many studies focused on improving the weights (Huang et al., 2006; Gretton et al., 2008; Yamada et al., 2013; Kanamori et al., 2009; Sugiyama et al., 2007, 2008; Loog, 2012; Aminian et al., 2022) in the same line of research. We refer the reader to Section 6 for more details of related work. Since we rarely know the model for how the input distributions can be shifted a priori, non-parametric methods are particularly useful for covariate shift adaptation. Some of the existing methods allow one to use non-parametric models through kernels. However, such kernel-based methods take at least quadratic times in computing kernel matrices. Some methods further need to solve linear systems and take cubic times in the sample size unless one resorts to approximations (Williams and Seeger, 2000; Le et al., 2013; Rudi and Rosasco, 2017). Moreover, their performance is often sensitive to the choice of hyper-parameters of the kernel. Typically, one performs a grid search K-fold cross-validation for selecting the hyper-parameters, which amplifies the running time by about $K|\Gamma|$, where Γ is the set of candidates for the hyper-parameters. Moreover, the criterion for the hyper-parameter selection is not obvious either because we do not have access to the labels for the target data. One can use weighted validation scores using the labeled source data with importance sampling, but it is not straightforward to choose what weights to be used for the cross-validation when we are choosing weights.

In this paper, we propose a non-parametric covariate shift adaptation method that is scalable and has no hyper-parameter. Our idea is to generate synthetic labels for unlabeled target data using a non-parametric conditional sampler constructed from source data. Under the assumption of covariate shift, the target data attached with the generated labels behave like labeled target data. This sampling technique allows any supervised learning method to be simply applied to the generated data to produce a model already adapted to the target distribution.

While the proposed approach is quite general and can be employed with various sampling methods for the synthetic labeling part, our main result shows that using k-nearest neighbor (k-NN) conditional sampling achieves an error of order $(k/n)^{1/d} + 1/\sqrt{n} + 1/\sqrt{m}$ for estimating an expectation on the target domain, where d is the data dimensionality, and n and m are the source and the target sample size, respectively. Importantly, our error bounds suggest that $k = 1$ is the most favorable. This property, which is revealed by a precise scaling of the variance term in $1/\sqrt{n}$, is a non-trivial and remarkable fact, given the usual $1/\sqrt{k}$-rate of the variance term associated to most k-NN estimators. In particular, it contrasts with well-known applications of k-NN such as density estimation (Dasgupta and Kpotufe, 2014), classification (Gadat et al., 2016; Cannings et al., 2020), regression (Devroye et al., 1994; Jiang, 2019) or conditional distribution function estimation (Portier, 2025), in which case choosing $k = 1$ would not provide a consistent estimation and letting k grow in a polynomial rate in the sample size is recommended to achieve a good balance between bias and variance. Textbooks dealing with the k-NN algorithm include Györfi et al. (2006); Devroye et al. (2013b); Biau and Devroye (2015).

This important difference in the rate of convergence, leading to the use of a single nearest neighbor ($k = 1$), has also been noticed in several estimation problems where a 1-NN estimator is used as a necessary step to estimate a parameter of interest. It includes entropy estimation (Berrett et al., 2019), integral approximation with control variates (Leluc et al., 2023; Blanchet et al., 2023) and residual variance estimation (Devroye et al., 2018, 2013a). For instance, in Devroye et al. (2018), using a 1-NN regression estimator allows achieving residual variance estimation with a variance term scaling as $1/\sqrt{n}$, as we obtain here. In the context of covariate shift, Loog (2012) proposed a 1-NN method to re-weight the source sample, in the spirit of the matching approach discussed below, and reported some encouraging empirical results.

The covariate shift problem is closely related to a well-known matching approach studied in the context of treatment effect estimation. In particular, k-NN estimators have been used to estimate the so-called average treatment effect to tackle missingness of potential outcomes. See, e.g., Rosenbaum (1995); Abadie and Imbens (2006), for an error bound obtained in the problem. More recently, Sharpnack (2022) proved consistency results on the 1-NN matching method under much milder assumptions. In Section 6, we discuss the main differences between the two problems and why their result is not generally applicable to ours.

In addition to being optimal with respect to the estimation error, setting $k = 1$ circumvent the cumbersome hyper-parameter tuning while providing computational efficiency at the same time. Our 1-NN-based algorithm takes only a quasi-linear time $\mathcal{O}((n + m) \log n)$ on average using the optimized k-d tree (Bentley, 1975; Friedman et al., 1977). Indeed, our experiments show that the proposed method terminates faster than previous methods, by large margins. Note that the problem of getting a computationally efficient method for covariate shift adaptation, in particular for scalability to large data sets, is a recurrent problem in the existing literature. In fact, many existing methods resorted to implementation heuristics such as using a fixed number of kernel centers for reducing the computational burden at the cost of statistical guarantee (Kanamori et al., 2009; Sugiyama et al., 2007, 2008; Yamada et al., 2013).

Note also that computational advantages of using $k = 1$ have been exploited in a recent line of work on classification, see for instance Kontorovich and Weiss (2015); Hanneke et al. (2021); Györfi and Weiss (2021), where constructing special Voronoï partitioning admits consistency results as well as finite-sample error bounds while keeping small complexities of training and prediction in time and space.

Our method simulates the missing labels of the target sample, which in turn can be used for a variety of downstream supervised learning tasks. Even though the main focus of this paper is the estimation of expectations in the target domain, for illustrating the usefulness of our method in a typical machine learning downstream task, we also demonstrate consistency properties of parametric M-estimators in the target domain (Section 5). This is particularly useful in regression with a mispecified parametric model. In this case, the minimizer projected onto the model depends on the covariate distribution even though the regression function stays the same.

In summary, the key contributions of this paper are the following. (i) Our method is non-parametric. It does not introduce a model in covariate shift adaptation so that it will have a minimum impact on the model trained for the downstream task. (ii) Our method is fast. Adaptation only takes a quasi-linear time. (iii) There is no hyper-parameter to be tuned. (iv) The proposed method only incurs an error of order $(k/n)^{1/d} + 1/\sqrt{n} + 1/\sqrt{m}$ for estimating an expectation on the target domain.

The outline is as follows. In Section 2, the problem of covariate shift adaptation is formally introduced along with the mathematical notation. Section 3 contains the description of the method. Section 4 is dedicated to the main theoretical results while Section 5 investigates the empirical risk minimization problem in presence of covariate shift adaptation. Section 6 provides a description of several alternative approaches to a similar type of problem as well as some points of comparison with our proposal. In Section 7, several avenues for further research are discussed and finally, the numerical experiments are provided in Section 8.

2. Problem setup

Let $\mathcal{X} = \mathbb{R}^d$ for some $d \geq 1$ and $\mathcal{Y}$ be a measurable space. Let $P \equiv P_{X,Y}$ and $Q \equiv Q_{X,Y}$ be probability distributions defined on $\mathcal{X} \times \mathcal{Y}$. Throughout the paper, we assume that P and Q admit the decomposition

$$P = P_{Y \mid X} P_X \quad \text{and} \quad Q = Q_{Y \mid X} Q_X,$$

where $P_{Y \mid X=x}$ and $Q_{Y \mid X=x}$ are probability distributions defined on $\mathcal{Y}$ for each $x \in \mathcal{X}$.[1] Here, P_X and Q_X are the marginal distributions of X when (X, Y) is distributed with P and Q, respectively. We shall simply call $P_{Y \mid X}$ (or $Q_{Y \mid X}$) the *conditional distribution* of Y given X in the source domain (or the target domain).

Definition 1 (Source sample, source distribution) *For each integer $n \geq 1$, let $(X_i, Y_i)_{i=1}^n$ be a collection of independent and identically distributed random variables with P. We refer to $(X_i, Y_i)_{i=1}^n$ as the (labeled)* source sample *and P as the* source distribution.

Definition 2 (Target sample, target distribution) *For each integer $m \geq 1$, let $(X_i^*)_{i=1}^m$ be a collection of independent and identically distributed random variables with Q_X. We refer to $(X_i^*)_{i=1}^m$ as the (unlabeled)* target sample *and Q as the* target distribution.

Definition 3 (Covariate shift) Covariate shift *is a situation in which the source and the target distribution have different marginal distributions for X while sharing a common conditional distribution:*

(C1) $P_{Y \mid X} = Q_{Y \mid X}$, P_X- and Q_X-a.s., but $P_X \neq Q_X$.

This paper focuses on the following simple but versatile estimation problem under covariate shift.

Definition 4 (Mean estimation under covariate shift) *For each pair of integers $n \geq 1$, $m \geq 1$, and a known integrable function $h \colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, the goal of* mean estimation under covariate shift *is to estimate the mean of h under the target distribution,*

$$Q(h) \equiv \int h(x, y) Q(dx, dy),$$

given access to the source sample $(X_i, Y_i)_{i=1}^n \sim P$ and the target sample $(X_i^*)_{i=1}^m \sim Q_X$ under Assumption (C1).

For instance, when $h(x, y) = \ell(f(x), y)$ for a loss function $\ell \colon \mathcal{Y}^2 \to \mathbb{R}$ and a hypothesis function $f \colon \mathcal{X} \to \mathcal{Y}$, estimation of $Q(h)$ becomes risk estimation, which is the central subtask in *empirical risk minimization.*

[1] More formally, we denote by $P_{Y \mid X=(\cdot)}(dy)$ a *regular conditional measure* (Bogachev and Ruas, 2007, Definition 10.4.1) such that the marginal distribution of Y can be expressed as $P_Y(dy) = \int P_{Y \mid X=x}(dy) P_X(dx)$. We also use $P_{Y \mid X}(dy \mid \cdot)$ for $P_{Y \mid X=(\cdot)}(dy)$. The same goes for Q.

3. Proposed method

The basic idea of our proposed method is to use the source sample for learning to *label the target data*. Specifically, using the source sample $(X_i, Y_i)_{i=1}^n$, we will construct a stochastic labeling function $\hat{\mathsf{S}}$ that inputs any target data point X_i^* and outputs a random label $Y_{n,i}^* \in \mathcal{Y}$. (The subscript n of $Y_{n,i}^*$ is for explicitly denoting the dependence on the source sample.) Once we succeed in generating labels for target data that behave like true target labels, we will be able to perform any supervised learning method directly on the target sample for the downstream task. For our mean estimation problem, we can simply average the output h evaluated at the target data with the generated labels.

When do the generated labels behave like the true target labels? Let $\hat{P}_{Y \mid X_i^*}$ denote the probability distribution of an output $Y_{n,i}^*$ of $\hat{\mathsf{S}}$ for input X_i^*. We wish to obtain $\hat{\mathsf{S}}$ such that the probability distribution $\hat{Q} \equiv \hat{P}_{Y \mid X} Q_X$ of $(X_i^*, Y_{n,i}^*)$ will give a good estimate $\hat{Q}(h)$ of $Q(h) = (Q_{Y \mid X} Q_X)(h)$. In this sense, our task boils down to designing a good conditional sampler $\hat{\mathsf{S}}$ that mimics sampling from $P_{Y|X}$. Algorithm 1 describes an outline of this general framework. In the analysis and design choices, it will be important to note that accurately estimating the integral $Q(h)$ does not necessarily mean that $P_{Y \mid X}$ should be accurately estimated.

Algorithm 1 Conditional Sampling Adaptation

```
**Input:** Conditional sampler $\hat{\mathsf{S}}$ and target sample $(X_j^*)_{j=1}^m$.
$Y_{n,j}^* \leftarrow \hat{\mathsf{S}}(X_j^*)$ for each $j \in \{1, \ldots, m\}$. // Generate a label conditioned on $X_j^*$.
**return** $m^{-1} \sum_{j=1}^m h(X_j^*, Y_{n,j}^*)$.
```

In this paper, we propose a method using a non-parametric conditional sampler $\hat{\mathsf{S}}$ based on the k-Nearest Neighbor (k-NN) method, which randomly picks one of the k-nearest neighbors of the input X_j^* among the source instances $(X_i)_{i=1}^n$ and output the corresponding label (Algorithm 2). We refer to this method as *k-NN-based Conditional Sampling Adaptation* (*k-NN-CSA*).

Algorithm 2 k-Nearest Neighbor Conditional Sampler

```
**Input:** Source sample $(X_i, Y_i)_{i=1}^n$ and target input $X_j^*$.
$(i_1, \ldots, i_k) \leftarrow$ the indices of the $k$-nearest neighbors of $X_j^*$ among $(X_i)_{i=1}^n$.
Pick $i^* \in \{i_1, \ldots, i_k\}$ uniformly at random.
**return** $Y_{n,j}^* := Y_{i^*}^*$.
```

Computing time Recent advances for nearest neighbor search rely on tree-search to reduce the computing time. The seminal paper by Bentley (1975) introduced the k-d tree method. Building such a tree requires $\mathcal{O}(n \log n)$ and once the tree is available, search for the nearest neighbor of a given point can be done in $\mathcal{O}(\log n)$ time (Friedman et al., 1977). As a consequence, the time complexity of k-NN-CSA is $\mathcal{O}(n \log n + km \log n)$.

4. Theoretical analysis

We now present the theory behind our approach in a didactic way by introducing a key decomposition first and then studying separately each of the terms involved: the *bootstrap sampling error* and the *nonparametric error* associated to k-NN. We will see that the k-NN-CSA with $k = 1$ (1-NN-CSA for short) achieves the best theoretical performance among those with other k's.

4.1 The key decomposition

For the analysis of k-NN-CSA, recall that $\hat{Q} = \hat{P}_{Y|X} Q_X$ is an estimate of the target distribution $Q = P_{Y|X} Q_X$ that depends on the source sample $(X_i, Y_i)_{i=1}^n$, whose probability distribution is P. We introduce the bootstrap sample as a collection of random variable generated according to $\hat{Q}$.

Definition 5 (Bootstrap sample) *For each $m \geq 1$ and $n \geq 1$, let $(X_i^*, Y_{n,i}^*)_{1 \leq i \leq m}$ be a collection of random variables identically distributed with $\hat{Q}$ and conditionally independent given $(X_i, Y_i)_{i=1}^n$.*

Let $h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a measurable function. The quantity of interest is

$$\hat{Q}^*(h) = m^{-1} \sum_{i=1}^{m} h(X_i^*, Y_{n,i}^*),$$

which is the CSA estimate of $Q(h) = \int h(x, y) Q(dx, dy)$ as introduced in Algorithm 1. The following decomposition is crucial in our analysis:

$$(\hat{Q}^* - Q)(h) = \underbrace{(\hat{Q}^* - \hat{Q})(h)}_{\textit{bootstrap sampling error}} + \underbrace{(\hat{Q} - Q)(h)}_{\textit{nonparametric error}} \tag{1}$$

$$\Big(= (\hat{Q}_X^* - Q_X)\hat{P}_{Y \mid X}(h) + Q_X(\hat{P}_{Y \mid X} - P_{Y \mid X})(h) \Big),$$

where $\hat{Q}_X^*(\cdot) \equiv \frac{1}{n} \sum_{i=1}^m \mathbb{1}_{X_i^* = (\cdot)}$ is the empirical measure defined with $(X_i^*)_{i=1}^m$. The first term is the error due to the use of $\hat{Q}_X^*$ in place of Q_X, which tends to zero as m grows. The second term represents the error due to the use of $\hat{P}_{Y \mid X}$ in place of $P_{Y \mid X}$. When using the k-nearest neighbor algorithm to obtain $\hat{P}_{Y \mid X}$, we show that this term is of order $(k/n)^{1/d} + 1/\sqrt{n}$, which differs from the standard non-parametric convergence rate in $(k/n)^{1/d} + 1/\sqrt{k}$ found in regression problems.

