Reweighting samples under covariate shift using a Wasserstein distance criterion

Julien Reygner, Adrien Touboul

Reweighting samples under covariate shift using a Wasserstein distance criterion

Julien Reygner and Adrien Touboul

ABSTRACT. Considering two random variables with different laws to which we only have access through finite size iid samples, we address how to reweight the first sample so that its empirical distribution converges towards the true law of the second sample as the size of both samples goes to infinity. We study an optimal reweighting that minimizes the Wasserstein distance between the empirical measures of the two samples, and leads to an expression of the weights in terms of Nearest Neighbors. The consistency and some asymptotic convergence rates in terms of expected Wasserstein distance are derived, and do not need the assumption of absolute continuity of one random variable with respect to the other. These results have some application in Uncertainty Quantification for decoupled estimation and in the bound of the generalization error for the Nearest Neighbor Regression under covariate shift.

1. Introduction

1.1. Covariate shift in UQ.

A common task in Uncertainty Quantification (UQ) for Computer Experiments [6, 8] is the evaluation of a quantity of interest QI of the form

$$\mathrm{QI} = \mathbb{E}[\phi(Y)],$$

where $Y \in \mathbb{R}^e$ is a random vector which is typically the output of a numerical simulation with uncertain inputs and parameters, and $\phi : \mathbb{R}^e \to \mathbb{R}$ is the *observable*. Generically, the random vector Y writes

$$Y = f(X, \Theta),$$

where $X \in \mathbb{R}^d$ represents the inputs of the numerical simulation, Θ is the set of parameters of this simulation (which takes its values in some measurable space $\mathbf{\Theta}$), and $f : \mathbb{R}^d \times \mathbf{\Theta} \to \mathbb{R}^e$ is the *numerical model*, which is the function actually evaluated by the computer code. We denote by μ_X and μ_Θ the respective probability distributions of X and Θ and assume that these variables are independent. Virtually, if one is able to sample iid realizations $(X_1, \Theta_1), \dots, (X_n, \Theta_n)$ from $\mu_X \otimes \mu_\Theta$, then QI can be estimated by the direct Monte Carlo estimator

$$\widehat{\mathrm{QI}}_n := \frac{1}{n} \sum_{i=1}^{n} f(X_i, \Theta_i).$$

The present work is motivated by the study of UQ in complex engineering systems, where

- the input X can be itself the output of possibly several other "upstream" numerical simulations,
- each evaluation of the function f is costly.

This research work has been carried out under the leadership of the Technological Research Institute SystemX, and therefore granted with public funds within the scope of the French Program "Investissements d'Avenir".

When X is modeled by a deterministic variable, this problem can be treated by the so-called Collaborative Optimization methods [4, 20] in Multidisciplinary Analysis and Optimization. When X is a random variable, the implementation of the direct Monte Carlo method is impossible because, in practice, the law μ_X is unknown and one cannot wait for a sample $X_1, \dots, X_n$ to be generated by the upstream numerical simulations before starting running one's own simulation. In contrast, we however assume that μ_Θ is known and that one is able to sample iid realizations $\Theta_1, \dots, \Theta_n$ from this distribution. This naturally leads one to generate a *synthetic* sample $X'_1, \dots, X'_{n_{\mathrm{off}}}$ according to some user-chosen probability measure $\mu_{X'}$ on $\mathbb{R}^d$, and evaluate the numerical model f on the sample $(X'_1, \Theta_1), \dots, (X'_{n_{\mathrm{off}}}, \Theta_{n_{\mathrm{off}}})$ to obtain a corresponding set of realizations $Y'_1, \dots, Y'_{n_{\mathrm{off}}}$ during some *offline* phase. Once actual realizations $X_1, \dots, X_{n_{\mathrm{on}}}$ become available in a subsequent *online* phase, they have to be used in combination with the synthetic sample to construct an estimator of QI, but evaluations of the numerical model f are no longer allowed.

The assumption that the sequence $\Theta_1, \dots, \Theta_{n_{\mathrm{off}}}$ be independent from $X'_1, \dots, X'_{n_{\mathrm{off}}}$ then ensures that for all $x \in \mathbb{R}^d$,

$$\mathrm{Law}(Y'|X' = x) = \mathrm{Law}(f(x, \Theta)) = \mathrm{Law}(Y|X = x).$$

This situation is known in the statistical learning literature as a *covariate shift* [4], [15, Section 1.4].

1.2. **Density ratio estimation.** Inspired by the *importance sampling* technique, an intuitive approach to estimate QI from the synthetic sample $\{(X'_j, \Theta_j; Y'_j), 1 \leq j \leq n_{\mathrm{off}}\}$ consists in writing

$$\begin{aligned} \mathrm{QI} &= \int_{\mathbb{R}^d \times \boldsymbol{\Theta}} \phi(f(x,\theta)) \mathrm{d}\mu_X(x) \mathrm{d}\mu_\Theta(\theta) \\ &= \int_{\mathbb{R}^d \times \boldsymbol{\Theta}} \phi(f(x',\theta)) \frac{\mathrm{d}\mu_X}{\mathrm{d}\mu_{X'}}(x') \mathrm{d}\mu_{X'}(x') \mathrm{d}\mu_\Theta(\theta), \end{aligned}$$

so that assuming that μ_X is absolutely continuous with respect to $\mu_{X'}$, an unbiased and consistent (in the $n_{\mathrm{off}} \to +\infty$ limit) estimator of QI is given by

$$\frac{1}{n_{\mathrm{off}}} \sum_{j=1}^{n_{\mathrm{off}}} \rho_{X,X'}(X'_j) \phi(Y'_j), \qquad \rho_{X,X'}(x') := \frac{\mathrm{d}\mu_X}{\mathrm{d}\mu_{X'}}(x').$$

Of course, the Radon–Nikodym derivative $\rho_{X,X'}$ is actually not known in this situation, and it has to be estimated in the online phase thanks to the sample $X_1, \dots, X_{n_{\mathrm{on}}}$. Observe that this problem no longer involves neither $\Theta_1, \dots, \Theta_{n_{\mathrm{off}}}$ nor $Y'_1, \dots, Y'_{n_{\mathrm{off}}}$.