4.2 Bootstrap sampling error

First, we will show that the bootstrap sampling error, $(\hat{Q}^* - \hat{Q})(h)$ (the first term in our decomposition (1)), is of order $1/\sqrt{m}$. The analysis relies on martingale tools. Define $\mathcal{F}_n = \sigma((X_1, Y_1), \ldots, (X_n, Y_n))$. For each $1 \leq i \leq m$, we have

$$\mathbb{E}[h(X_i^*, Y_{n,i}^*) \mid \mathcal{F}_n] = \hat{Q}(h).$$

This property implies that $\sum_{i=1}^m \{h(X_i^*, Y_{n,i}^*) - \hat{Q}(h)\}$ is a martingale and therefore can be analyzed using the Lindeberg Central Limit Theorem (CLT) conditionally on the initial sample hence fixing the distribution $\hat{Q}$. The next property is reminiscent of certain results about the bootstrap method where sampling is done with the basic empirical measure; see e.g., Van der Vaart (2000). We need this type of results without specifying the measure $\hat{Q}$ so that we can incorporate a variety of sampling schemes such as $\hat{Q} = \hat{P}_{Y \mid X} Q_X$. The proof is given in Appendix B.1.

Proposition 1 *Suppose that $\hat{Q}$ satisfies the following strong law of large number: for each h such that $Q(h) < \infty$, we have $\lim_{n\to\infty} \hat{Q}(h) = Q(h)$ almost surely. Then, if $m := m_n \to \infty$ as $n \to \infty$, we have the following central limit theorem: for each function such that $Q(h^2) < \infty$, we have, conditionally to $\mathcal{F}_n$, almost surely,*

$$\sqrt{m}\{\hat{Q}^*(h) - \hat{Q}(h)\} \rightsquigarrow \mathcal{N}(0, V) \qquad \text{as } n \to \infty,$$

where $V = \lim_{n\to\infty}\{\hat{Q}(h^2) - \hat{Q}(h)^2\}$.

As a corollary of the previous results, we can already deduce that if m goes to ∞ and $\hat{Q}$ satisfies a strong law of large numbers, then $\hat{Q}^*(h)$ converges to $Q(h)$ provided that $Q(h^2)$ exists. This is a general consistency result that justifies the use of any resampling distribution $\hat{Q}$ that converges to Q. In practical situations, it is useful to know a finite-sample bound on the error. This is the purpose of the next proposition, in which we give a non-asymptotic control of the bootstrap sampling error. A proof is given in Appendix B.2.

Proposition 2 *Suppose that h is bounded by a constant $U_h > 0$. Let $\delta \in (0, 1)$. Then with probability greater than $1 - \delta$,*

$$\left|\hat{Q}^*(h) - \hat{Q}(h)\right| \leq \frac{U_h}{m}\log(2/\delta) + \sqrt{2\frac{\hat{v}_n}{m}\log(2/\delta)},$$

where $\hat{v}_n = \hat{Q}(h^2) - (\hat{Q}h)^2$.

Notes. A natural "averaging" alternative to the above "bootstrap sampling" estimator $\hat{Q}^*$ can also be investigated using the same tools. Instead of sampling $Y_{n,i}^*$ according to $\hat{P}_{Y|X}(dy|X_i^*)$, one might consider taking the expectation, leading to

$$\overline{Q}(h) = \frac{1}{m}\sum_{i=1}^m \int h(X_i^*, y)\hat{P}_{Y|X}(dy|X_i^*).$$

This estimate can be studied in a similar way as before and the two above results are still valid with small changes. In particular Proposition 2 holds true with smaller variance term as, by Jensen's inequality, $\mathbb{E}[\int h(X_i^*, y)\hat{P}_{Y|X}(dy|X_i^*)^2 \mid \mathcal{F}_n] \leq \hat{Q}(h^2)$. This alternative $\overline{Q}(h)$ requires more computing time (when measured in terms of the number of evaluations of h) and is less appealing for stochastic gradient descent algorithm or in semiparametric estimation problems, as discussed in Section 7. Estimators similar to $\overline{Q}(h)$ have been studied in average treatment effects literature (Rosenbaum, 1995; Abadie and Imbens, 2006) as well as in covariate shift or bias sampling problems (Loog, 2012; Sharpnack, 2022); see Section 6 for precise discussions.

4.3 Nonparametric error of the nearest neighbor estimator

In this section, we obtain several bounds on the nonparametric error $\hat{Q}(h)-Q(h)$, the second term in our decomposition (1), when $\hat{P}_{Y|X}$ is the k-NN estimator of the conditional measure $P_{Y|X}$. The obtained bounds will be then combined with the ones from previous section to provide a guarantees on the global estimation error $\hat{Q}^*(h) - Q(h)$.

Let $x \in \mathbb{R}^d$ and $\|\cdot\|$ be the Euclidean norm on $\mathbb{R}^d$. Denote the closed ball of radius $\tau \geq 0$ around x by $B(x,\tau) := \{z \in \mathbb{R}^d \mid \|x-z\| \leq \tau\}$. For $n \geq 1$ and $k \in \{1, \dots, n\}$, the k-nearest neighbor (k-NN for short) radius at x is denoted by $\hat{\tau}_{n,k,x}$ and defined as the smallest radius $\tau \geq 0$ such that the ball $B(x,\tau)$ contains at least k points from the collection $\{X_1, \dots, X_n\}$. That is,

$$\hat{\tau}_{n,k,x} := \inf \left\{ \tau \geq 0 \,:\, \sum_{i=1}^n 1_{B(x,\tau)}(X_i) \geq k \right\},$$

where $1_A(x)$ is 1 if $x \in A$ and 0 elsewhere. The k-NN estimate of $P_{Y\,|\,X}(dy \mid x)$ is given by

$$\hat{P}_{Y\,|\,X}(dy \mid x) = k^{-1} \sum_{i=1}^n 1_{\|X_i - x\| \leq \hat{\tau}_{n,k,x}} \delta_{Y_i}(dy),$$

where $\delta_y(\cdot)$ is the Dirac measure at $y \in \mathcal{Y}$ defined by $\delta_y(A) = 1_A(y)$ for any measurable set $A \subseteq \mathcal{Y}$. Consequently, the k-NN estimate of the integral $\int h(y,x) P_{Y\,|\,X}(dy \mid x)$ is then defined as

$$\int h(x,y) \hat{P}_{Y\,|\,X}(dy \mid x) = k^{-1} \sum_{i=1}^n 1_{\|X_i - x\| \leq \hat{\tau}_{n,k,x}} h(x, Y_i).$$

Our result on $\hat{Q}(h)-Q(h)$ is established under the following set of assumptions. Consider the case where source covariates X admit a density with respect to the Lebesgue measure. We will need in addition that the support S_X is well shaped and that the density is lower bounded. These are standard regularity conditions to obtain some upper bound on the k-NN radius.

(X1) The random variable X admits a density p_X with compact support $S_X \subset \mathbb{R}^d$.

(X2) There is $c > 0$ and $T > 0$ such that

$$\lambda_d(S_X \cap B(x,\tau)) \geq c \lambda_d(B(x,\tau)), \qquad \forall \tau \in (0,T],\, \forall x \in S_X,$$

where λ_d is the Lebesgue measure on $\mathbb{R}^d$.

(X3) There are $0 < b_X \leq U_X < +\infty$ such that $b_X \leq p_X(x) \leq U_X$, for all $x \in S_X$.

In addition, we will use the following assumption on the target measure Q_X.

(X4) The probability measure Q_X admits a density q_X with support included in S_X. Moreover, there exists a measurable mapping $C\colon S_X \to [0,\infty)$ such that $\int C(x) Q_X(dx) < \infty$ and that for any $\tau > 0$ and $x \in S_X$,

$$Q_X\left(B(x,\tau)\right) \leq C(x) \tau^d.$$

Assumption (X4) is satisfied in particular when the density q_X is upper bounded. In this case, one can simply take $C(x) = V_d \sup_{x \in S_X} q_X(x)$, where $V_d := \lambda_d(B(0,1))$ denotes the volume of the unit ball of $\mathbb{R}^d$. However, it is possible to consider unbounded densities q_X. The simple example $q_X(x) = (1-\alpha)x^{-\alpha}$ for $x \in S_X = (0,1)$ satisfies Assumption (X4) as soon as $\alpha \in (0, 1/2)$. See Appendix G for details. In particular, the density ratio q_X/p_X is not necessarily bounded on S_X.

Two additional assumptions will be needed to deal with the function h and the probability distribution of (Y, X).

(H1) There exists a measurable function $g_h \colon S_X \to \mathbb{R}$ such that $\int g_h^2(x) Q_X(dx) < \infty$ and for any x in S_X,

$$|\mathbb{E}[h(x,Y) \mid X = x] - \mathbb{E}[h(x,Y) \mid X = x+u]| \leq g_h(x)\|u\|.$$

(H2) There exists $\sigma_+^2 > 0$ such that $\sup_{x \in S_X} \mathrm{Var}(h(x,Y)|X) \leq \sigma_+^2$ a.s., where $\mathrm{Var}(h(x,Y)|X)$ is the conditional variance of $h(x,Y)$ given X.

We are in position to state our result on $\hat{Q}(h) - Q(h)$. It consists in an upper bound on the mean squared error with explicit constants with respect to the dimension d and the parameters of the problem. Additionally, we provide a lower bound for the variance term which implies a standard parametric rate of convergence (see the notes below the statement). The proof is given in Appendix C.1. Let $\lfloor x \rfloor$ be the integer part of a real number x and define the gamma function Γ by $\Gamma(x) := \int_0^\infty u^{x-1}\exp(-u)du$ for $x > 0$.

Proposition 3 *Suppose that Assumptions (X1), (X2), (X3), (X4), (H1), and (H2) are fulfilled. We have*

$$\hat{Q}h - Qh = S_h + B_h,$$

where B_h is a bias term (defined in the proof) that satisfies, for any $n \geq 1$,

$$\mathbb{E}|B_h|^2 \leq \frac{2\Gamma\left(1 + \lfloor 2/d \rfloor\right)}{M_{1,d}^{2/d}} \int g_h^2(x) Q_X(dx) \cdot \frac{k^{2/d}}{n^{2/d}},$$

and S_h is a variance term (defined in the proof) that satisfies, for any $n \geq 2$,

$$\frac{\sigma_-^2 M_{1,d}^2}{4M_{2,d}^2} n^{-1} \leq \min_{1 \leq k \leq n} \mathbb{E}\left[S_h^2\right] \leq \max_{1 \leq k \leq n} \mathbb{E}\left[S_h^2\right] \leq \frac{2^{d+3}\sigma_+^2 M_{2,d} \int C(x) Q_X(dx)}{M_{1,d}^2} n^{-1},$$

with $M_{1,d} = cb_X V_d$ and $M_{2,d} = U_X V_d$. For the lower bound to be true, it is assumed that the mapping h does not depend on x, i.e., $h(x,y) = h(y)$ and $\sigma_-^2 = \inf_{x \in S_X} \mathrm{Var}\left(h(Y)|X = x\right)$.

Notes. (i) The two terms B_h and S_h correspond respectively to the bias term and the variance term. The upper bound obtained for the bias term is usual in k-NN regression analysis. However, the upper and lower bound on the variance are particular to our framework as they show that the variance behaves as in usual parametric estimation. A similar result was obtained for residual variance estimation in Devroye et al. (2018). Consequently,

our rates of convergence are sharper than the optimal rate of convergence $n^{-\frac{1}{2+d}}$ for nonparametric estimation of Lipschitz functions. This can be explained by the fact that several k-NN estimators are averaged to estimate Qh, which is a standard expectation and not a conditional expectation.
(ii) Since the rate of convergence of the variance term S_h does not depend on k, k might be chosen according to the upper bound on the bias term, which gives $k = 1$. One can deduce the following convergence rates, depending on the dimension. For $d = 1$, we get the rate $n^{-1/2}$. For $d = 2$, the contributions of both terms, B_h and S_h, coincide and we get the rate $n^{-1/2}$. For $d \geq 3$, the rate is $n^{-1/d}$.

For the global mean squared error, which incorporates the bootstrap sampling error $\hat{Q}^*(h) - \hat{Q}(h)$ as well as the k-NN nonparametric error $\hat{Q}(h) - Q(h)$, we give the following result in the optimal case $k = 1$. The proof can be found in Appendix D.

Theorem 1 *Suppose that Assumptions (X1), (X2), (X3), (X4) and (H1) hold true with* $\sup_{x\in S_X} \mathbb{E}\left[h^2(x, Y)|X\right]$ *bounded. If* $k = 1$*, there then exists* $C > 0$ *only depending on the distribution of* (X, Y)*,* X^**, and on* h *such that*

$$\mathbb{E}\left|\hat{Q}^*(h) - Q(h)\right|^2 \leq C\left\{\frac{1}{m} + \frac{1}{n^{2/d}} + \frac{1}{n}\right\}.$$

We next give a non-asymptotic control of $\hat{Q}h - Qh$ when h is a bounded function using Bernstein's concentration inequality. This bound affords a complement with respect to the bound for the MSE. However, for technical reasons, this high-probability bound requires that k grows at least logarithmically with respect to n, in contrast to Proposition 3. In our numerical experiments, we will also include the case $k = \log n$ for comparison. Additionally, we impose an upper bound for the target density q_X (for clarity reason, we assume that q_X is bounded by U_X, as for p_X). The proof of the next result is given in Appendix C.2.

Proposition 4 *Suppose that Assumptions (X1), (X2), (X3), (H1), and (H2) are fulfilled and assume that* Q_X *has a density* q_X *bounded by* U_X *with support included in* S_X*. Suppose that there exists a constant* $C > 0$ *such that* $C \log n \leq k \leq n/2$ *and that* h *is bounded by* U_h*. Let* $\delta \in (0, 1/3)$*. With probability greater than* $1 - 3\delta$*, we have*

$$\left|\hat{Q}h - Qh\right| \leq L_0\left(\frac{k}{n}\right)^{1/d} + \frac{2L_2}{n}\log(2/\delta) + \sqrt{\frac{2L_1\sigma^2}{n}\log(2/\delta)},$$

where

$$L_0 = \left(\frac{2}{cb_xV_d}\right)^{1/d}\int g_h(x)Q_X(dx), \quad L_1 = \frac{4\sigma_+^2U_X^2}{b_X^2c^2}, \quad L_2 = \frac{4U_hU_X}{3b_Xc}.$$

Notes. (i) The proof needs a bound on $\sup_{x\in S_X}\hat{\tau}_{n,k,x}$ which is given in (Portier, 2025, Lemma 3). For this we need that k grows logarithmically with respect to n as stated in the assumptions.
(ii) The sum of the two last terms in the upper-bound, which corresponds to the variance of our estimator, is of order $1/\sqrt{n}$ and the conditional variance of h appears as a multiplicative factor. Combining Proposition 2 and Proposition 4, we finally obtain that $\hat{Q}^*(h) - Q(h)$ is of order $1/\sqrt{m} + 1/\sqrt{n} + (k/n)^{1/d}$ (up to log factors).

(iii) In Corollary 1 in Appendix E, the upper bound given in Proposition 2 is refined using Proposition 4 that allows to (roughly speaking) replace the empirical variance by the true variance.

5. Applications to empirical risk minimization

In this section, we illustrate our results with some applications to empirical risk minimization. This is of particular interest in our context as the optimal linear model for the source distribution might be different from the ideal linear model for the target. In such a case, using covariate shift adaptation is necessary for consistency as the source minimizer will be away from the target minimizer.

5.1 Mathematical background

Suppose that $\mathcal{R}^*_{m,n}(\theta) = \frac{1}{m}\sum_{i=1}^m m_\theta\left(Y^*_{n,i}, X^*_i\right)$, where for each $\theta \in \Theta \subset \mathbb{R}^p$, m_θ is a measurable function from $\mathcal{Y} \times \mathcal{X}$ to $\mathbb{R}$. Set

$$\hat{\theta}^* \in \arg\min_{\theta\in\Theta} \mathcal{R}^*_{m,n}(\theta).$$

Similarly, we define

$$\theta^* \in \arg\min_{\theta\in\Theta} \mathcal{R}^*\left(\theta\right)$$

with $\mathcal{R}^*(\theta) = \mathbb{E}m_\theta\left(Y^*, X^*\right)$ and (Y^*, X^*) is a copy of (Y^*_i, X^*_i). Note that the expected value is taken for the unobserved label Y^*_i and not the generated label $Y^*_{n,i}$. We assume here that for a reference measure μ on $\mathcal{Y}$, there exists for each $x \in \mathcal{X}$ a conditional density $p(\cdot|x)$ such that $(x, y) \mapsto p(y|x)$ is jointly measurable and for any Borel set B,

$$\mathbb{P}\left(Y \in B | X = x\right) = \int_B p(y|x)\mu(dy).$$

One can then include the cases of classification ($\mu = \delta_0 + \delta_1$), counts ($\mu$ is the counting measure on the set of nonnegative integers), or regression (μ is the Lebesgue measure on $\mathbb{R}$).

5.2 Consistency of general empirical risk minimizers

We will use the following assumptions. We denote by $|\cdot|$ an arbitrary norm on $\mathbb{R}^p$.

(A1) There exist a measurable function $h : \mathcal{Y} \to \mathbb{R}_+ := (0, +\infty)$ and $\eta : \mathbb{R}_+ \to \mathbb{R}_+$ such that for any $y \in \mathcal{Y}$

$$\sup_{x\in\mathcal{X}} \sup_{\theta\in\Theta} |m_\theta(y, x)| \leq h(y),$$
$$\sup_{x\in\mathcal{X}} \sup_{|\theta-\theta'|\leq\delta} |m_\theta(y, x) - m_{\theta'}(y, x)| \leq h(y)\eta(\delta),$$

and such that $\mathbb{E}\left[h(Y)^2|X\right]$ is a bounded random variable and $\lim_{\delta\to 0}\eta(\delta) = 0$.

(A2) There exists a measurable function $g_h : \mathcal{X} \to \mathbb{R}_+$ such that $\int g_h(x)^2 Q_X(dx) < \infty$ and

$$\int h(y)\left|p(y|x+u) - p(y|x)\right|\mu(dy) \leq g_h(x)\|u\|, \quad (x, x+u) \in S_X^2.$$

The above assumptions are satisfied, for instance, in the logistic regression framework with compact covariates. In this case, h is a constant function and $\eta(\delta) = \delta$. Note also that $p(1|x)$ could be different form $\left(1 + \exp\left(-x^T\theta\right)\right)^{-1}$ as soon as $x \mapsto p(1|x) \in (0, 1)$ is Lipschitz on S_X.

In what follows, an assertion of the form $X_{m,n} = o_{\mathbb{P}}(1)$ as $m, n \to \infty$ means that for any $\epsilon, \zeta > 0$, there exists $A > 0$ such that

$$\min(m, n) \geq A \Rightarrow \mathbb{P}\left(|X_{m,n}| > \epsilon\right) \leq \zeta.$$

Additionally, the assertion $X_{m,n} = O_{\mathbb{P}}(1)$ means that for any $\varepsilon > 0$, there exist $A, M > 0$ such that

$$\sup_{m,n \geq A} \mathbb{P}\left(|X_{m,n}| > M\right) \leq \varepsilon.$$

The proof of the following result is in Appendix F.1.

Proposition 5 *Suppose that Assumptions (X1), (X2), (X3), (X4), and (A1), (A2) hold true with a compact subset Θ of $\mathbb{R}^p$ and assume that $\mathcal{R}^*$ has a unique minimizer denoted by θ^*. Then $\hat{\theta}^* - \theta^* = o_{\mathbb{P}}(1)$ as $m, n \to \infty$. Moreover, the excess risk satisfies $\mathcal{R}^*(\hat{\theta}^*) - \mathcal{R}^*(\theta^*) = o_{\mathbb{P}}(1)$.*

5.3 Convergence rate for linear least-squares estimators

We now illustrate our results with an upper-bound on the excess risk for linear least-squares estimators in the misspecified case. Here, the target risk is given by

$$\mathcal{R}^*(\theta) = \mathbb{E}\left[\left(Y^* - X^{*T}\theta\right)^2\right]$$

and any optimal linear rule should simply be satisfied:

$$\theta^* \in \arg\min_{\theta \in \mathbb{R}^d} \mathcal{R}^*(\theta).$$

Note that θ^* is unique the matrix $\mathbb{E}\left[X^* X^{*T}\right]$ is of full rank. The empirical risk is defined by

$$\mathcal{R}^*_m(\theta) = \frac{1}{m} \sum_{i=1}^{m} \left(Y^*_{n,i} - X_i^{*T}\theta\right)^2$$

and $\hat{\theta}^*$, the empirical risk minimizer, is given by

$$\hat{\theta}^* \in \arg\min_{\theta \in \mathbb{R}^d} \mathcal{R}^*_m(\theta).$$

The excess risk satisfies the following upper bound whose proof is given in Appendix F.1.

Proposition 6 *Suppose that Assumptions (X1), (X2), (X3), (X4) hold true. Suppose that the mapping $x \mapsto \mathbb{E}\left[Y|X = x\right]$ is Lipschitz and that the conditional expectation $\mathbb{E}[Y^2|X]$ is bounded. Suppose also that $\Gamma = \mathbb{E}[X^* X^{*T}]$ is positive definite. Then, we have*

$$\mathcal{R}^*(\hat{\theta}^*) - \mathcal{R}^*(\theta^*) = O_{\mathbb{P}}\left(m^{-1} + n^{-1} + n^{-2/d}\right).$$

Notes. The assumptions do not require the linear model for the (X_i, Y_i)'s to be valid, i.e., one can consider cases where $\mathbb{E}[Y|X]$ is not linear. Also, when the source data follows a non-linear model of the form $Y = r\left(X\right) + \varepsilon$ where ε and X are independent, our regularity assumptions means that r is Lipschitz on the compact set S_X.

6. Related work

A standard approach to covariate shift problems is to use some re-weighting in order to "transfer" the source distribution with density p_X to the target distribution with density q_X. This approach relies on the following type of estimates:

$$\hat{Q}_w(h) = n^{-1} \sum_{i=1}^{n} w(X_i) h(X_i, Y_i),$$

where ideally the function w would take the form q_X/p_X. Such a choice has the nice property that the expected value $\mathbb{E}[w(X_i)h(X_i, Y_i)]$ is equal to the targeted quantity $Q(h)$. This however often cannot be directly computed as p_X and q_X are unknown in practice. There are actually different ways to estimate w, and our goal here is to distinguish between two leading approaches.

Plug-in approach The plug-in approach is when the weights are computed using two estimates $\hat{p}_X$ and $\hat{q}_X$ in place of p_X and q_X, respectively; i.e., simply use $\hat{w} = \hat{q}_X/\hat{p}_X$ instead w in the above formula, see for instance (Shimodaira, 2000; Sugiyama et al., 2007, 2008). Note that the selection of hyper-parameters for $\hat{q}_X$ and $\hat{p}_X$ is needed and the n evaluation $(\hat{q}_X(X_i), \hat{p}_X(X_i))$ might be heavy in terms of computing time.

For the sake of clarity, we focus on a specific instance of covariate shift problem in which the target probability density q_X is known and $\hat{p}_X$ is the kernel density estimate (KDE), i.e., $\hat{p}_X^{KDE}(x) = (1/n)\sum_{i=1}^{n} K_b(x - X_i)$, where K_b typically is a Gaussian density with mean 0 and variance b^2 (a hyper-parameter to be tuned). This framework is clearly advantageous for weighted approaches as one usually unknown quantity, q_X, is now given. In this case, the analysis of $\hat{Q}_w(h)$ can be carried out using the decomposition $\hat{Q}_{\hat{w}}(h) - Q(h) = \hat{Q}_{\hat{w}}(h - Th) + \hat{Q}_{\hat{w}}(Th) - Q(h)$, with operator T defined by $Th(x) = \mathbb{E}[h(X,Y)|X = x]$. The first term above is a sum of centered random variable which (provided some conditions) satisfies the so-called Lindeberg condition, so that the central limit theorem implies that $\sqrt{n}\hat{Q}_{\hat{w}}(h - Th)$ is asymptotically Gaussian. The second term above is more complicated and the analysis can be derived using results in Delyon and Portier (2016); Clémençon and Portier (2018). Those results assert (under some conditions) that $\hat{Q}_{\hat{w}}(Th) - Q(h) = O_p(nb^d + b)$ (in case p_X is Lipschitz). As a consequence, we obtain, optimizing over b, that $\hat{Q}_{\hat{w}}(h) - Q(h) = O_p(n^{-1/2} + n^{-1/(1+d)})$. This is easily compared to our bound, when q_X is known, $n^{-1/2} + n^{-1/d}$, which is smaller than the one given before.

Direct weight estimation Huang et al. (2006) proposed *Kernel Mean Matching (KMM)* for estimating the ratios of the probability density functions of the source and the target distribution. They used the estimated ratios for weighting the source sample. Gretton et al. (2008) further studied this method theoretically and empirically. Sugiyama et al. (2007, 2008) proposed a method that estimates the ratios as a function by minimizing the Kullback-Leibler divergence between the source density function multiplied by the ratio function and the target density function. The estimated function can predict ratios even outside of the source sample, which enables cross-validation for hyper-parameter tuning. Kanamori et al. (2009) proposed constrained and unconstrained least squares methods for estimating the ratio function called *Least-Squares Importance Fitting (LSIF)* and *unconstrained LSIF (uLSIF)*. Yamada et al. (2013) developed its variant called *Relative uLSIF (RuLISF)*, which replaces the denominator of the ratio with a convex mixture of the source and the target density functions to circumvent issues caused by near-zero denominators. Zhang et al. (2021) proposed a covariate shift adaptation method that directly minimizes an upper bound of the target risk in order to avoid estimation of weights. The method shows great empirical performance while it does not exactly minimize the target risk and hence the minimizer converges to a biased solution.

Connection to treatment effect estimation One of the quantities of great interest in treatment effect estimation is the average treatment effect on the treated (ATT), $\mathbb{E}[Y^{(1)} - Y^{(0)} \mid W = 1]$, where $W \in \{0,1\}$ is a treatment assignment variable, $Y^{(1)}$ and $Y^{(0)}$ are potential outcomes corresponding to the treatment 1 and 0.[2] Suppose that we wish to estimate the ATT using i.i.d. observations of W and its outcome $Y := Y^{(W)}$ together with covariates X, $(Y_i, W_i, X_i)_{i=1}^N$. Under the standard assumptions (see e.g., Hernan and Robins (2023)) including the *conditional exchangeability* $Y^{(w)} \perp\!\!\!\perp W \mid X$, the *positivity* $P(\{W = w \mid X\}) > 0$, and the *consistency* $W = w \implies Y^{(w)} = Y^{(W)} = Y$, for each $w \in \{0,1\}$, the ATT equals the difference between

$$\begin{aligned}\mathbb{E}[Y^{(1)} \mid W = 1] &= \int y P_{Y^{(1)} \mid X, W=1}(dy) P_{X \mid W=1}(dx) \\ &= \int y P_{Y \mid X, W=1}(dy) P_{X \mid W=1}(dx) \\ &= \mathbb{E}[Y \mid W = 1] \end{aligned} \tag{2}$$

and

$$\begin{aligned}\mathbb{E}[Y^{(0)} \mid W = 1] &= \int y P_{Y^{(0)} \mid X, W=0}(dy) P_{X \mid W=1}(dx) \\ &= \int y P_{Y \mid X, W=0}(dy)\, r(x)\, P_{X \mid W=0}(dx) \\ &= \mathbb{E}[r(X) Y \mid W = 0], \end{aligned} \tag{3}$$

where $r(x)$ is the density ratio defined such that $r(x) dP_{X \mid W=0}(x) = dP_{X \mid W=1}(x)$. We can easily estimate the first term $\mathbb{E}[Y^{(1)} \mid W = 1]$ (Eq. (2)) by the conditional sample average $\frac{1}{N_1} \sum_{i=1}^N Y_i \cdot \mathbb{1}_{W_i=1}$, where $N_1 := \sum_{i=1}^N \mathbb{1}_{W_i=1}$ and $\mathbb{1}_E$ denotes the indicator random variable for any event E. Estimating the second term $\mathbb{E}[Y^{(0)} | W = 1]$ (Eq. (3)) is more involved. The sample average with the condition $W = 0$, $\frac{1}{N_0} \sum_{i=1}^M Y_i \cdot \mathbb{1}_{W_i=0}$, where $N_0 := \sum_{i=1}^N \mathbb{1}_{W_i=0}$, would be biased to $\mathbb{E}[Y \mid W = 0] \neq \mathbb{E}[Y^{(0)} \mid W = 1]$, but the bias is only due to the change in the conditional distributions of X given $W = 1$ and X given $W = 0$ quantified by $r(X)$, similarly to the covariate shift (see Eq. (3)). One way to correct the bias is to use an

[2] A common scenario is that we have a treated group (represented by treatment 1) and a non-treated, or controlled group (represented by treatment 0).

estimate $\widehat{r}$ of the ratio r for the weighted average $\frac{1}{N_0}\sum_{i=1}^{N}\widehat{r}(X_i)\cdot Y_i \cdot \mathbb{1}_{W_i=0}$, similarly to the reweighting approach to covariate shift adaptation, leading to the following estimate:

$$\widehat{\mathrm{ATT}} = \frac{1}{N_1}\sum_{i\colon W_i=1} Y_i - \frac{1}{N_0}\sum_{i\colon W_i=0} \widehat{r}(X_i)\cdot Y_i.$$

Another popular approach is the nearest neighbor matching Abadie and Imbens (2006). See also Rosenbaum (1995) for a broad introduction to matching problems for evaluating treatment effects. In Abadie and Imbens (2006), the ATT is estimated by

$$\overline{\mathrm{ATT}} = \frac{1}{N_1}\sum_{i\colon W_i=1}\left[Y_i - \hat{Y}_i(0)\right],$$

where $\hat{Y}_i(0) = \frac{1}{k}\sum_{j=1}^{N} Y_j(1-W_j)\mathbb{1}_{\|X_i-X_j\|\leq\tau_{n,k,X_i}}$ is the average of Y_j's over the k first NNs of X_i in the untreated group. The estimator takes the form

$$\overline{\mathrm{ATT}} = \frac{1}{N_1}\sum_{i\colon W_i=1} Y_i - \frac{1}{N_1}\sum_{i\colon W_i=0}\frac{K_k(i)}{k}Y_i,$$

where $K_k(i) = \sum_{j=1}^{N} W_j\mathbb{1}_{\|X_i-X_j\|\leq\tau_{n,k,X_j}}$ is the number of times observation i is used as a match, i.e., the number of times observation i is among the k NNs of X_j's found in the treated group. Note that $\overline{\mathrm{ATT}}$ coincides with $\widehat{\mathrm{ATT}}$ if $\hat{r}(X_i)$ is defined as $\frac{K_k(i)N_0}{kN_1}$. To see an analogy with our method, one can consider the case in which h does not depend on x, i.e. $h(x,y)=h(y)$. Using the notation from the present paper ($W=0$ and $W=1$ indicate the target and the source domain, respectively, with $N_0=n$ and $N_1=m$) the second term of $\overline{\mathrm{ATT}}$ above generalizes to the form

$$\frac{1}{m}\sum_{i=1}^{n}\frac{K_k(i)}{k}h(Y_i) = \frac{1}{m}\sum_{i=1}^{m}\int h(y)\hat{P}_{Y|X}(dy|X_i^*) \tag{4}$$

with $K_k(i) = \sum_{j=1}^{m} 1_{\{\|X_i-X_j^*\|\leq\hat{\tau}_{n,k,X_j^*}\}}$. The estimate (4) corresponds to the matching method mentioned in the notes following Proposition 2. On the other hand, our estimator applied to this case is given by

$$\frac{1}{m}\sum_{i=1}^{m}h(Y_{n,i}^*). \tag{5}$$

Both estimators are different when $k>1$ but they coincide as soon as $k=1$, for h only depending on y. In fact, $\hat{P}_{Y|X}(dy|X_i^*)$ has one single atom when $k=1$, so that sampling from it and evaluating the average are the same. Here are a few remarks:

- When $k>1$, Eq. (5) requires fewer evaluations of h than Eq. (4). This is relevant when evaluation of the function is time-consuming or costly such as observation from physical experiments.

- Our theoretical analysis is rather different from that of Abadie and Imbens (2006). Since they rely on the expression on the left side of Eq. (4), it is unclear whether they can or not handle the case when h depends on x (required for prediction purposes). In contrast, our approach is based on the decomposition given in Section 4.2, with bootstrap sampling error and nonparametric error, leveraging $\int\int h(y)\hat{P}_n(dy|x)Q(dx)$ as a centering term. Our results are more general because they include the case when h depends on x and also we can deal with both estimates (4) and (5) in the meantime, as mentioned in the notes following Proposition 2. Moreover, Proposition 3 implies a lower bound for Eq. (4) and we believe this result to be new in treatment effect literature.