The theoretical issue of estimating the Radon–Nikodym derivative $\rho_{X,X'}$ from independent samples $\mathbf{X}_{n_{\mathrm{on}}} := (X_1, \dots, X_{n_{\mathrm{on}}})$ and $\mathbf{X}'_{n_{\mathrm{off}}} := (X'_1, \dots, X'_{n_{\mathrm{off}}})$ is known in the statistical learning literature as *density ratio estimation* [16]. A rather generic procedure consists in fixing some distance-like function d on the set of probability measures on $\mathbb{R}^d$, writing

$$\rho_{X,X'} = \arg\min_\rho \mathsf{d}\left(\rho \mu_{X'}, \mu_X\right),$$

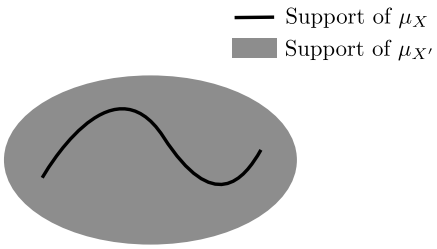


FIGURE 1. Example in which μ_X is not absolutely continuous with respect to $\mu_{X'}$ but its support is included in the support of $\mu_{X'}$.

and estimating $\rho_{X,X'}$ by

$$\widehat{\rho}_{\mathbf{X}_{n_{\mathrm{on}}},\mathbf{X}'_{n_{\mathrm{off}}}} := \arg\min_{\rho} \mathsf{d}\left(\rho\widehat{\mu}_{\mathbf{X}'_{n_{\mathrm{off}}}}, \widehat{\mu}_{\mathbf{X}_{n_{\mathrm{on}}}}\right),$$

with the empirical measures

$$\widehat{\mu}_{\mathbf{X}_{n_{\mathrm{on}}}} := \frac{1}{n_{\mathrm{on}}}\sum_{i=1}^{n_{\mathrm{on}}} \delta_{X_i}, \qquad \widehat{\mu}_{\mathbf{X}'_{n_{\mathrm{off}}}} := \frac{1}{n_{\mathrm{off}}}\sum_{j=1}^{n_{\mathrm{off}}} \delta_{X'_j}.$$

Since the quantity which is minimized only depends on ρ through the measure $\rho\widehat{\mu}_{\mathbf{X}'_{n_{\mathrm{off}}}}$, and thus through the values $\rho(X'_1), \ldots, \rho(X'_{n_{\mathrm{off}}})$, the actual output is a vector of *weights* $\widehat{w}_{n_{\mathrm{off}}} := (\widehat{w}_1, \ldots, \widehat{w}_{n_{\mathrm{off}}})$ which approximate the values of $\rho_{X,X'}$ at the points $X'_1, \ldots, X'_{n_{\mathrm{off}}}$, and therefore yield the estimator

$$\widehat{\mathrm{QI}}_{n_{\mathrm{off}},n_{\mathrm{on}}} := \frac{1}{n_{\mathrm{off}}}\sum_{j=1}^{n_{\mathrm{off}}} \widehat{w}_j \phi(Y'_j) \tag{1}$$

of QI. This approach has been applied with several choices of distance-like functions d, such as moment/kernel matching, L^2 distance, Kullback–Leibler divergences; we refer to [16] for an extensive review supplemented with a detailed list of references. Since the primary purpose of these methods is the approximation of the density ratio $\rho_{X,X'}$, the existence of this ratio (and often the existence of positive densities for μ_X and $\mu_{X'}$ with respect to the Lebesgue measure, at least on some bounded subset of $\mathbb{R}^d$) is almost always a necessary condition for their theoretical analysis.

However, in the Computer Experiment context in which we are interested, this ratio need not exist. Indeed, while some prior information on the law μ_X may be known, such as bounds on its support, mean or dispersion, it may happen for example that some components of the vector X be tied to each other by deterministic relations of the form $h(X) = 0$, so that the actual support of μ_X might be contained in a low-dimensional manifold and difficult to determine precisely, see Figure 1.

Therefore, designing a synthetic probability distribution $\mu_{X'}$ with respect to which μ_X is absolutely continuous may actually turn out to be impossible. Nevertheless, one may retain the idea to approximate QI by an estimator of the form (1), where the weights

$(\widehat{w}_1, \dots, \widehat{w}_{n_{\rm off}})$ only depend on the samples $\mathbf{X}_{n_{\rm on}}$ and $\mathbf{X}'_{n_{\rm off}}$, and are determined by minimizing some distance between the empirical measure $\widehat{\mu}_{\mathbf{X}_{n_{\rm on}}}$ and weighted empirical measures of the form

$$\widehat{\mu}^{\mathbf{w}_{n_{\rm off}}}_{\mathbf{X}'_{n_{\rm off}}} = \frac{1}{n_{\rm off}} \sum_{j=1}^{n_{\rm off}} w_j \delta_{X'_j}, \qquad \mathbf{w}_{n_{\rm off}} := (w_1, \dots, w_{n_{\rm off}}).$$