- Recently, Lin et al. (2023) showed that the quantity $K_k(i)$, when properly normalized, is an estimator of the density ratio (between target and source in our case). The consistency as well as the minimax optimality require $k \to \infty$ while the previous work by Abadie and Imbens (2006) considered a fixed value of k, as in our case. Note that the same NN-based density ratio estimator was introduced by Loog (2012) in a covariate shift context. Consistency results for this density ratio estimator (without rates) are given in Sharpnack (2022) while rates and their minimax properties are obtained in Lin et al. (2023).

Other references The idea of nonparametric sampling is a standard one in the field of texture synthesis. In particular, the choice of 1-NN resampling was often used as a fast method to generate new textures from a small sample. See Truquet (2011) for a literature review in this context. Our conditional sampling framework bears resemblance with traditional bootstrap sampling (Efron, 1992) as there is random generation according to some estimated distribution. In contrast, the original bootstrap method is usually made up using draws from the standard empirical measure $(1/n)\sum_{i=1}^{n}\delta_{X_i,Y_i}$. Here another distribution, $\hat{Q}$, has been used to generate new samples. Moreover, our goal is totally different here. While the bootstrap technique was initially introduced for making inference, here the goal is to estimate an unknown quantity $Q(h)$ which appears in many machine learning tasks. Kpotufe and Martinet (2021) theoretically study covariate shift adaptation under the assumption that we have access to a labeled sample both from the source and the target distribution. Although they consider a k-nearest-neighbor-based method, it is essentially different from ours since they perform the k-NN method on the union of the source and the target sample. Lee (2013) proposed pseudo-labeling unlabeled data in the context of semi-supervised learning. Wang (2023) proposed a hyper-parameter selection method for kernel ridge regression under covariate shift using pseudo-labeling. The author focuses on model selection in regression problems while we study the mean estimation that can be applied to a wider range of supervised learning problems.

7. Extensions

Several ways to extend our method beyond the mean estimation problem are considered in this section.

Heterogeneity in target distributions The case where the target covariates distribution Q_X changes across the data might be of interest if one wishes to aggregate several pieces of target data whose covariates distributions are not necessarily the same. This might occur when the target data is obtained by gathering individuals from different countries, and consequently, the distributions are not the same anymore, or when the time between the measurements has caused some changes in the distribution.

While such an heterogeneity in target data might be seen as more complicated at first glance, it actually can be examined using a similar decomposition and the same tools as the ones used to obtain the non-asymptotic bound in Theorem 1. More formally, the target distribution here is $Q = (1/m)\sum_{i=1}^m Q_i$ with $Q_i = P_{Y|X}Q_{X,i}$. For each $m \geq 1$ and $n \geq 1$, let $(X_i^*, Y_{n,i}^*)_{1\leq i\leq m}$ be a collection of random variables conditionally independent given $(X_i, Y_i)_{i=1}^n$ and such that for each $i = 1, \dots, m$, $(X_i^*, Y_{n,i}) \sim \hat{Q}_{i,n}$ with $\hat{Q}_{i,n} = \hat{P}_{Y|X}Q_{X,i}$. The quantity of interest and the proposed estimator are therefore slightly different from before bug given by

$$Q(h) = m^{-1}\sum_{i=1}^m Q_i(h), \qquad \hat{Q}^*(h) = m^{-1}\sum_{i=1}^m h(X_i^*, Y_{n,i}^*),$$

respectively. The decomposition is

$$\hat{Q}^*(h) - Q(h) = m^{-1}\sum_{i=1}^m \left\{h(X_i^*, Y_{n,i}^*) - \hat{Q}_{i,n}(h)\right\} + m^{-1}\sum_{i=1}^m \left\{\hat{Q}_{i,n}(h) - Q_i(h)\right\}.$$

The non-asymptotic analysis of the bootstrap sampling error (first term above) is similar to before as the Bernstein inequality is tailored to non-identically distributed variables. We obtain that the rate $O(m^{-1/2})$ as before by simply requiring a bound on the variance of each random variables. The other term, concerning the conditional distribution, error can be analyzed by writing

$$\left\| m^{-1}\sum_{i=1}^m \left\{\hat{Q}_{i,n}(h) - Q_i(h)\right\} \right\|_{L_2} \leq m^{-1}\sum_{i=1}^m \|\hat{Q}_{i,n}(h) - Q_i(h)\|_{L_2}$$

where $\|W\|_{L_2} = \sqrt{\mathbb{E}W^2}$ and therefore we can directly apply Proposition 3 (given the assumptions are satisfied for each i uniformly). We finally obtain the rate $m^{-1/2} + n^{-1/d} + n^{-1/2}$, similar to the one obtained before.

Stochastic gradient descent Our sampling approach can be easily combined with the well-known stochastic gradient descent algorithm (and more generally with stochastic approximation), where only a small part of the data is used at each step to update the estimator, reducing the number of operations at each iteration compared with the (batched) gradient descent.

To illustrate this idea, consider the empirical risk minimization problem described in Section 5 where one is interested in solving $\min_\theta\{\mathcal{R}^*(\theta) := \mathbb{E}[m_\theta(Y^*, X^*)]\}$ where $\theta \mapsto m_\theta(y, x)$ is differentiable. Suppose that n source samples have been obtained making the conditional distribution $\hat{P}_{Y|X}$ available for sampling new points. Then the algorithm at step $i \geq 1$, might proceed by first generating X_i^* and then $Y_{n,i}^* \sim \hat{P}_{Y|X_i^*}$. This means finding the nearest neighbor to X_i^* among the source data and represents only $\log(n)$ operations using the k-d tree. Having this been done, the update is simply

$$\theta_i = \theta_{i-1} - \gamma_i \nabla_\theta m_\theta(Y_{n,i}^*, X_i^*).$$

It results that each iteration in the above is similar to standard stochastic gradient descent, the only difference being the additional 1-nearest neighbor search. We stress that this is contrasting with the re-weighting approach using kernels for which a new sample, say X_i, would require evaluating $\hat{p}_X(X_i)$ and therefore would need to compute all n distances between X_j, $j = 1, \dots, n$ and the new X_i.

Semiparametric estimation Simulating the labels to obtain a new sample is also convenient in semiparametric problems where quantities of interest often involve additional estimated parameters. Typical semiparametric problems involve expectations of functions that are indexed by an unknown parameter, $\mathbb{E}[h_\theta(X, Y)]$, and θ is estimated from the data using some transformation $\hat{\theta}$ of the sample. In such a situation, while estimating θ using reweighting is unclear without more information on $\hat{\theta}$, one can directly use our sampling approach by introducing $m^{-1} \sum_{j=1}^{m} h_{\hat{\theta}^*} \left(X_j^*, Y_j^* \right)$ where $\hat{\theta}^* = \hat{\theta}\left(X_1^*, Y_1^*, \dots, X_m^*, Y_m^*\right)$. This allows to obtain a semiparametric estimate with covariate shift adaptation. See Van der Vaart (2000), Chapters 19.4 and 25.8 for more details and examples in parametric or semiparametric estimation.

8. Experiments

The main purpose of the experiments is to compare our k-NN-CSA approach with several state-of-the-art competitors when facing multiple situations from mean estimation to empirical risk minimization with synthetic and real-world data.

We consider the following instances of our proposed method.

1-NN-CSA: the *Conditional Sampling Covariate-shift Adaptation (CSA)* (Algorithm 1) with k-Nearest Neighbor (k-NN) conditional sampler (Algorithm 2) with $k = 1$.

$\log n$**-NN-CSA:** the same as above but with $k = \log n$.

We use the Python module `cKDTree` (Archibald, 2008) from SciPy (Virtanen et al., 2020) for nearest neighbor search in our methods. We compare them with the following existing covariate-shift adaptation methods.

KDE-R-W (KDE-Ratio-Weighting): the weighting method using the ratio of the Kernel Density Estimates (KDEs) of p_X and q_X (see Section 6).

KMM-W (KMM-Weighting): the weighting method directly estimating q_X/p_X using the *Kernel Mean Matching (KMM)* (Huang et al., 2006; Gretton et al., 2008). We use the Gaussian kernel.

KLIEP-W (KLIEP-Weighting): the weighting method estimating q_X/p_X by the *Kullback-Leibler Importance Estimation Procedure (KLIEP)* (Sugiyama et al., 2007, 2008). The linear combination of the Gaussian basis functions centered at the sample points are used for modeling the weight function.

KLIEP100-W: the same as KLIEP-W but using only 100 randomly subsampled kernel centers (Sugiyama et al., 2007, 2008) for reducing the time- and space-complexities.

RuLSIF-W (RuLSIF-Weighting): the weighting method using $q_X/(\alpha p_X + (1-\alpha)q_X)$ estimated by *Relative unconstrained Least-Squares Importance Fitting (RuLSIF)* (Yamada et al., 2013), where $\alpha \in [0,1]$ is a hyper-parameter. We use the default value $\alpha = 0.1$. As a model of the weight function, the Gaussian basis functions centered at the sample points are used.

RuLSIF100-W: the same as RuLSIF-W but using only 100 randomly subsampled kernel centers (Yamada et al., 2013) for reducing the time- and space-complexities.

See Section 6 for more explanations of those methods. For KMM-W, KLIEP-W, and RuLSIF-W, we used the implementations from *Awesome Domain Adaptation Python Toolbox (ADAPT)* (de Mathelin et al., 2021). All the computations were performed on the cluster, *Grid5000* (Balouek et al., 2013). For the methods using Gaussian basis functions (KLIEP-W, KLIEP100-W, RuLSIF-W, RuLSIF100-W), we use 5-fold cross-validation for choosing the Gaussian bandwidth from $\{0.001, 0.01, 0.1, 1, 10\}$. KMM-W does not offer a way to do cross-validation, and we fixed to 1. More details are in the supplementary material.

Furthermore, we also report the results for the following baseline method and ideal method.

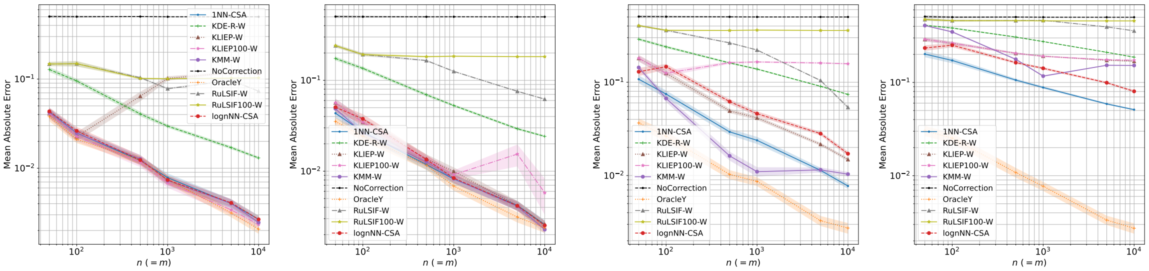
NoCorrection: the method that takes the average $\frac{1}{n}\sum_{i=1}^{n} h(X_i, Y_i)$ only using the source sample $(X_i, Y_i)_{i=1}^{n} \sim P^m$, ignoring the target sample.

OracleY: the result for taking the average $\frac{1}{m}\sum_{i=1}^{m} h(X_i^*, Y_i^\circ)$ using a sample $(X_i^*, Y_i^\circ)_{i=1}^{m} \sim Q^m$. Note that Y_i° are not available in practical scenarios of our interest and made invisible to other methods.

We conduct experiments in three setups, detailed below, with different sample sizes $n\ (= m)$ and data dimensionalities d: $(n, d) \in \{50, 100, 500, 1000, 5000, 10000\} \times \{1, 2, 5, 10\}$. Each experiment is repeated 50 times with different random seeds.

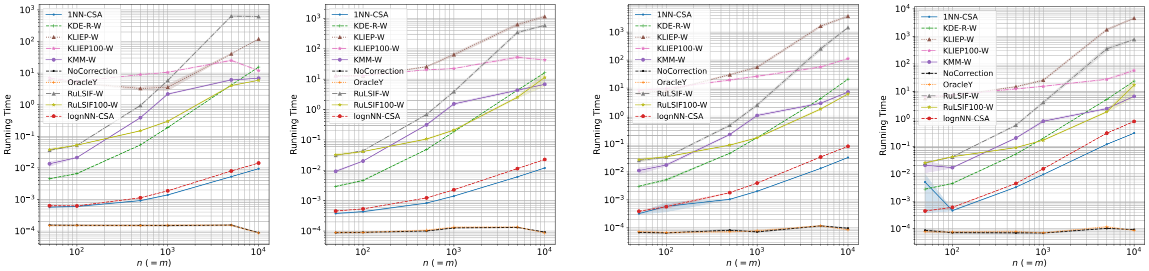
Setup of Experiment E1 (mean estimation with synthetic data): The task here is to estimate $Q(h) = \iint y\, P_{Y|X}(dy \mid x) Q_X(dx)$ under the following setup. We define h by $h(x, y) = y$, P_X as the uniform distribution over $[-1, 1]^d$, Q_X as that over $[0, 1] \times [-1, 1]^{d-1}$, and $P_{Y|X=x}$ as the normal distribution with mean x_1 and standard deviation 0.1. Figure 7a in Appendix H shows an illustration of the setup. In this setup, we have $P(h) = \iint y\, P_{Y|X}(dy \,|\, x) P_X(dx) = 0$ while $Q(h) = \iint y\, P_{Y|X}(dy \,|\, x) Q_X(dx) = 0.5$. Because of this difference, covariate shift adaptation is essential for correctly estimating $Q(h)$.

Comparison of estimation errors for Experiment E1: The results are presented in Figure 1. First, the errors for NoCorrection are not decreasing as the sample sizes increase, ending up with large errors in all cases, because of the bias due to the covariate shift. Other methods with covariate-shift adaptation had always smaller errors than that of this baseline. Excluding OracleY, an ideal method unavailable in practice, KLIEP100-W, KMM-W, 1-NN-CSA, and $\log n$-NN-CSA were among the best for smaller dimensionalities $d \in \{1, 2\}$ (Figures 1a and 1b). For the larger dimensionalites $d \in \{5, 10\}$, KMM-W and 1-NN-CSA outperformed other methods. In particular, 1-NN-CSA gave outstanding performances in many cases except $d = 5$ and $n \in \{100, 500, 1000\}$, for which KMM-W was even better. The errors of most methods roughly follow power laws, where the slope of a line corresponds



(a) $d = 1$ (b) $d = 2$ (c) $d = 5$ (d) $d = 10$

Figure 1: Mean Absolute Errors (MAEs) for Experiment E1 (estimation of $\int y\, Q(dy)$). The horizontal axis is for the sample sizes n $(= m)$, and the vertical axis is for the mean absolute error of each estimate. The four figures are for different data dimensionalities.



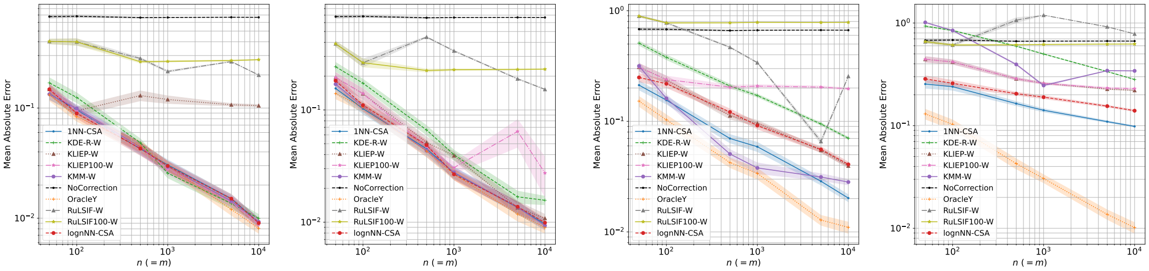
(a) $d = 1$ (b) $d = 2$ (c) $d = 5$ (d) $d = 10$

Figure 2: Running times for Experiment E1. The horizontal axis is for the sample sizes n $(= m)$, and the vertical axis is for the mean running time of each method. The four figures are for different data dimensionalities.

to the power of the convergence rate (steeper is better). 1-NN-CSA and $\log n$-NN-CSA seem to have the steepest slopes for $d = 10$, although comparison is difficult for the lower dimensionalities.

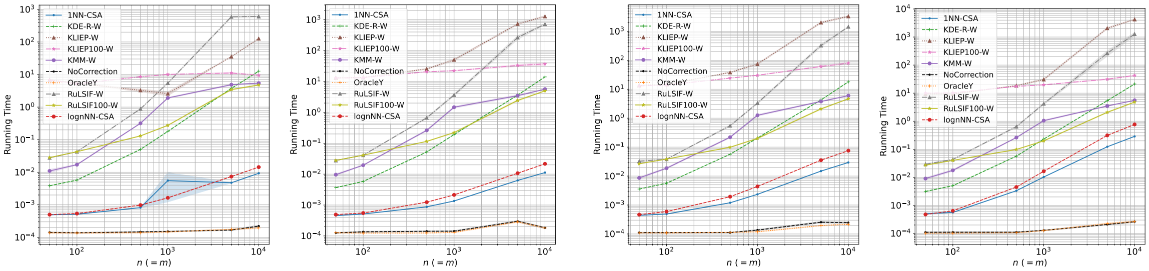
Comparison of running times in Experiment E1: Figure 2 shows the comparison in running times. 1- and $\log n$-NN-CSA were much faster than other methods in all cases except for $(d, n) = (10, 50)$. Their advantage is most pronounced for larger sample sizes. For instance, 1- and $\log n$-NN-CSA were at least 100 times faster than other methods for $(n, d) = (10000, 1)$ (Figure 2a).

Setup of Experiment E2 (risk estimation with synthetic data): In this experiment, we compare the methods in the context of risk estimation of a fixed function f_0. Setting $f_0 \colon x \mapsto -x_1$, where x_1 is the first coordinate of x, we estimate the expected loss (i.e., risk) of $f_0(X)$ with the square loss in predicting the response Y when (X, Y) follows Q. In other words, we set h as $h(x, y) := (y - f_0(x))^2$, and the goal is to estimate the risk $Q(h) = \int (y - f_0(x))^2 Q(dx, dy)$. We use the uniform distribution over $[-1, 1]^d$ for P_X and that over $[0, 1] \times [0, 1]^{d-1}$ for Q_X. The conditional distribution $P_{Y|X=x}$ is the normal distribution with mean $|x_1|$ and standard deviation 0.1 for any $x \in \mathcal{X} := [-1, 1]^d$. Under



(a) $d = 1$ (b) $d = 2$ (c) $d = 5$ (d) $d = 10$

Figure 3: Estimation errors for Experiment E2 (estimation of $\int (y - f_0(x))^2\, Q(dx, dy)$)



(a) $d = 1$ (b) $d = 2$ (c) $d = 5$ (d) $d = 10$

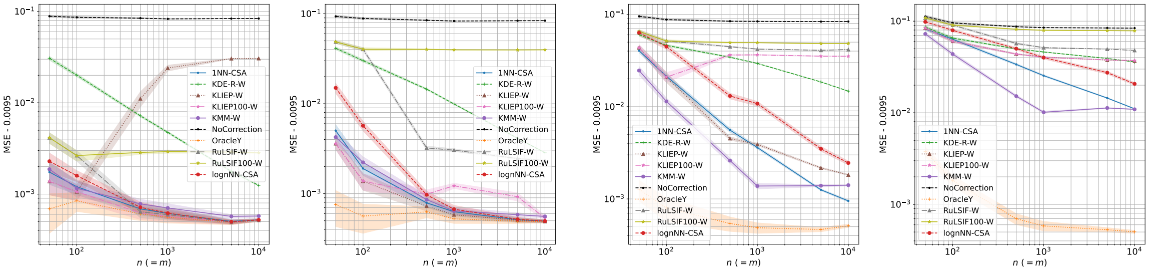
Figure 4: Running times in Experiment E2

this setup, the function f_0 performs poorly on the support of Q_X and should incur a large risk. See Figure 7b for an illustration of the setup. In this setup, the risks under P_X and Q_X largely differ because f_0 fits (X, Y) well in a half of the support of P_X but not in that of Q_X.

Comparison of estimation errors for Experiment E2: We present the estimation errors for Experiment E2 in Figure 3. KMM-W, 1-NN-CSA, $\log n$-NN-CSA gave similar results, almost matching those of OracleY, while KMM-W and 1-NN-CSA were advantageous for $d = 5$, and 1-NN-CSA outperformed other methods for $d = 10$. We can notice that KDE-W, KMM-W, RuLSIF-W, and RuLSIF100-W did not always improve errors over NoCorrection (Figures 1c and 1d). Some methods such as KMM-W and KLIEP-W showed great performance in some cases while giving poor results in other cases. In contrast, 1-NN-CSA showed stable and often best performances in these experiments.

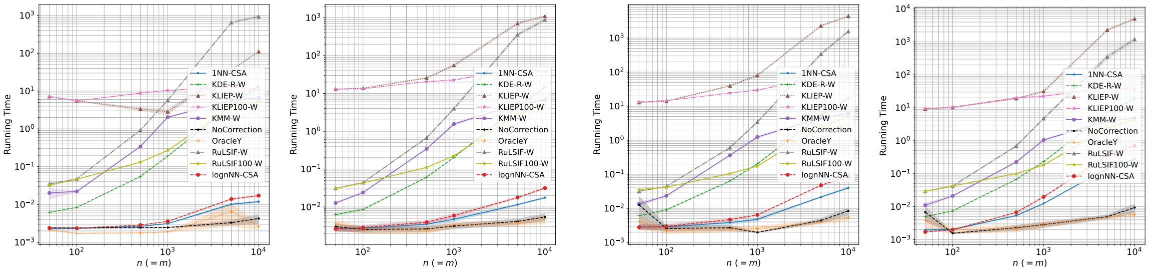
Comparison of running times in Experiment E2: The running times in Experiments E2 (Figure 4) were very similar to those in Experiment E1 (Figure 4), but we can see more clearly that 1- and $\log n$-NN-CSA outperform other methods even for the smallest sample size.

Setup of Experiment E3 (linear regression with synthetic data): Next, we present experiments of linear regression. Using samples from the same source and test distributions as in Experiment E2, we perform the ordinary least squares after covariate adaptation. More precisely, we aim to optimize the parameters $\theta \in \mathbb{R}^d$ of the model $f_\theta \colon \mathbb{R}^d \to \mathbb{R}, x \mapsto \theta^\top x$



(a) $d = 1$ (b) $d = 2$ (c) $d = 5$ (d) $d = 10$

Figure 5: Mean Squared Errors (MSE) (subtracted by 0.0095) for Experiment E3 (linear regression)



(a) $d = 1$ (b) $d = 2$ (c) $d = 5$ (d) $d = 10$

Figure 6: Computation times (seconds) in Experiment E3

so that the Mean Squared Error (MSE) $\mathbb{E}[(Y^* - f_\theta(X^*))^2]$ in the target domain will be minimized. To do so, we minimize the MSE estimated by each covariate shift adaptation method.

Comparison of estimation errors for Experiment E3: The results are summarized in Figures 5.[3] KMM-W performed better than any other methods for the higher dimensions $d \in \{5, 10\}$ and the small-to-moderate sample sizes $50 \leq n \leq 500$, 1-NN-CSA being the second best. For $n = 10000$, 1-NN-CSA showed performance better than or comparable to KMM-W.

Comparison of running times in Experiment E3: As in Experiments E1–E2, 1-NN-CSA and $\log n$-NN-CSA finished their computations faster than the other adaptation methods by large margins (Figure 6).

In Experiments E1–E3, the proposed methods, 1- and $\log n$-NN-CSA were able to finish computation much faster than other adaptation methods without compromising on the statistical performance. $\log n$-NN-CSA did not show advantages in accuracy, with increased computation costs. We can conclude that 1-NN-CSA is preferred over $\log n$-NN-CSA. A reason that we were not able to conduct experiments with larger sample sizes than 10000 is that the existing adaptation methods have too demanding computational requirements.

[3] We plot the MSEs subtracted by 0.0095 to better present the curves in the region close to the minimum population MSE 0.01 while keeping values positive.

For instance, the running times of RuLSIF-W in Figure 6c grows about 100 times as the sample size increases by 10 times, taking more than 10^3 seconds for $n = 10000$. For $n = 10^5$, we would need at least $10^3 \times 100$ seconds, that is 27 hours of compute for a single run. In contrast, the time complexity of 1-NN-CSA being $\mathcal{O}(n \log n)$ and its running time less than one second for $n = 10^4$, we can estimate its running time for $n = 10^5$ as $1 \times (10^5 \log 10^5)/(10^4 \log 10^4) = 12.5$ seconds. 1-NN-CSA would stay feasible in applications of even larger scales.

The previous methods construct the distance matrix between pairs of data points, which takes running time and memory space quadratic in the sample size. Additionally, RuLSIF-W computes the inverse of the distance matrix, taking cubic running time. KMM-W and KLIEP-W solve convex optimization problems with iterative procedures, for which the implementations from de Mathelin et al. (2021) use stopping criteria based on objective function values. This resulted in good accuracy and milder growth in running time in our experiments. However, tuning the solvers can be involved in practice. In contrast, k-NN-CSA does not have such subtle issues around optimization solvers: we only have to perform nearest neighbor search.

In all cases, we can observe that 1-NN-CSA showed clear power-law, with nearly straight lines in the logarithmic scales. This is a significant advantage in predicting returns when one invests on increasing the sample size.

Experiment E4 (linear regression and logistic regression with benchmark datasets): We use regression benchmark datasets, `diabetes`[4], `california` (Pace and Barry, 1997)[5] and classification datasets, `twonorm` (Breiman, 1996)[6] and `breast_cancer`[4]. We apply the ridge regression and the logistic regression, respectively. The evaluation metric is the mean squared error for the regression tasks and the classification accuracy for classification tasks. We synthetically introduce covariate shift by subsampling test data. See Appendix I for more details.

Remark: For fair comparison, the benchmark experiments presented in this paper follow the standard protocol used in the literature as similarly done in previous research (Gretton et al., 2008; Kanamori et al., 2009; Yamada et al., 2013; Sugiyama et al., 2007, 2008): we apply biased resampling to synthetically simulate a target dataset under covariate shift. It is thus important to note that they are not completely real-world data. Nevertheless, this ensures that the methods are tested in isolation from other types of distribution shifts while using real data for the source covariate distribution as well as the conditional distributions.

Results for Experiment E4: Table 1 shows the obtained MSEs and classification accuracies. 1-NN-CSA and $\log n$-NN-CSA gave the best performance for `california` and performances comparable to the best for `breast_cancer`. For the other datasets, different methods performed the best depending on the dataset. On the other hand, in terms of running time, 1NN-CSA was consistently faster than the previous methods (Table 2).

Our experiments show that the proposed method is almost always faster than the previous methods and gives great accuracy in many cases, even though it is not always the best.

[4] Available at `https://archive.ics.uci.edu/ml/index.php`.

[5] Available at `https://www.dcc.fc.up.pt/~ltorgo/Regression/cal_housing.html`.

[6] Available at `https://www.cs.utoronto.ca/~delve/data/datasets.html`.

Table 1: MSE/accuracy for regression/classification benchmark datasets. We repeat the experiment using 50 different random subsamples and calculate the average scores (and standard errors). The results comparable to the best in terms of Wilcoxon's signed rank test with significance level 1% are shown in bold fonts.

	Regression (MSE)		Classification (accuracy)	
	`diabetes`	`california`	`breast_cancer`	`twonorm`
1NN-CSA	3470 (35)	**0.146** (0.001)	**0.9633** (0.002)	0.9327 (0.002)
lognNN-CSA	3605 (40)	0.150 (0.001)	0.9595 (0.002)	0.9293 (0.002)
KDE-R-W	3673 (52)	3.864 (1.067)	0.9596 (0.002)	0.5260 (0.009)
KMM-W	3831 (60)	3.702 (1.160)	0.9594 (0.002)	**0.9583** (0.001)
KLIEP-W	**3221** (31)	2.896 (0.798)	**0.9648** (0.002)	0.9482 (0.001)
KLIEP100-W	**3223** (31)	3.034 (0.843)	**0.9648** (0.002)	0.9480 (0.001)
RuLSIF-W	3235 (31)	3.039 (0.843)	0.7794 (0.015)	0.9512 (0.001)
RuLSIF100-W	3238 (31)	3.045 (0.844)	0.7794 (0.015)	0.9539 (0.001)

Table 2: Total running times in seconds spent for the training including the hyper-parameter tuning (if any) for benchmark datasets. We repeat the experiment using 50 different random subsamples dataset and calculate the average running times (and standard errors). The results comparable to the best in terms of Wilcoxon's signed rank test with significance level 1% are shown in bold.

	`diabetes`	`california`	`breast_cancer`	`twonorm`
1NN-CSA	**0.0015** (0.0000)	**0.0084** (0.0001)	**0.0036** (0.0000)	**0.0051** (0.0000)
lognNN-CSA	0.0016 (0.0000)	0.0128 (0.0001)	0.0037 (0.0000)	0.0052 (0.0000)
KDE-R-W	0.0078 (0.0000)	0.2121 (0.0008)	0.0117 (0.0000)	0.0124 (0.0000)
KMM-W	0.0373 (0.0015)	0.4067 (0.0038)	0.0542 (0.0014)	0.0220 (0.0006)
KLIEP-W	7.602 (0.051)	29.98 (0.34)	8.67 (0.07)	8.86 (0.16)
KLIEP100-W	7.501 (0.045)	16.91 (0.07)	8.68 (0.07)	8.26 (0.10)
RuLSIF-W	0.0575 (0.0014)	1.686 (0.011)	0.0529 (0.0020)	0.2014 (0.0016)
RuLSIF100-W	0.0401 (0.0007)	0.1237 (0.0004)	0.0454 (0.0014)	0.0391 (0.0002)

1-NN-CSA is highly recommended as an off-the-shelf method applicable even in larger scales, although the previous methods such as KMM-W, KLIEP-W, and RuLSIF-W should not be neglected, as far as the computational budget allows. The times spent for adaptation are summarized in Table 2, showing that the proposed methods 1-NN-CSA and $\log n$-NN-CSA are much faster than other methods.

9. Conclusion

We proposed a k-NN-based covariate shift adaptation method. We provided error bounds, which suggest setting $k = 1$ is among the best choices. This resulted in a scalable nonparametric method with no hyper-parameter. For future research directions, one could complete our results for the parametric inference on the target domain, in particular for finding the asymptotic distribution of M-estimators. For the average treatment effect, Abadie and Imbens (2006) derived asymptotic normality of their estimator and it could be interesting to get a similar result in our sampling context as well as investigating efficiency properties as conducted by Lin et al. (2023) relying on de-biased machine learning approach Chernozhukov et al. (2018). Non-parametric estimation on the target domain could be also an interesting direction. Ma et al. (2023) showed that in the reproducing kernel Hilbert space framework, getting optimal rates of convergence requires to reweight the estimator when the density ratio is unbounded. A natural question is whether optimal rates can be reachable with our k-NN conditional sampling instead of reweighting. Finally, it may be useful to extend our approach with approximate nearest neighbor methods for further scalability.

Acknowledgments

Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see `https://www.grid5000.fr`). IY was supported by the *Allocation d'Installation Scientifique (AIS) 2023* from Rennes Métropole.

References

Alberto Abadie and Guido W Imbens. Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267, 2006.

Gholamali Aminian, Mahed Abroshan, Mohammad Mahdi Khalili, Laura Toni, and Miguel Rodrigues. An information-theoretical approach to semi-supervised learning under covariate-shift. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 7433–7449, 28–30 Mar 2022.

Anne M. Archibald. cKDTree, 2008. URL `https://github.com/scipy/scipy/blob/main/scipy/spatial/_ckdtree.pyx`.

Daniel Balouek, Alexandra Carpen Amarie, Ghislain Charrier, Frédéric Desprez, Emmanuel Jeannot, Emmanuel Jeanvoine, Adrien Lèbre, David Margery, Nicolas Niclausse, Lucas Nussbaum, Olivier Richard, Christian Pérez, Flavien Quesnel, Cyril Rohr, and Luc Sarzyniec. Adding virtualization capabilities to the Grid'5000 testbed. In Ivan I. Ivanov, Marten van Sinderen, Frank Leymann, and Tony Shan, editors, *Cloud Computing and Services Science*, volume 367 of *Communications in Computer and Information Science*, pages 3–20. Springer International Publishing, 2013.

Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517, 1975.

Thomas B. Berrett, Richard J. Samworth, and Ming Yuan. Efficient multivariate entropy estimation via k-nearest neighbour distances. *The Annals of Statistics*, 47(1):288–318, 2019.

Gérard Biau and Luc Devroye. *Lectures on the nearest neighbor method*, volume 246. Springer, 2015.

Jose Blanchet, Haoxuan Chen, Yiping Lu, and Lexing Ying. When can regression-adjusted control variate help? Rare events, Sobolev embedding and minimax optimality. In *Advances in Neural Information Processing Systems 36*, NeurIPS 2023, 2023.

Vladimir Igorevich Bogachev and Maria Aparecida Soares Ruas. *Measure theory*, volume 2. Springer Science & Business Media, 2007.

Leo Breiman. Bias, variance, and arcing classifiers. *Technical Report 460, Statistics Department, University of California, Berkeley*, 1996.

Timothy I Cannings, Thomas B Berrett, and Richard J Samworth. Local nearest neighbour classification with applications to semi-supervised learning. *The Annals of Statistics*, 48 (3):1789–1814, 2020.

Lingjiao Chen, Matei Zaharia, and James Y. Zou. Estimating and explaining model performance when both covariates and labels shift. In *Advances in Neural Information Processing Systems 35*, NeurIPS 2022, pages 11467–11479, 2022.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018.

Stéphan Clémençon and François Portier. Beating Monte Carlo integration: a nonasymptotic study of kernel smoothing methods. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84, pages 548–556. PMLR, 09–11 Apr 2018.

Sanjoy Dasgupta and Samory Kpotufe. Optimal rates for k-NN density and mode estimation. In *Advances in Neural Information Processing Systems 27*, NIPS 2014, 2014.

Antoine de Mathelin, Mounir Atiq, Guillaume Richard, Alejandro de la Concha, Mouad Yachouti, François Deheeger, Mathilde Mougeot, and Nicolas Vayatis. ADAPT : Awesome Domain Adaptation Python Toolbox, 2021. arXiv:2107.03049 [cs.LG].

Bernard Delyon and François Portier. Integral approximation by kernel smoothing. *Bernoulli*, 22(4):2177–2208, 2016.

Luc Devroye, László Györfi, Adam Krzyżak, and Gábor Lugosi. On the strong universal consistency of nearest neighbor regression function estimates. *The Annals of Statistics*, 22(3):1371–1385, 1994.

Luc Devroye, Paola G Ferrario, László Györfi, and Harro Walk. Strong universal consistent estimate of the minimum mean squared error. *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, pages 143–160, 2013a.

Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013b.

Luc Devroye, László Györfi, Gábor Lugosi, and Harro Walk. A nearest neighbor estimate of the residual variance. *Electronic Journal of Statistics*, 12:1752–1778, 2018.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in Statistics: Methodology and Distribution*, pages 569–593. Springer, 1992.

Jerome H. Friedman, Jon Louis Bentley, and Raphael Ari Finkel. An algorithm for finding best matches in logarithmic expected time. *ACM Transactions on Mathematical Software (TOMS)*, 3(3):209–226, 1977.

Sébastien Gadat, Thierry Klein, and Clément Marteau. Classification in general finite dimensional spaces with the k-nearest neighbor rule. *The Annals of Statistics*, 44(3):982–1009, 2016.

Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten Borgwardt, and Bernhard Schölkopf. Covariate Shift by Kernel Mean Matching. In *Dataset Shift in Machine Learning*, pages 131–160. The MIT Press, December 2008.

László Györfi and Roi Weiss. Universal consistency and rates of convergence of multiclass prototype algorithms in metric spaces. *Journal of Machine Learning Research*, 22(151): 1–25, 2021.

László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression.* Springer Science & Business Media, 2006.

Steve Hanneke, Aryeh Kontorovich, Sivan Sabato, and Roi Weiss. Universal bayes consistency in metric spaces. *The Annals of Statistics*, 49(4):2129–2150, 2021.

Miguel A. Hernan and James M. Robins. *Causal Inference: What If.* Chapman & Hall/CRC Monographs on Statistics & Applied Probab. CRC Press, 2023.

Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex Smola. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems 19*, NIPS 2006, 2006.

Heinrich Jiang. Non-asymptotic uniform rates of consistency for k-NN regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3999–4006, 2019.

Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10(48):1391–1445, 2009.

Aryeh Kontorovich and Roi Weiss. A Bayes consistent 1-NN classifier. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pages 480–488. PMLR, 2015.

Samory Kpotufe and Guillaume Martinet. Marginal singularity and the benefits of labels in covariate-shift. *The Annals of Statistics*, 49(6):3299–3323, 2021.

Quoc Le, Tamás Sarlós, Alex Smola, et al. Fastfood—approximating kernel expansions in loglinear time. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, 2013.

Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, 2013.

Rémi Leluc, François Portier, Johan Segers, and Aigerim Zhuman. Speeding up monte carlo integration: Control neighbors for optimal convergence. *To appear in Bernoulli. ArXiv:2305.06151*, 2023.

Zhexiao Lin, Peng Ding, and Fang Han. Estimation based on nearest neighbor matching: from density ratio to average treatment effect. *Econometrica*, 91(6):2187–2217, 2023.

Marco Loog. Nearest neighbor-based importance weighting. In *2012 IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6. IEEE, 2012.

Cong Ma, Reese Pathak, and Martin J. Wainwright. Optimally tackling covariate shift in RKHS-based nonparametric regression. *The Annals of Statistics*, 51(2):738–761, 2023.

R. Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297, 1997.

François Portier. Nearest neighbor empirical processes. *Bernoulli*, 31(1):312–332, 2025.

Paul R. Rosenbaum. *Observational Studies*. Springer, 1995.

Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems 30*, NIPS 2017, 2017.

James Sharpnack. On L^2-Consistency of Nearest Neighbor Matching. *IEEE Transactions on Information Theory*, 69(6):3978–3988, 2022.

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, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in Neural Information Processing Systems 20*, NIPS 2007, pages 1433–1440, 2007.

Masashi Sugiyama, Taiji Suzuki, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699–746, December 2008.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12:389–434, 2012.

Lionel Truquet. On a nonparametric resampling scheme for Markov random fields. *Electronic Journal of Statistics*, 5:1503–1536, 2011.

Aad W. Van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge University Press, 2000.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

Kaizheng Wang. Pseudo-Labeling for Kernel Ridge Regression under Covariate Shift, March 2023. arXiv:2302.10160 [cs, math, stat].

James G. Wendel. Note on the gamma function. *The American Mathematical Monthly*, 55 (9):563, 1948.

Christopher K. I. Williams and Matthias Seeger. Using the Nyström Method to Speed up Kernel Machines. In *Advances in Neural Information Processing Systems 13*, NIPS 2000, pages 661–667, 2000.

Makoto Yamada, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Masashi Sugiyama. Relative Density-Ratio Estimation for Robust Distribution Comparison. *Neural Computation*, 25(5):1324–1370, May 2013.

Tianyi Zhang, Ikko Yamane, Nan Lu, and Masashi Sugiyama. A One-Step Approach to Covariate Shift Adaptation. *SN Computer Science*, 2(4):319, 2021.

Appendix A. Preliminary results

The first preliminary result is concerned about the order of magnitude of $P_X(B(x,\tau))$ for which we obtain a lower bound and an upper bound.

Lemma 1 *Under (X1), (X2), and (X3), it holds, for every $x \in S_X$ and $\tau \in [0,T]$,*

$$M_{1,d}\tau^d \le P_X(B(x,\tau)) \le M_{2,d}\tau^d,$$

with $M_{1,d} = cb_X V_d$ and $M_{2,d} = U_X V_d$.

Proof The proof of the lower bound follows from

$$P_X(B(x,\tau)) = \int_{B(x,\tau)\cap S_X} p_X(y)dy \geq b_X \int_{B(x,\tau)\cap S_X} dy \geq b_X c \int_{B(x,\tau)} dy,$$

where we have used (X3) to get the first inequality and then (X2) to obtain the second one. We conclude by change of variable $y = x + \tau u$. The proof of the upper bound is similar:

$$P_X(B(x,\tau)) = \int_{B(x,\tau)\cap S_X} p_X(y)dy \leq U_X \int_{B(x,\tau)\cap S_X} dy \leq U_X \int_{B(x,\tau)} dy.$$

■

The same type of result can be obtained for $P_X(B(x_1,\tau_1) \cup B(x_2,\tau_2))$ as follows.

Lemma 2 *Under (X1), (X2), and (X3), it holds, for every $(x_1,x_2) \in S_X \times S_X$ and $(\tau_1,\tau_2) \in [0,T]^2$,*

$$\frac{1}{2} M_{1,d}(\tau_1^d + \tau_2^d) \leq P_X(B(x_1,\tau_1) \cup B(x_2,\tau_2)) \leq M_{2,d}(\tau_1^d + \tau_2^d),$$

with $M_{1,d} = cb_X V_d$ and $M_{2,d} = U_X V_d$.

Proof The proof of the upper bound follows from the union bound and Lemma 1. For the lower bound, start noting that for any events A and B, $\mathbb{1}_{A\cup B} \geq (\mathbb{1}_A + \mathbb{1}_B)/2$. Then the conclusion follows from Lemma 1.

■

Based on the previous results, an upper and a lower bound are obtained on the moments of the nearest neighbor radius $\hat{\tau}_{n,k,x}$. A similar upper bound is stated as Lemma 3 in Leluc et al. (2023).

Lemma 3 *Let q be a positive real number. Under (X1), (X2), and (X3), there exist two positive real numbers $c_{q,d}$ and $C_{q,d}$, depending on q, d and on the distribution of X such that for all $x \in S_X$*

$$c_{q,d} \frac{k^{q/d}}{(n+1)^{q/d}} \leq \mathbb{E}\hat{\tau}_{n,k,x}^q \leq C_{q,d} \frac{k^{q/d}}{(n+1)^{q/d}}. \tag{6}$$

A more precise expression of the two constants are

$$c_{q,d} = \frac{M_{2,d}^{-q/d}}{2\Gamma\left(1 + \lfloor q/d \rfloor\right)}, \quad C_{q,d} = 2\Gamma\left(1 + \lfloor q/d \rfloor\right) M_{1,d}^{-q/d},$$

where $\lfloor x \rfloor$ denotes the integer part of the real number x.

Proof We have $\hat{\tau}_{n,k,x} = Z_{(k)}(x)$ the kth-order statistics of $Z_i(x) = |x - X_i|$. Moreover, for any measurable and non-negative function f,

$$\mathbb{E} f\left(Z_{(k)}(x)\right) = \mathbb{E} f \circ F_x^{-1}\left(U_{(k)}\right),$$

where $U_{(k)}$ is the kth-order statistics of a n sample of uniform random variables and since $F_x(z) = \mathbb{P}\left(X_1 \in B(x,z)\right) \in \left[M_{1,d} z^d, M_{2,d} z^d\right]$,

$$F_x^{-1}(u) = \inf\left\{z \in \mathbb{R} : F_x(z) \geq u\right\} \in \left[\frac{u^{1/d}}{M_{2,d}^{1/d}}, \frac{u^{1/d}}{M_{1,d}^{1/d}}\right].$$

Note that the range of $Z_i(x)$ is $[0, \mathrm{diam}(S_X)]$ and we use a constant c in the definition of $M_{1,d} = c b_X V_d$ such that

$$\inf_{x \in S_X} \lambda_d\left(B(x,z)\right) \geq c V_d z^d \text{ for } 0 \leq z \leq \mathrm{diam}(S_X).$$

If $z = u^{1/d}/M_{1,d}^{1/d} \geq \mathrm{diam}(S_X)$, we have $F_x(z) = 1 \geq u$ and we still have $F_x^{-1}(u) \leq z$.

Moreover, if g is measurable and nonnegative,

$$\mathbb{E} g\left(U_{(k)}\right) = n! \int g(u_k) \mathbb{1}_{0 \leq u_1 \leq \cdots \leq u_n \leq 1} du_1 \cdots du_n = \frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n-k+1)} \int_0^1 g(u) u^{k-1} (1-u)^{n-k} du.$$

When $f(z) = z^q$ for some $q > 0$, we get

$$\mathbb{E}\left[Z_{(k)}^q\right] \leq M_{1,d}^{-q/d} \mathbb{E}\left[U_{(k)}^{q/d}\right] = M_{1,d}^{-q/d} \frac{\Gamma(n+1)\Gamma(k+q/d)}{\Gamma(k)\Gamma(n+q/d+1)},$$

$$\mathbb{E}\left[Z_{(k)}^q\right] \geq M_{2,d}^{-q/d} \mathbb{E}\left[U_{(k)}^{q/d}\right] = M_{2,d}^{-q/d} \frac{\Gamma(n+1)\Gamma(k+q/d)}{\Gamma(k)\Gamma(n+q/d+1)}.$$

For $x \geq 1$ and $s > 0$, let $N_{1,s} = \inf_{x \geq 1} \frac{\Gamma(x+s)}{x^s \Gamma(x)}$ and $N_{2,s} = \sup_{x \geq 1} \frac{\Gamma(x+s)}{x^s \Gamma(x)}$. We then get

$$M_{2,d}^{-q/d} \frac{N_{1,q/d}}{N_{2,q/d}} \frac{k^{q/d}}{(n+1)^{q/d}} \leq \mathbb{E}\left[Z_{(k)}(x)^q\right] \leq M_{1,d}^{-q/d} \frac{N_{2,q/d}}{N_{1,q/d}} \frac{k^{q/d}}{(n+1)^{q/d}}.$$

By Wendel's inequality (Wendel, 1948), for $s \in (0,1)$, we have $N_{1,s} \geq \inf_{x \geq 1} \left(\frac{x}{x+s}\right)^{1-s} \geq 1/2$ and $N_{2,s} \leq 1$. For $s \geq 1$, using the equality $\Gamma(z+1) = z\Gamma(z)$, one can deduce that

$$N_{1,s} \geq 1/2, \quad N_{2,s} \leq \Gamma(2 + \lfloor s \rfloor).$$

Indeed if $s = s' + \ell$ with $\ell \in \mathbb{N}$ and $0 \leq s' < 1$,

$$\frac{\Gamma(x+s)}{x^s \Gamma(x)} = \prod_{j=1}^{\ell} \left(1 + \frac{j + s' - 1}{x}\right) \frac{\Gamma(x+s')}{x^{s'} \Gamma(x)}$$

and

$$1 \leq \prod_{j=1}^{\ell} \left(1 + \frac{j + s' - 1}{x}\right) \leq \prod_{j=1}^{\ell} (1+j) = \Gamma(\ell + 2).$$

This completes the proof of Lemma 3. ■

Appendix B. Proofs of the results on the bootstrap sampling error (Section 4.2)

B.1 Proof of Proposition 1

The proof relies on the Lindeberg central limit theorem as given in Proposition 2.27 in Van der Vaart (2000) conditionally to $\mathcal{F}_n$. We need to show the two properties:

$$m^{-1}\sum_{i=1}^m \mathbb{E}[(h(X_i^*, Y_{n,i}^*) - \mathbb{E}[h(X_i^*, Y_{n,i}^*) \mid \mathcal{F}_n])^2 \mid \mathcal{F}_n] \to V,$$
$$m^{-1}\sum_{i=1}^m \mathbb{E}[h(X_i^*, Y_{n,i}^*)^2 \mathbb{1}_{|h(X_i^*, Y_{n,i}^*)|>\epsilon\sqrt{n}} \mid \mathcal{F}_n] \to 0,$$

where each convergence needs to happen with probability 1. Equivalently, using that $(X_i^*, Y_{n,i}^*)_{i=1,\dots,m}$ is identically distributed according to $\hat{Q}$, we need to show that

- $V := \lim_{m\to\infty}\{\hat{Q}(h^2) - \hat{Q}(h)^2\}$ exists; and
- $\hat{Q}(h^2\mathbb{1}_{|h|>\epsilon\sqrt{n}}) \to 0$ for each $\epsilon > 0$.