This idea was for example applied in the UQ context in [1, 2]. Notice that for $\widehat{\mu}^{\mathbf{w}_{n_{\rm off}}}_{\mathbf{X}'_{n_{\rm off}}}$ to be a probability measure, the weights $(w_1, \dots, w_{n_{\rm off}})$ must satisfy

$$\forall j \in [\![1, n_{\rm off}]\!], \quad w_j \geq 0, \qquad \text{and} \quad \sum_{j=1}^{n_{\rm off}} w_j = n_{\rm off}. \tag{2}$$

In this paper, we follow this approach and study the estimator $\widehat{\rm QI}_{n_{\rm off}, n_{\rm on}}$ of QI obtained by minimizing the *Wasserstein distance*, whose definition is recalled below, between $\widehat{\mu}_{\mathbf{X}_{n_{\rm on}}}$ and $\widehat{\mu}^{\mathbf{w}_{n_{\rm off}}}_{\mathbf{X}'_{n_{\rm off}}}$. The main reason for this choice is that, unlike Kullback–Leibler or more general ϕ-divergences, or L^p distances, the Wasserstein distance between two probability measures on $\mathbb{R}^d$ is not sensitive to whether these measures have densities with respect to the Lebesgue measure, or are absolutely continuous with respect to one another. The optimal weights can be expressed terms of *Nearest Neighbor* and our estimator can be interpreted as the Monte Carlo evaluation of a Nearest Neighbor Regression under covariate shift, for which we bound the error explicitly.

1.3. **Organization of the paper.** The Wasserstein distance is introduced in Section 2, as well as the explicit form of the optimal weights and their reformulation in terms of Nearest Neighbor. Section 3 is devoted to the analysis of the convergence of the weighted empirical measure to μ_X, in terms of Wasserstein distance. The consistency is studied in Section 3.1 and we state our main result in Section 3.2, namely the asymptotic rates of convergence. The link between these results and the estimation of QI is discussed in Sections 4.1, 4.2 and 4.3, with the computation of rates of convergence for $\widehat{\rm QI}_{n_{\rm off}, n_{\rm on}}$. Some links with the Nearest Neighbor literature are highlighted in Section 4.4. Numerical experiments are performed in Section 5, in which the impact of the difference between μ_X and $\mu_{X'}$ is investigated.

1.4. **Notation.** Throughout this paper, we denote by $\mathbb{N}$ the set of the natural integers including zero and by $\mathbb{N}^* = \mathbb{N} \setminus \{0\}$ the set of the positive integers. Given two integers $n_1 \leq n_2$, the set of the integers between n_1 and n_2 is written $[\![n_1, n_2]\!] = \{n_1, \dots, n_2\}$. For $x \in \mathbb{R}$, $\lceil x \rceil$ (resp. $\lfloor x \rfloor$) is the unique integer verifying $x \leq \lceil x \rceil < x + 1$ (resp. $x - 1 < \lfloor x \rfloor \leq x$). For $(x, y) \in \mathbb{R}^2$, we use the join and meet notation $x \wedge y = \min(x, y)$ and $x \vee y = \max(x, y)$. The supremum norm of $\phi : \mathbb{R}^d \to \mathbb{R}$ is denoted by $\|\phi\|_\infty = \sup_{x \in \mathbb{R}^d} |\phi(x)|$.

2. Wasserstein distance minimization and Nearest Neighbor Regression

2.1. **Optimal weights for Wasserstein distances.** We begin by recalling the definition of the Wasserstein distance. Throughout this article, we fix a norm $|\cdot|$ on $\mathbb{R}^d$, which need not be the Euclidean norm.

Definition 2.1 (Wasserstein distance)**.** *Let $\mathcal{P}(\mathbb{R}^d)$ be the set of probability measures on $\mathbb{R}^d$ and, for any $q \in [1, +\infty)$, let*

$$\mathcal{P}_q(\mathbb{R}^d) = \left\{ \nu \in \mathcal{P}(\mathbb{R}^d) : \int_{\mathbb{R}^d} |x|^q \mathrm{d}\nu(x) < +\infty \right\}.$$

The Wasserstein distance of order q between μ and $\nu \in \mathcal{P}_q(\mathbb{R}^d)$ is defined as

$$W_q(\mu, \nu) = \inf \left\{ \int_{\mathbb{R}^d \times \mathbb{R}^d} |x - x'|^q \mathrm{d}\gamma(x, x') : \gamma \in \Pi(\mu, \nu) \right\}^{1/q},$$

where $\Pi(\mu, \nu)$ is the set of probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals μ and ν.

We refer to [19, Section 6] for a general introduction to Wasserstein distances.

This definition allows for an explicit resolution of the minimization problem on $\mathbf{w}_{n_{\text{off}}}$, which relies on the notion of *Nearest Neighbor*. For $x \in \mathbb{R}^d$ and $k \in [\![1, n_{\text{off}}]\!]$, we denote by $\mathrm{NN}^{(k)}_{\mathbf{X}'_{n_{\text{off}}}}(x)$ the *k-th Nearest Neighbor* (k-NN) of x among the sample $\mathbf{X}'_{n_{\text{off}}}$, that is to say the k-th closest point to x among $X'_1, \ldots, X'_{n_{\text{off}}}$ for the norm $|\cdot|$. If there are several such points, we define $\mathrm{NN}^{(k)}_{\mathbf{X}'_{n_{\text{off}}}}(x)$ to be the point X'_j with lowest index j. We omit the superscript notation (k) when referring to the 1-NN, *i.e.*

$$\mathrm{NN}_{\mathbf{X}'_{n_{\text{off}}}}(x) = \mathrm{NN}^{(1)}_{\mathbf{X}'_{n_{\text{off}}}}(x).$$

In the next statement, for any $i \in [\![1, n_{\text{on}}]\!]$ and $l \in [\![1, n_{\text{off}}]\!]$, we denote by $j_i^{(l)}$ the (lowest) index j such that $X'_j = \mathrm{NN}^{(l)}_{\mathbf{X}'_{n_{\text{off}}}}(X_i)$.