The first assumption of the proposition ensures that $Q(h^2) < \infty$ implies the existence of $\lim_{m\to\infty}\hat{Q}(h^2)$ and that $Q(h) < \infty$ implies the existence of $\lim_{m\to\infty}\hat{Q}(h)$ and thus of $\lim_{m\to\infty}\hat{Q}(h)^2$. From the assumptions $Q(h) < \infty$ and $Q(h^2) < \infty$, V exists. For the second result, fix $M > 0$. For all n sufficiently large, we have $M \le \epsilon\sqrt{n}$, implying that

$$\hat{Q}(h^2\mathbb{1}_{|h|>\epsilon\sqrt{n}}) \le \hat{Q}(h^2\mathbb{1}_{|h|>M}),$$

which converges to $Q(h^2\mathbb{1}_{|h|>M})$ by assumption. Since $Q(h^2)$ is finite, one can choose M large enough to make $Q(h^2\mathbb{1}_{|h|>M})$ arbitrarily small.

B.2 Proof of Proposition 2

Set $Z_{n,i}^* = h\left(X_i^*, Y_{n,i}^*\right) - \hat{Q}(h)$. We have $\left|Z_{n,i}^*\right| \le 2U_h$ and Var $\left(Z_{n,i}^* \,\middle|\, \mathcal{F}_n\right) = \hat{v}_n$. Note that $\hat{Q}^*(h) - \hat{Q}(h) = \frac{1}{m}\sum_{i=1}^m Z_{n,i}^*$. Bernstein's concentration inequality leads to

$$\mathbb{P}\left(\left|\hat{Q}^*h - \hat{Q}h\right| > u \,\middle|\, \mathcal{F}_n\right) \le 2\exp\left(-\frac{1/2u^2m}{\hat{Q}(h^2) - (\hat{Q}(h))^2 + 2/3U_hu}\right).$$

Then setting

$$\hat{u}_h(\delta) = \frac{4/3U_h}{m}\log(2/\delta) + \sqrt{2\frac{\hat{Q}h^2 - (\hat{Q}h)^2}{m}\log(2/\delta)},$$

we get

$$\mathbb{P}\left(\left|\hat{Q}^*h - \hat{Q}h\right| > \hat{u}_h(\delta) \,\middle|\, \mathcal{F}_n\right) \le \delta$$

and then integrate both sides to obtain

$$\mathbb{P}\left(\left|\hat{Q}^*h - \hat{Q}h\right| > \hat{u}_h(\delta)\right) \le \delta,$$

which leads to the stated bound.

Appendix C. Proofs of the results on the k-NN nonparametric error (Section 4.3)

Here, we give proofs of the results on the k-NN nonparametric error appearing in Section 4.3.

C.1 Proof of Proposition 3

We start with a useful bias-variance decomposition. Introduce

$$\begin{aligned} \epsilon_i(x) &= h(x, Y_i) - \int h(x,y) P_{Y\,|\,X}(dy \mid X_i), \\ \Delta(x, X_i) &= \int h(x,y)(P_{Y\,|\,X}(dy \mid X_i) - P_{Y\,|\,X}(dy \mid x)). \end{aligned}$$

We have

$$\begin{aligned} &\int h(x,y)(\hat{P}_{Y\,|\,X}(dy \mid x) - P_{Y\,|\,X}(dy \mid x)) \\ &= k^{-1} \sum_{i=1}^{n} \epsilon_i(x) 1_{B(x, \hat{\tau}_{n,k,x})}(X_i) + k^{-1} \sum_{i=1}^{n} \Delta(x, X_i) 1_{B(x, \hat{\tau}_{n,k,x})}(X_i). \end{aligned}$$

Integrating with respect to $Q_X(dx)$, we obtain

$$(\hat{Q} - Q)(h) = B_h + S_h \tag{7}$$

with

$$\begin{aligned} B_h &= k^{-1} \sum_{i=1}^{n} \int \Delta(x, X_i) 1_{B(x, \hat{\tau}_{n,k,x})}(X_i) Q_X(dx), \\ S_h &= k^{-1} \sum_{i=1}^{n} \int \epsilon_i(x) 1_{B(x, \hat{\tau}_{n,k,x})}(X_i) Q_X(dx). \end{aligned}$$

The term B_h is a bias term and the term S_h (which has mean 0) is a variance term.

The proof is divided into 3 steps. The first step takes care of bounding the bias term. The second step deals with the variance upper-bound. The third step is concerned with the variance lower bound.

The bias. First (H1) gives that for any $X_i \in S_X \cap B(x, \hat{\tau}_{n,k,x})$ and $x \in S_X$,

$$|\Delta(x, X_i)| \leq g_h(x) \hat{\tau}_{n,k,x}.$$

Consequently, using (X4) and the fact that $\sum_{i=1}^{n} \mathbb{1}_{\|X_i - x\| \leq \hat{\tau}_{n,k,x}} = k$, we have

$$\left| k^{-1} \sum_{i=1}^{n} \int \Delta(x, X_i) 1_{\{B(x, \hat{\tau}_{n,k,x})\}}(X_i) Q_X(dx) \right| \leq \int \hat{\tau}_{n,k,x} g_h(x) Q_X(dx)$$

and from Jensen inequality

$$\left| k^{-1} \sum_{i=1}^{n} \int \Delta(x, X_i) 1_{\{B(x,\hat{\tau}_{n,k,x})\}}(X_i) Q_X(dx) \right|^2 \leq \int \hat{\tau}_{n,k,x}^2 g_h^2(x) Q_X(dx)$$

From Lemma 3, we have $\sup_{x \in S_X} \mathbb{E}[\hat{\tau}_{n,k,x}^2] \leq C_{2,d} k^{2/d} (n+1)^{-2/d}$ and the control of the bias is given by

$$\mathbb{E}\left[\left| B_h^2 \right| \right] \leq C_{2,d} \int g_h^2(x) Q_X(dx) \cdot \frac{k^{2/d}}{(n+1)^{2/d}}.$$

The variance upper-bound. For the proof, we assume that $1 \leq k < n/2$. We have for each $(x, x') \in S_X^2$ and $(i,j) \in \{1, \ldots, n\}^2$,

$$\mathbb{E}[\epsilon_i(x)\epsilon_j(x') \mid X_1, \ldots, X_n] = \begin{cases} 0 & \text{if } i \neq j, \\ \mathbb{E}\left[\epsilon_i(x)\epsilon_i(x') \mid X_i\right] \leq \sqrt{\mathbb{E}\left[\epsilon_i(x)^2 \mid X_i\right] \mathbb{E}\left[\epsilon_i(x')^2 \mid X_i\right]} \leq \sigma_+^2 & \text{if } i \neq j. \end{cases}$$

For the second case, we used (H2) and the Cauchy-Schwarz inequality. As a consequence, the variance is given by

$$\begin{aligned}
\mathbb{E}\left[S_h^2\right] &= \mathbb{E}\left[\left(k^{-1} \sum_{i=1}^{n} \int \mathbb{1}_{\|X_i - x\| \leq \hat{\tau}_{n,k,x}} \epsilon_i(x) Q_X(dx) \right)^2 \right] \\
&= k^{-2} \sum_{i,j}^{n} \mathbb{E}\left[\left(\int \mathbb{1}_{\|X_i - x\| \leq \hat{\tau}_{n,k,x}} \epsilon_i(x) Q_X(dx) \right) \left(\int \mathbb{1}_{\|X_j - x'\| \leq \hat{\tau}_{n,k,x'}} \epsilon_j(x') Q_X(dx') \right) \right] \\
&= k^{-2} \sum_{i,j}^{n} \mathbb{E}\left[\left(\int \int \mathbb{1}_{\|X_i - x\| \leq \hat{\tau}_{n,k,x}} \mathbb{1}_{\|X_j - x'\| \leq \hat{\tau}_{n,k,x'}} \mathbb{E}\left[\epsilon_i(x)\epsilon_j(x') \mid X_1, \ldots, X_n\right] Q_X(dx) Q_X(dx') \right) \right] \\
&\leq k^{-2} \sigma_+^2 \sum_{i=1}^{n} \mathbb{E}\left[\int \int \mathbb{1}_{\|X_i - x\| \leq \hat{\tau}_{n,k,x}} \mathbb{1}_{\|X_i - x'\| \leq \hat{\tau}_{n,k,x'}} Q_X(dx) Q_X(dx') \right] \\
&= k^{-2} \sigma_+^2 \sum_{i=1}^{n} \mathbb{E}\left[\hat{Y}_i^2 \right]
\end{aligned}$$

with $\hat{Y}_i = \int \mathbb{1}_{\|X_i - x\| \leq \hat{\tau}_{n,k,x}} Q_X(dx)$. Let $Z_i(x) = \|x - X_i\|$ and $Z_{(k)}(x)$, the kth order statistics of the sample $(Z_i(x))_{1 \leq i \leq n}$. One can observe that

$$Z_i(x) \leq Z_{(k)}(x) \iff Z_i(x) < Z_{(k)}^{-i}(x),$$

where $Z_{(k)}^{-i}(x)$ is the kth order statistics of the sample $(Z_j(x))_{1 \leq j \neq i \leq n}$. Note that the two sigma fields generated respectively by $\{Z_i(x) : x \in S_X\}$ and $\left\{ Z_{(k)}^{-i}(x) : x \in S_X \right\}$ are independent. For one mapping $\rho : S_X \to \operatorname{diam}(S_X)$, we first bound

$$\begin{aligned}
I &= \mathbb{E}\left|\int \mathbb{1}_{Z_1(x)\le\rho(x)} Q_X(dx)\right|^2 \\
&= \int\int \mathbb{P}\left(Z_1(x)\le\rho(x), Z_1(y)\le\rho(y)\right) Q_X(dx)Q_X(dy) \\
&= 2\int\int \mathbb{1}\{\rho(x)\le\rho(y)\}\mathbb{P}\left(Z_1(x)\le\rho(x), Z_1(y)\le\rho(y)\right) Q_X(dx)Q_X(dy) \\
&= 2\int\int \mathbb{1}\{\rho(x)\le\rho(y), \|x-y\|\le 2\rho(y)\}\mathbb{P}\left(Z_1(x)\le\rho(x), Z_1(y)\le\rho(y)\right) Q_X(dx)Q_X(dy) \\
&\le 2M_{2,d}\int\int \mathbb{1}\{\rho(x)\le\rho(y), \|x-y\|\le 2\rho(y)\}\rho(y)^d Q_X(dx)Q_X(dy) \\
&\le 2M_{2,d}\int Q_X\left(B(y,2\rho(y))\right)\rho(y)^d Q_X(dy) \\
&\le 2^{d+1}M_{2,d}\int C(y)\rho(y)^{2d} Q_X(dy)
\end{aligned}$$

The fourth inequality is due to the fact that when $\|x-y\| > 2\rho(y)$, the two balls $B(x,\rho(x))$ and $B(y,\rho(y))$ do not intersect. The last inequality is a consequence of Assumption (X4). We then get using Lemma 3,

$$E\hat{Y}_i^2 \le 2^{d+1}M_{2,d}\int C(y)\mathbb{E}\left\{Z_{(k)}^{-i}(y)\right\}^{2d} Q_X(dy) \le \frac{2^{d+3}M_{2,d}\int C(y)Q_X(dy)}{M_{1,d}^2}\frac{k^2}{n^2}.$$

This leads to the variance upper-bound.

The variance lower-bound. Here we assume the $\sigma_-^2 := \inf_{x\in S_X}\operatorname{Var}\left(h(Y)|X=x\right) > 0$. From previous computations, we have

$$\operatorname{Var}\left(\hat{Q}h\right) \ge \frac{\sigma_-^2}{k^2}\sum_{i=1}^n \mathbb{E}\left[\hat{Y}_i^2\right].$$

We have $\mathbb{E}\hat{Y}_i^2 \ge \mathbb{E}^2\hat{Y}_i$ and we have to find a lower bound for $\mathbb{E}\hat{Y}_i$. But from the arguments used in the proof for the upper-bound, we have

$$\mathbb{E}\hat{Y}_i = \int_{S_X}\mathbb{P}\left(Z_i(x)\le Z_{(k)}^{-i}(x)\right)Q_X(dx) \ge M_{1,d}\int_{S_X}\mathbb{E}\left|Z_{(k)}^{-i}(x)\right|^d Q_X(dx) \ge \frac{k}{2nM_{2,d}},$$

where the last inequality follows from Lemma 3. This shows the lower bound and the proof of Proposition 3 is now complete.

C.2 Proof of Proposition 4

Define

$$\overline{\tau}_{n,k} = \left(\frac{2k}{nM_{1,d}}\right)^{1/d}.$$

The following Lemma (Portier, 2025, Lemma 3) controls the size of the k-NN balls uniformly over all $x \in S_X$.

Lemma 4 (Portier (2025, Lemma 3)) *Suppose that (X1) (X2) and (X3) hold true. Then for all* $n \geq 1$, $\delta \in (0,1)$ *and* $1 \leq k \leq n$ *such that* $16d \log(12n/\delta) \leq k \leq T^d n b_X c V_d/2$ *, it holds, with probability at least* $1-\delta$*:*

$$\sup_{x \in S_X} \hat{\tau}_{n,k,x} \leq \overline{\tau}_{n,k}.$$

We now deal with the variance term of our estimator. The variance term of the nearest-neighbors estimator is given by $V_n = k^{-1} \sum_{i=1}^n \hat{Y}_i$, where

$$\hat{Y}_i = \int \epsilon_i(x) \mathbb{1}_{B(x, \hat{\tau}_{n,k,x})}(X_i) Q_X(dx).$$

and Set $\hat{\tau}_k = \sup_{x \in S_X} \hat{\tau}_{n,k,x}$. From our assumptions and Jensen's inequality, we have

$$|\hat{Y}_i| \leq 2U_h M_{2,d} \hat{\tau}_k, \quad \operatorname{Var}\left(\hat{Y}_i | X_1, \ldots, X_n\right) \leq \sigma_+^2 M_{2,d}^2 \hat{\tau}_k{}^{2d}.$$

Applying Bernstein's inequality for i.i.d. random variables (we recall that the Y_i's are independent conditionally on the X_i's), we get for $t > 0$,

$$\mathbb{P}\left(|V_n| > t | X_1, \ldots, X_n\right) \leq 2 \exp\left(-\frac{\frac{1}{2} k^2 t^2}{n \sigma_+^2 M_{2,d}^2 \hat{\tau}_k^{2d} + \frac{2}{3} U_h M_{2,d} \hat{\tau}_k^d k}\right).$$

This leads to

$$\mathbb{P}\left(|V_n| > t, \hat{\tau}_k \leq \overline{\tau}_{n,k} | X_1, \ldots, X_n\right) \leq 2 \exp\left(-\frac{\frac{1}{2} t^2 n}{L_1 \sigma_+^2 + L_2 t}\right).$$

Note that this upper-bound is not random and we get

$$\mathbb{P}\left(|V_n| > t, \hat{\tau}_k \leq \overline{\tau}_{n,k}\right) \leq 2 \exp\left(-\frac{\frac{1}{2} t^2 n}{L_1 \sigma_+^2 + L_2 t}\right). \tag{8}$$

Setting

$$\widetilde{t}_n(\delta, h) = \frac{L_2}{n} \log(2/\delta) + \sqrt{\frac{L_2^2}{n^2} \log^2(2/\delta) + \frac{2 L_1 \sigma_+^2}{n} \log(2/\delta)},$$

which is smaller than

$$t_n(\delta, h) = L_0 \left(\frac{k}{n}\right)^{1/d} + \frac{2L_2}{n} \log(2/\delta) + \sqrt{\frac{2 L_1 \sigma_+^2}{n} \log(2/\delta)}$$

we then get

$$\begin{aligned} \mathbb{P}\left(|V_n| > t_n(\delta, h)\right) &\leq \mathbb{P}\left(|V_n| > \widetilde{t}_n(\delta, h), \hat{\tau}_k \leq \overline{\tau}_{n,k}\right) + P\left(\hat{\tau}_k > \overline{\tau}_{n,k}\right) \\ &\leq 2\delta, \end{aligned}$$

where the last inequality is a consequence of (8) and Lemma 4. Moreover, from the proof of Proposition 3 and Lemma 4, the bias part can be dominated by $L_0(k/n)^{1/d}$ with probability at least $1-\delta$. This concludes the proof.