Proposition 2.2 (Optimal vector of weights)**.** *Let the k-NN vector of weights $\mathbf{w}^{(k)}_{n_{\text{off}}} = (w^{(k)}_1, \ldots, w^{(k)}_{n_{\text{off}}})$ be defined by, for all $j, k \in [\![1, n_{\text{off}}]\!]$,*

$$w_j^{(k)} := \frac{n_{\text{off}}}{k n_{\text{on}}} \sum_{i=1}^{n_{\text{on}}} \sum_{l=1}^{k} \mathbb{1}_{\{j = j_i^{(l)}\}}. \tag{3}$$

The vector $\mathbf{w}^{(k)}_{n_{\text{off}}}$ satisfies (2) and verifies, for all $q \in [1, +\infty)$,

$$W_q^q \left(\widehat{\mu}_{\mathbf{X}_{n_{\text{on}}}}, \widehat{\mu}^{\mathbf{w}^{(k)}_{n_{\text{off}}}}_{\mathbf{X}'_{n_{\text{off}}}} \right) \leq \frac{1}{k n_{\text{on}}} \sum_{i=1}^{n_{\text{on}}} \sum_{l=1}^{k} \left| X_i - \mathrm{NN}^{(l)}_{\mathbf{X}'_{n_{\text{off}}}}(X_i) \right|^q. \tag{4}$$

For $k = 1$, the equality is reached

$$W_q^q \left(\widehat{\mu}_{\mathbf{X}_{n_{\text{on}}}}, \widehat{\mu}^{\mathbf{w}^{(1)}_{n_{\text{off}}}}_{\mathbf{X}'_{n_{\text{off}}}} \right) = \frac{1}{n_{\text{on}}} \sum_{i=1}^{n_{\text{on}}} \left| X_i - \mathrm{NN}_{\mathbf{X}'_{n_{\text{off}}}}(X_i) \right|^q, \tag{5}$$

and the vector is optimal in the sense that for any $\mathbf{w}_{n_{\text{off}}} = (w_1, \ldots, w_{n_{\text{off}}})$ which also satisfies (2), we have

$$W_q \left(\widehat{\mu}_{\mathbf{X}_{n_{\text{on}}}}, \widehat{\mu}^{\mathbf{w}^{(1)}_{n_{\text{off}}}}_{\mathbf{X}'_{n_{\text{off}}}} \right) \leq W_q \left(\widehat{\mu}_{\mathbf{X}_{n_{\text{on}}}}, \widehat{\mu}^{\mathbf{w}_{n_{\text{off}}}}_{\mathbf{X}'_{n_{\text{off}}}} \right). \tag{6}$$

In other words, for a given $j \in [\![1, n_{\text{off}}]\!]$, $w_j^{(k)}$ is proportional to the number of points X_i of which X'_j is one of the first k Nearest Neighbors. We refer to [13] for a numerical illustration of the use of the vector of weights $\mathbf{w}^{(1)}_{n_{\text{off}}}$ in the context of classification under covariate shift.

Proof. For a general vector of weights $\mathbf{w}_{n_{\mathrm{off}}} = (w_1, \dots, w_{n_{\mathrm{off}}})$ which satisfies (2), the Wasserstein distance $W_q^q\left(\widehat{\mu}_{\mathbf{X}_{n_{\mathrm{on}}}}, \widehat{\mu}_{\mathbf{X}'_{n_{\mathrm{off}}}}^{\mathbf{w}_{n_{\mathrm{off}}}}\right)$ is the solution of the following optimal transport problem

$$
\begin{aligned}
&\inf_{(\gamma_{i,j})_{(i,j)\in[\![1,n_{\mathrm{on}}]\!]\times[\![1,n_{\mathrm{off}}]\!]}} \sum_{i=1}^{n_{\mathrm{on}}}\sum_{j=1}^{n_{\mathrm{off}}} \gamma_{i,j}|X_i - X'_j|^q, \\
&\forall i \in [\![1, n_{\mathrm{on}}]\!], \quad \sum_{j=1}^{n_{\mathrm{off}}} \gamma_{i,j} = \frac{1}{n_{\mathrm{on}}} \quad \text{(marginal condition on } \widehat{\mu}_{\mathbf{X}_{n_{\mathrm{on}}}}\text{)}, \\
&\forall j \in [\![1, n_{\mathrm{off}}]\!], \quad \sum_{i=1}^{n_{\mathrm{on}}} \gamma_{i,j} = \frac{w_j}{n_{\mathrm{off}}} \quad \text{(marginal condition on } \widehat{\mu}_{\mathbf{X}'_{n_{\mathrm{off}}}}^{\mathbf{w}_{n_{\mathrm{off}}}}\text{)}, \\
&\forall (i,j) \in [\![1, n_{\mathrm{on}}]\!] \times [\![1, n_{\mathrm{off}}]\!], \quad \gamma_{i,j} \geq 0,
\end{aligned}
\tag{7}
$$

where $\gamma_{i,j}$ is the coefficient of the discrete transport plan between δ_{X_i} and $\delta_{X'_j}$. For the k-NN vector of weights $\mathbf{w}_{n_{\mathrm{off}}}^{(k)}$ defined by (3), the transport plan

$$
\gamma_{i,j}^{(k)} = \frac{1}{k n_{\mathrm{on}}} \sum_{l=1}^{k} \mathbb{1}_{\{j = j_i^{(l)}\}}
$$

satisfies the two marginal conditions. Reordering the terms in the associated cost gives the upper bound of Equation (4).

We now prove the equality (5) and optimality (6) of $\mathbf{w}_{n_{\mathrm{off}}}^{(1)}$ at the same time. For a given $\mathbf{w}_{n_{\mathrm{off}}}$, if we drop the marginal condition on $\widehat{\mu}_{\mathbf{X}'_{n_{\mathrm{off}}}}^{\mathbf{w}_{n_{\mathrm{off}}}}$, the values of $(\gamma_{i,j})_{j\in[\![1,n_{\mathrm{off}}]\!]}$ for a given i do not constrain the values of $(\gamma_{i',j})_{j\in[\![1,n_{\mathrm{off}}]\!]}$ for another $i' \neq i$. Thus, the optimal values can be found by minimizing separately the following subproblem for $i \in [\![1, n_{\mathrm{on}}]\!]$

$$
\begin{aligned}
&\inf_{(\gamma_{i,j})_{j\in[\![1,n_{\mathrm{off}}]\!]}} \sum_{j=1}^{n_{\mathrm{off}}} \gamma_{i,j}|X_i - X'_j|^q, \\
&\sum_{j=1}^{n_{\mathrm{off}}} \gamma_{i,j} = \frac{1}{n_{\mathrm{on}}}, \\
&\forall j \in [\![1, n_{\mathrm{off}}]\!], \quad \gamma_{i,j} \geq 0,
\end{aligned}
$$

the solution of which is trivially $\frac{1}{n_{\mathrm{on}}}|X_i - \mathrm{NN}_{\mathbf{X}'_{n_{\mathrm{off}}}}(X_i)|^q$. As a consequence, we get the estimate

$$
W_q^q\left(\widehat{\mu}_{\mathbf{X}_{n_{\mathrm{on}}}}, \widehat{\mu}_{\mathbf{X}'_{n_{\mathrm{off}}}}^{\mathbf{w}_{n_{\mathrm{off}}}}\right) \geq \frac{1}{n_{\mathrm{on}}} \sum_{i=1}^{n} |X_i - \mathrm{NN}_{\mathbf{X}'_{n_{\mathrm{off}}}}(X_i)|^q \tag{8}
$$

for any $\mathbf{w}_{n_{\mathrm{off}}}$ satisfying Equation (2). Taking $\mathbf{w}_{n_{\mathrm{off}}} = \mathbf{w}_{n_{\mathrm{off}}}^{(1)}$ in the left-hand side and combining this inequality with (4) for $k = 1$, we obtain both the equality (5) and optimality (6). □

Remark 2.3. *In order to alleviate notation, from now on we shall write* $\widehat{\mu}_{\mathbf{X}'_{n_{\mathrm{off}}}}^{(k)} = \widehat{\mu}_{\mathbf{X}'_{n_{\mathrm{off}}}}^{\mathbf{w}_{n_{\mathrm{off}}}^{(k)}}$.

2.2. **NNR reformulation.** With the choice of weights $\mathbf{w}_{n_{\mathrm{off}}}^{(k)}$ introduced in Proposition 2.2, the resulting estimator of QI writes

$$
\widehat{\mathrm{QI}}_{n_{\mathrm{off}},n_{\mathrm{on}}}^{(k)} = \frac{1}{n_{\mathrm{on}}} \sum_{i=1}^{n_{\mathrm{on}}} \frac{1}{k} \sum_{l=1}^{k} \phi\left(Y'_{j_i^{(l)}}\right),
$$

which makes the method very close to Nearest Neighbor Regression (NNR) [3, Chapter 9], since it may be reformulated as the following two-step procedure:

(1) define the *regression function* ψ of $\phi(Y)$ on X by, for any $x \in \mathbb{R}^d$,

$$\psi(x) := \mathbb{E}\left[\phi(Y)|X = x\right] = \mathbb{E}\left[\phi\left(f(x, \Theta)\right)\right], \tag{9}$$

and let the k-NNR estimator of $\psi(x)$ be given by

$$\widehat{\psi}^{(k)}_{n_{\rm off}}(x) := \frac{1}{k}\sum_{l=1}^{k} \phi\left(Y'_{j^{(l)}(x)}\right), \tag{10}$$

where $j^{(l)}(x)$ is the (lowest) index j such that $X'_j = \mathrm{NN}^{(l)}_{\mathbf{X}'_{n_{\rm off}}}(x)$;

(2) approximate the expectation

$$\mathrm{QI} = \mathbb{E}[\phi(Y)] = \mathbb{E}[\psi(X)]$$

by the empirical mean

$$\frac{1}{n_{\rm on}}\sum_{i=1}^{n_{\rm on}} \widehat{\psi}^{(k)}_{n_{\rm off}}(X_i) = \widehat{\mathrm{QI}}^{(k)}_{n_{\rm off},n_{\rm on}}.$$

In this context, the peculiar fact that the law of the *evaluation* set $X_1, \ldots, X_{n_{\rm on}}$ differs from the law of the *training* set $X'_1, \ldots, X'_{n_{\rm off}}$ is referred to as *domain adaptation* [15]. From a UQ point of view, the first step may be reinterpreted as the construction, based on the Nearest Neighbor approach, of a metamodel for the regression function ψ.