Appendix D. Proof of Theorem 1

First, note that the boundedness of $\sup_{x\in S_X}\mathbb{E}\left[h^2(x,Y)|X\right]$ entails **H2**. Setting $A_{m,n}=\hat{Q}^*(h)-\hat{Q}(h)$ and $B_n=\hat{Q}(h)-Q(h)$, Proposition 3 guarantees that

$$\mathbb{E}B_n^2\leq C_1\left\{\frac{1}{n^{2/d}}+\frac{1}{n}\right\},$$

for some $C_1>0$ only depending on the distribution of (X,Y), X^* and on h. It remains to show that the same bound can obtained for $A_{m,n}$. Since,

$$\mathbb{E}\left[A_{m,n}^2|\mathcal{F}_n\right]\leq\frac{\hat{Q}(h^2)-\hat{Q}(h)^2}{m}\leq\frac{\hat{Q}(h^2)}{m}.$$

It only remains to show that $\mathbb{E}\hat{Q}(h^2)$ is bounded with respect to n. The approach is similar to the control of the variance term for $k=1$ studied in the proof of Proposition 3. If L is an upper-bound for $\sup_{x\in S_X}\mathbb{E}\left[h^2(x,Y)|X\right]$, we have

$$\begin{aligned}\mathbb{E}\hat{Q}(h^2) &= \sum_{i=1}^n\int_{S_X}Q_X(dx)\mathbb{E}\left[h^2(x,Y_1)\mathbb{1}_{\|X_1-x\|\leq\min_{2\leq j\leq n}\|X_j-x\|}\right]\\ &\leq nL\int_{S_X}\mathbb{P}\left(\|X_1-x\|\leq\min_{2\leq j\leq n}\|X_j-x\|\right)Q_X(dx)\\ &\leq nLM_{2,d}\int_{S_X}\mathbb{E}\min_{2\leq j\leq n}\|X_j-x\|^dQ_X(dx)\\ &\leq \frac{2LM_{2,d}}{M_{1,d}}.\end{aligned}$$

The last upper bound is obtained from Lemma 3 using the fact that $\min_{2\leq j\leq n}\|X_j-x\|$ has the same probability distribution as $\hat{\tau}_{n-1,1,x}$. We deduce the result taking C as the maximum between C_1 and $2LM_{2,d}/M_{1,d}$. ∎

Appendix E. A corollary bounding the bootstrap sampling error when using k-NN sampling

Corollary 1 *Suppose that (X1), (X2), (X3) are fulfilled and assume that Q_X has a density q_X bounded by U_X with support included in S_X. Suppose also that (H1) and (H2) are fulfilled for both h and h^2 and h is bounded. Let $k=k_n$ satisfying the condition of Lemma 4 and $\delta\in(0,1/7)$ and set $s^2(h)=Qh^2-(Qh)^2$ and $\hat{s}^2(h)=\hat{Q}h^2-(\hat{Q}h)^2$. Let $\delta\in(0,1/7)$. Then with probability greater than $1-7\delta$,*

$$\hat{Q}^*h-\hat{Q}h\leq\frac{4/3U_h}{m}\log(2/\delta)+\sqrt{2\frac{v_n^2(\delta,h)}{m}\log(2/\delta)}+\sqrt{2\frac{s^2(h)}{m}\log(2/\delta)},\qquad(9)$$

where

$$v_n^2(\delta,h)=t_n(\delta,h^2)+t_n(\delta,h)^2+2t_n(\delta,h)Q|h|$$

and $t_n(\delta,h)$ is defined in the statement of Proposition 4.

Proof We first use the result of Proposition 2. In particular, setting

$$\hat{u}_h(\delta) = \frac{4/3U_h}{m}\log(2/\delta) + \sqrt{2\frac{\hat{Q}h^2 - (\hat{Q}h)^2}{m}\log(2/\delta)},$$

we have

$$\mathbb{P}\left(\left|\hat{Q}^*h - \hat{Q}h\right| > \hat{u}_h(\delta)\right) \leq \delta,$$

We then use the decomposition

$$\hat{Q}h^2 - (\hat{Q}h)^2 = s^2(h) + \hat{Q}h^2 - Qh^2 - \left(\hat{Q}h - Qh\right)^2 - 2Qh\left(\hat{Q}h - Qh\right).$$

From Proposition 4, we know that

$$\left|\hat{Q}h^2 - Qh^2\right| \leq t_n(\delta, h^2)$$

with probability greater than $1 - 3\delta$ and

$$\left|\hat{Q}h - Qh\right| \leq t_n(\delta, h)$$

with probability greater than $1 - 3\delta$. Collecting these three bounds, we easily obtain the conclusion of the second point of Corollary 1. ■

Appendix F. Proofs of the results on the empirical risk minimization (Section 5)

Here, we present proofs of the result on the application to empirical risk minimization.

F.1 Proof of Proposition 5.

From (A1), $\sup_{\theta\in\Theta}|m_\theta(Y_1^*, X_1^*)|$ is integrable and $\theta \mapsto \mathcal{R}^*(\theta)$ is continuous over the compact set Θ. As a consequence, weak consistency will follow from Theorem 5.7 in Van der Vaart (2000) if we show that

$$\sup_{\theta\in\Theta}\left|\mathcal{R}^*_{m,n}(\theta) - \mathcal{R}^*(\theta)\right| = o_{\mathbb{P}}(1). \tag{10}$$

Pointwise convergence holds true from assumptions (A1), (A2) as each mapping m_θ satisfies Assumptions (H1), (H2). One can then apply Theorem 1 and the Markov inequality to get for any $\theta \in \Theta$,

$$\mathcal{R}^*_{m,n}(\theta) - \mathcal{R}^*(\theta) = o_{\mathbb{P}}(1), \quad m, n \to \infty.$$

We now prove uniform convergence. Let $\delta > 0$. One can cover the compact set Θ with finitely many balls $B(\theta_i, \delta)$, $1 \leq i \leq k$. For $\theta \in \Theta \cap B(\theta_i, \delta)$, we have

$$\left|\mathcal{R}^*_{m,n}(\theta) - \mathcal{R}^*_{m,n}(\theta_i)\right| \leq \eta(\delta)\frac{1}{m}\sum_{i=1}^{m} h\left(Y^*_{n,i}\right).$$

Moreover, from Assumptions (A2) and Theorem 1 with the Markov inequality, we know that

$$\frac{1}{m}\sum_{i=1}^{m} h\left(Y_{n,i}^{*}\right) \to \mathbb{E}h\left(Y_1^{*}\right), \quad m,n\to\infty,$$

in probability. We also have

$$|\mathcal{R}^*(\theta)-\mathcal{R}^*(\theta_i)| \leq \eta(\delta)\mathbb{E}h\left(Y_1^*\right).$$

Finally, one can use the bound

$$\begin{aligned} \sup_{\theta\in\theta}\left|\mathcal{R}_{m,n}^*(\theta)-\mathcal{R}^*(\theta)\right| &\leq \max_{1\leq i\leq k}\left|\mathcal{R}_{m,n}^*(\theta_i)-\mathcal{R}^*(\theta_i)\right|+\eta(\delta)\left\{\frac{1}{m}\sum_{i=1}^{m}h\left(Y_{n,i}^*\right)+\mathbb{E}h\left(Y_1^*\right)\right\} \\ &= 2\eta(\delta)\mathbb{E}h\left(Y_1^*\right)+o_{\mathbb{P}}(1). \end{aligned}$$

Given that δ is arbitrary, the above implies (10) and the weak consistency of $\hat{\theta}^*$ follows. The second assertion about the excess risk then follows easily using that

$$\mathcal{R}^*(\hat{\theta}^*)-\mathcal{R}^*(\theta^*) \leq 2\sup_{\theta\in\Theta}\left|\mathcal{R}_{m,n}^*(\theta)-\mathcal{R}^*(\theta)\right|.$$

■

F.2 Proof of Proposition 6.

Let $Z_i^* = \Gamma^{-1/2}X_i^*$ and define

$$\Sigma_m = \frac{1}{m}\sum_{i=1}^{m} Z_i^* Z_i^{*T}, \quad N_m = \frac{1}{m}\sum_{i=1}^{m} Z_i^*\left[Y_{n,i}^* - X_i^{*T}\theta^*\right].$$

The proof first requires some analysis of the smallest eigenvalues of Σ_m. From the matrix Chernoff inequality given in Tropp (2012), see Corollary 5.2 and Remark 5.3, we have

$$P(\lambda_{\min}(\Sigma_m)\leq 1-\eta) \leq d\exp(-\eta^2 m/2B)$$

where $B=\lambda_{\min}(\Gamma)^{-1}d\max_{x\in S_X}|x|_\infty^2$ is defined so as to satisfy $\|Z\|_2^2\leq\lambda_{\min}(\Gamma)^{-1}\|X\|_2^2\leq B$, with probability 1. Inverting the previous we obtain that with probability at least $1-\delta$,

$$\lambda_{\min}(\Sigma_m) > 1-\sqrt{(2B/m\log(d\delta^{-1}))}$$

and therefore as soon as $(8B/n)\log(d\delta^{-1})\leq 1$, we have that $\lambda_{\min}(\Sigma_m)\geq 1/2$. On the previous event, we have that

$$\Gamma^{1/2}(\hat{\theta}^*-\theta^*) = \Sigma_m^{-1}N_m$$

It follows that

$$\mathcal{R}^*(\hat{\theta}^*)-\mathcal{R}^*(\theta^*) = \|\Gamma^{1/2}(\hat{\theta}^*-\theta^*)\|_2^2 = \|\Sigma_m^{-1}N_m\|_2^2 \leq 2\|N_m\|_2^2.$$

We conclude using Theorem 1 with $h(y,x)=\Gamma^{-1/2}x\left(y-x^T\theta^*\right)$. Note that $\mathbb{E}\left[h\left(Y^*,X^*\right)\right]=0$ by definition of θ^*. ■

Appendix G. An unbounded density satisfying Assumption (X4)

Suppose that $q_X(y) = (1-\alpha)y^{-\alpha}$ for every $y \in S_X = (0,1)$, for some $\alpha \in (0,1/2)$. We will show that this example satisfies Assumption (X4) with $C(x) = C_\alpha x^{-\alpha}$ for some $C_\alpha > 0$. It is only necessary to bound $Q_X(B(x,\tau))$ for some $\tau_0 \in (0,1)$. Indeed, for $\tau > \tau_0$, one can use the inequalities

$$Q_X(B(x,\tau)) \le 1 \le \tau/\tau_0 \le x^{-\alpha}\tau/\tau_0.$$

We choose $\tau_0 = 1/6$. We have

$$\begin{aligned} Q_X(B(x,\tau)) &= \int_{(0,1)\cap[x-\tau,x+\tau]} (1-\alpha)y^{-\alpha}dy \\ &\le \min(x+\tau,1)^{1-\alpha} - \max(x-\tau,0)^{1-\alpha}. \end{aligned}$$

We consider three cases.

- If $0 < x \le 2\tau$, we have

$$Q_X(B(x,\tau)) \le (x+\tau)^{1-\alpha} \le 3^{1-\alpha}2^\alpha x^{-\alpha}\tau.$$

- If now $1 - 2\tau \le x < 1$, using the mean value theorem, we have

$$Q_X(B(x,\tau)) \le 1 - (1-3\tau)^{1-\alpha} \le 3(1-\alpha)2^\alpha\tau \le 3(1-\alpha)2^\alpha x^{-\alpha}\tau.$$

- Finally, if $2\tau \le x \le 1-2\tau$, we get

$$Q_X(B(x,\tau)) \le \frac{1-\alpha}{(x-\tau)^\alpha}2\tau \le \frac{(1-\alpha)2^{\alpha+1}}{x^\alpha}\tau.$$

We deduce that $C(x) = C_\alpha x^{-\alpha}$ for some $C_\alpha > 0$ not depending on $x \in S_X$. Assumption (X4) is then satisfied.

Appendix H. Illustration for Experiments E1–E3

Illustrations of data used in Experiments E1–E3 can be found in Figure 7.

Appendix I. Details of the benchmark data experiments

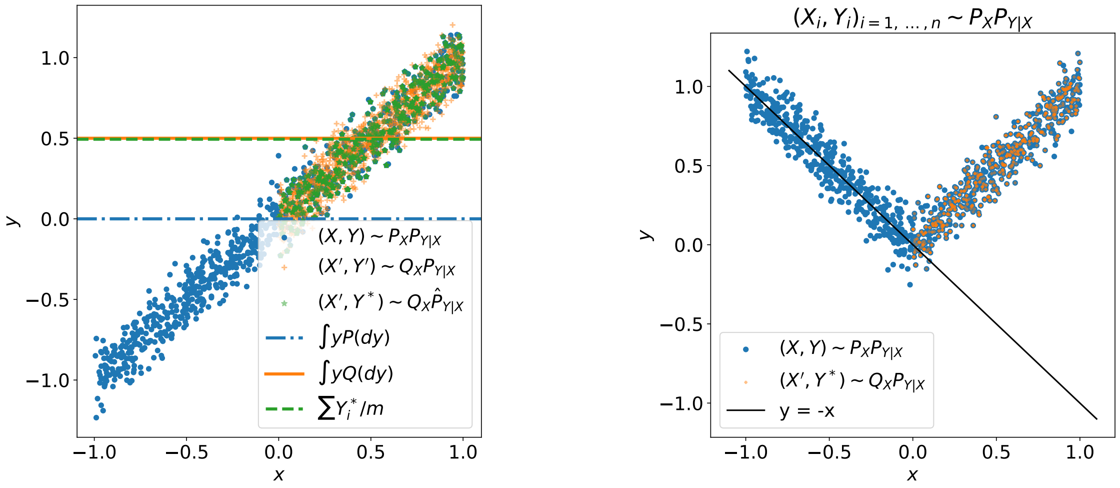
We use the following datasets.

- `california`: Regression dataset called "California Housing" available from `https://www.dcc.fc.up.pt/~ltorgo/Regression/cal_housing.html`.

- `diabetes`: Regression dataset available from `https://archive.ics.uci.edu/ml/index.php` (Dua and Graff, 2017).

- `breast cancer`: Classification dataset available from `https://archive.ics.uci.edu/ml/index.php` (Dua and Graff, 2017).

- `twonorm`: Classification dataset available from `https://www.cs.utoronto.ca/~delve/data/datasets.html`.



(a) Data for Experiment E1

(b) Data for Experiment E2 and E3

Figure 7: Visualization of data for Experiments E1–E3

Data splitting and sampling bias simulation We split the original to the training and test set and simulate covariate shift by rejection sampling from the test set with rejection probability determined according to the value of a covariate. For `california`, `twonorm`, `breast cancer`, we follow the procedure of Sugiyama et al. (2007): we include each target data point X_i to the target set with probability $\min(1, 4X_{i,c}^2)$ or reject it otherwise, where $X_{i,c}$ is the c-th attribute of X_i. For `diabetes`, we used a different biasing procedure for this data set because the technique of Sugiyama et al. (2007) rejects too many data points to perform our experiment for this dataset. We instead use the procedure of an example from the ADAPT package de Mathelin et al. (2021)[7] for `diabetes`: for each data point X_i, we accept it with probability proportional to $\exp(-20 \times |X_{i,\text{age}} + 0.06|)$, where $X_{i,\texttt{age}}$ is the `age` attribute of X_i and reject (i.e., exclude) otherwise.

Pre-processing We use the hot-encoding for all categorical features. We center and normalize all the data using the mean and the dimension-wise standard deviation of the source set. We do the same centering and normalization for the output variables for regression datasets.

After training and prediction, we post-process the output using the inverse operation. Table 3 shows basic information about the datasets after the bias-sampling and pre-processing.

[7] `https://adapt-python.github.io/adapt/examples/Sample_bias_example.html`

Table 3: Basic information of the datasets

	california	twonorm	diabetes	breast cancer
Input dimension d	8	20	10	9
source sample size n	1000	100	150	200
Target sample size m	1000	500	150	100