3. **Convergence analysis**

As is evidenced by its reformulation in terms of NNR, the method does not actually depend on the choice of the observable $\phi \circ f$, and its primary purpose is rather the direct estimation of the law μ_X by the weighted empirical measure $\widehat{\mu}^{(k)}_{\mathbf{X}'_{n_{\rm off}}}$. Rewriting

$$\mathrm{QI} = \mathbb{E}[\phi(Y)] = \int_{\mathbb{R}^d} \psi(x)\mathrm{d}\mu_X(x),$$

we observe that estimates on the approximation of QI by $\widehat{\mathrm{QI}}^{(k)}_{n_{\rm off},n_{\rm on}}$ which are uniform in ϕ can be obtained from estimates on the approximation of μ_X by $\widehat{\mu}^{(k)}_{\mathbf{X}'_{n_{\rm off}}}$. Therefore, we turn our attention to the convergence, when the sizes $n_{\rm off}$ and $n_{\rm on}$ of the two samples go to ∞, of $\widehat{\mu}^{(k)}_{\mathbf{X}'_{n_{\rm off}}}$ to μ_X. We naturally work with Wasserstein distances.

3.1. **Consistency.**

Let us fix $q \in [1, +\infty)$ and use Jensen's inequality to write

$$\mathbb{E}\left[W_q^q\left(\mu_X, \widehat{\mu}^{(k)}_{\mathbf{X}'_{n_{\rm off}}}\right)\right] \leq 2^{q-1}\left(\mathbb{E}\left[W_q^q\left(\mu_X, \widehat{\mu}_{\mathbf{X}_{n_{\rm on}}}\right)\right] + \mathbb{E}\left[W_q^q\left(\widehat{\mu}_{\mathbf{X}_{n_{\rm on}}}, \widehat{\mu}^{(k)}_{\mathbf{X}'_{n_{\rm off}}}\right)\right]\right). \tag{11}$$

As soon as there exists $s > q$ such that $\mathbb{E}[|X|^s] < +\infty$, the first term $\mathbb{E}[W_q^q\left(\mu_X, \widehat{\mu}_{\mathbf{X}_{n_{\rm on}}}\right)]$ is known to converge to 0 when $n_{\rm on} \to +\infty$ and explicit rates are available [9], see also the discussion in Subsection 4.1 below. We therefore focus on the second term and first observe that, by Proposition 2.2, we have

$$\begin{aligned}\mathbb{E}\left[W_q^q\left(\widehat{\mu}_{\mathbf{X}_{n_{\rm on}}}, \widehat{\mu}^{(1)}_{\mathbf{X}'_{n_{\rm off}}}\right)\right] &= \mathbb{E}\left[\frac{1}{n_{\rm on}}\sum_{i=1}^{n_{\rm on}}\left|X_i - \mathrm{NN}_{\mathbf{X}'_{n_{\rm off}}}(X_i)\right|^q\right] \\ &= \mathbb{E}\left[\left|X - \mathrm{NN}_{\mathbf{X}'_{n_{\rm off}}}(X)\right|^q\right],\end{aligned} \tag{12}$$

for $k = 1$, and

$$\mathbb{E}\left[W_q^q\left(\widehat{\mu}_{\mathbf{X}_{n_{\mathrm{on}}}}, \widehat{\mu}^{(k)}_{\mathbf{X}'_{n_{\mathrm{off}}}}\right)\right] \leq \mathbb{E}\left[\frac{1}{kn_{\mathrm{on}}}\sum_{i=1}^{n_{\mathrm{on}}}\sum_{l=1}^{k}\left|X_i - \mathrm{NN}^{(l)}_{\mathbf{X}'_{n_{\mathrm{off}}}}(X_i)\right|^q\right] = \frac{1}{k}\sum_{l=1}^{k}\mathbb{E}\left[\left|X - \mathrm{NN}^{(l)}_{\mathbf{X}'_{n_{\mathrm{off}}}}(X)\right|^q\right], \tag{13}$$

for $k \geq 2$. Observe that the right-hand side of both (12) and (13) no longer depend on n_{on}.

We now formulate two crucial assumptions and then state our first main result. For all $x \in \mathbb{R}^d$ and $r \geq 0$, we denote $B(x,r) := \{x' \in \mathbb{R}^d : |x - x'| \leq r\}$, and recall that the *support* of a probability measure $\nu \in \mathcal{P}(\mathbb{R}^d)$ is defined by

$$\mathrm{supp}(\nu) := \left\{x \in \mathbb{R}^d : \forall r > 0, \nu(B(x,r)) > 0\right\}.$$

Assumption 3.1 (Support condition)**.** *We have* $\mathrm{supp}(\mu_X) \subset \mathrm{supp}(\mu_{X'})$.

Assumption 3.2 (Min-integrability)**.** *There exists an integer* $m_0 \geq 1$ *such that*

$$\mathbb{E}\left[\min_{j \in [\![1, m_0]\!]} |X'_j|\right] < +\infty.$$

Theorem 3.3 (Consistency)**.** *Let Assumptions 3.1 and 3.2 hold. For all* $q \in [1, +\infty)$ *such that* $\mathbb{E}[|X|^q] < +\infty$*, and any sequence of positive integers* $(k_n)_{n \geq 1}$ *such that* $k_n/n \to 0$ *when* $n \to \infty$*, we have*

$$\lim_{n_{\mathrm{off}} \to +\infty} \mathbb{E}\left[W_q^q\left(\widehat{\mu}_{\mathbf{X}_{n_{\mathrm{on}}}}, \widehat{\mu}^{(k_{n_{\mathrm{off}}})}_{\mathbf{X}'_{n_{\mathrm{off}}}}\right)\right] = 0,$$

uniformly in n_{on}.

Remark 3.4 (On Assumption 3.2)**.** *Assumption 3.2 is obviously satisfied if* X' *has a finite first order moment, but also for some heavy-tailed distributions. It writes under the equivalent form*

$$\int_0^\infty \mathbb{P}(|X'| > r)^{m_0} \mathrm{d}r < +\infty,$$

which may be easier to check. An example of a random variable which does not satisfy this assumption, in dimension $d = 1$*, is* $X' = \exp(1/U)$ *where* U *is a uniform random variable on* $[0, 1]$.

Theorem 3.3 is proved in Subsection 3.3.

3.2. **Rates of convergence.** The next step of our study consists in complementing Theorem 3.3 with a rate of convergence. We first discuss the case $k = 1$. Following (12), we start by writing

$$\mathbb{E}\left[W_q^q\left(\widehat{\mu}_{\mathbf{X}_{n_{\mathrm{on}}}}, \widehat{\mu}^{(1)}_{\mathbf{X}'_{n_{\mathrm{off}}}}\right)\right] = \mathbb{E}\left[\left|X - \mathrm{NN}_{\mathbf{X}'_{n_{\mathrm{off}}}}(X)\right|^q\right] = \mathbb{E}\left[\mathbb{E}\left[\left|X - \mathrm{NN}_{\mathbf{X}'_{n_{\mathrm{off}}}}(X)\right|^q \middle| X\right]\right], \tag{14}$$

and observe that for any $x \in \mathrm{supp}(\mu_X)$, $|x - \mathrm{NN}_{\mathbf{X}'_{n_{\mathrm{off}}}}(x)| = \min_{j \in [\![1, n_{\mathrm{off}}]\!]} |x - X'_j|$. If there is an open set U of $\mathbb{R}^d$ containing x and such that $\mu_{X'}(\cdot \cap U)$ has a density $p_{X'}$ with

respect to the Lebesgue measure which is continuous at x, then an elementary computation shows that, for all $r \geq 0$,

$$\lim_{n_{\text{off}} \to +\infty} \mathbb{P}\left({n_{\text{off}}}^{1/d} \min_{j \in [\![1, n_{\text{off}}]\!]} |x - X'_j| > r \right) = \exp\left(-r^d v_d p_{X'}(x)\right),$$

where v_d denotes the volume of the unit sphere of $\mathbb{R}^d$ for the norm $|\cdot|$. If $p_{X'}(x) > 0$ then this indicates that the correct order of convergence in Theorem 3.3 should be ${n_{\text{off}}}^{-q/d}$. If $p_{X'}(x) = 0$, or if the measure $\mu_{X'}(\cdot \cap U)$ is not absolutely continuous with respect to the Lebesgue measure, it is easy to construct elementary examples yielding different rates of convergence; see also [3, Chapter 2] for the singular case. We leave these peculiarities apart and work under the following strengthening of the support condition of Assumption 3.1.

Assumption 3.5 (Strong support condition)**.** *There exists an open set $U \subset \mathbb{R}^d$ which contains $\operatorname{supp}(\mu_X)$ and such that:*

(i) the measure $\mu_{X'}(\cdot \cap U)$ has a density $p_{X'}$ with respect to the Lebesgue measure;

(ii) the density $p_{X'}$ is continuous and positive on U;

(iii) there exist $\kappa \in (0, 1]$ and $r_\kappa > 0$ such that, for any $x \in U$, for any $r \in [0, r_\kappa]$,

$$\mathbb{P}\left(X' \in B(x, r)\right) \geq \kappa p_{X'}(x) v_d r^d.$$

Obviously, Assumption 3.5 implies Assumption 3.1 because then $\operatorname{supp}(\mu_X) \subset U \subset \operatorname{supp}(\mu_{X'})$. Part (iii) of Assumption 3.5 was introduced in [10] in the context of Nearest Neighbor Classification, and called *Strong minimal mass assumption* there.

Under Assumption 3.5, for all $x \in \operatorname{supp}(\mu_X)$, a positive random variable Z such that $\mathbb{P}(Z > r) = \exp(-r^d v_d p_{X'}(x))$ has moments

$$\mathbb{E}\left[Z^q\right] = \frac{\Gamma(1 + q/d)}{(v_d p_{X'}(x))^{q/d}},$$

where Γ denotes Euler's Gamma function. Therefore, one may expect the normalized quantity

$${n_{\text{off}}}^{q/d} \mathbb{E}\left[W_q^q \left(\widehat{\mu}_{\mathbf{X}_{n_{\text{on}}}}, \widehat{\mu}^{(1)}_{\mathbf{X}'_{n_{\text{off}}}} \right) \right]$$

to converge, when $n_{\text{off}} \to +\infty$, to

$$\frac{\Gamma(1 + q/d)}{v_d^{q/d}} \mathbb{E}\left[\frac{1}{p_{X'}(X)^{q/d}} \right],$$

so that as soon as the expectation in the right-hand side is finite, the rate of convergence in Theorem 3.3 is ${n_{\text{off}}}^{-q/d}$. In order to prove this convergence we shall actually need the following stronger integrability assumption.

Assumption 3.6 (Moments)**.** *Under Assumption 3.5, we have*

$$\mathbb{E}\left[\frac{1 + |X|^q}{p_{X'}(X)^{q/d}} \right] < +\infty.$$

Assumptions 3.5 and 3.6 are discussed in more detail below. We now state our second main result.

Theorem 3.7 (Convergence rates for $k = 1$)**.** *Let Assumptions 3.2 and 3.5 hold, and let $q \in [1, +\infty)$ be such that Assumption 3.6 holds. Then we have*

$$\lim_{n_{\mathrm{off}} \to +\infty} {n_{\mathrm{off}}}^{q/d} \mathbb{E}\left[W_q^q \left(\widehat{\mu}_{\mathbf{X}_{n_{\mathrm{on}}}}, \widehat{\mu}^{(1)}_{\mathbf{X}'_{n_{\mathrm{off}}}} \right) \right] = \frac{\Gamma(1+q/d)}{v_d^{q/d}} \mathbb{E}\left[\frac{1}{p_{X'}(X)^{q/d}} \right].$$

Theorem 3.7 is proved in Subsection 3.3.

We now discuss the estimation of $\widehat{\mu}_{\mathbf{X}_{n_{\mathrm{on}}}}$ by the weighted empirical measure $\widehat{\mu}^{(k)}_{\mathbf{X}'_{n_{\mathrm{off}}}}$ for an arbitrary $k \in [\![1, n_{\mathrm{off}}]\!]$. By (8), we first observe that we always have

$$W_q \left(\widehat{\mu}_{\mathbf{X}_{n_{\mathrm{on}}}}, \widehat{\mu}^{(k)}_{\mathbf{X}'_{n_{\mathrm{off}}}} \right) \geq W_q \left(\widehat{\mu}_{\mathbf{X}_{n_{\mathrm{on}}}}, \widehat{\mu}^{(1)}_{\mathbf{X}'_{n_{\mathrm{off}}}} \right),$$

so that the estimation of $\widehat{\mu}_{\mathbf{X}_{n_{\mathrm{on}}}}$ is deteriorated by increasing the number of neighbors. Still, in the asymptotic regime of Theorem 3.3, a bound of the same order of magnitude as Theorem 3.7 may be obtained.

Corollary 3.8 (Convergence rates for k-NN)**.** *Under the assumptions of Theorem 3.7, for any nondecreasing sequence of positive integers $(k_n)_{n \geq 1}$ such that $k_n/n \to 0$ when $n \to \infty$, we have*

$$\limsup_{n_{\mathrm{off}} \to +\infty} \left(\frac{n_{\mathrm{off}}}{k_{n_{\mathrm{off}}}} \right)^{q/d} \mathbb{E}\left[W_q^q \left(\widehat{\mu}_{\mathbf{X}_{n_{\mathrm{on}}}}, \widehat{\mu}^{(k_{n_{\mathrm{off}}})}_{\mathbf{X}'_{n_{\mathrm{off}}}} \right) \right] \leq c_{d,q} \frac{\Gamma(1+q/d)}{v_d^{q/d}} \mathbb{E}\left[\frac{1}{p_{X'}(X)^{q/d}} \right],$$

with some constant $c_{d,q} > 1$.

Corollary 3.8 is proved in Subsection 3.3, where the explicit expression of the constant $c_{d,q}$ is also given.

Let us conclude this subsection with some comments on Assumptions 3.5 and 3.6. When X has a compact support, Assumptions 3.5 and 3.6 are verified as soon as $\mu_{X'}$ has a continuous density $p_{X'}$ which is bounded from below on U; these results are similar to the case $\mu_X = \mu_{X'}$ in [3, Section 2]. It is however interesting to note that these assumptions also hold in some nontrivial noncompact cases.

An example of a sufficient condition for Assumption 3.5 is given in the next statement, which is proved in Subsection 3.3.

Lemma 3.9 (Sufficient condition for Assumption 3.5)**.** *Let $\| \cdot \|$ be a norm on $\mathbb{R}^d$, not necessarily identical to $| \cdot |$, induced by an inner product. If $\mu_{X'}$ has a density $p_{X'}$ with respect to the Lebesgue measure on $\mathbb{R}^d$, which writes $p_{X'}(x) = h(\|x - x_0\|)$ for some $x_0 \in \mathbb{R}^d$ and $h : [0, +\infty) \to \mathbb{R}$ continuous, positive and nonincreasing, then Assumption 3.5 holds with $U = \mathbb{R}^d$.*

We also refer to [10, Section 2.4] for a discussion of this assumption.

Assumption 3.6 gives a relationship between μ_X and $p_{X'}$ to ensure the convergence. In essence, it asserts that the tail of μ_X must be quite lightweight compared to the tail of $p_{X'}$. For instance, if X and X' are centered Gaussian vectors with respective covariance $\sigma^2 I_d$ and $\sigma'^2 I_d$, then by Lemma 3.9, Assumption 3.5 is satisfied with $U = \mathbb{R}^d$, and it is easy to check that for $q \in [1, +\infty)$, Assumption 3.6 holds if and only if $\sigma'^2 > \sigma^2 q/d$.

3.3. **Proofs.** In this subsection, we present the proofs of Theorems 3.3 and 3.7, Corollary 3.8 and Lemma 3.9.

Proof of Theorem 3.3. We begin our proof with the constant case $k_n = 1$ for all n and then extend it to the general case. We recall that by (12),

$$\mathbb{E}\left[W_q^q\left(\widehat{\mu}_{\mathbf{X}_{n_{\text{on}}}}, \widehat{\mu}^{(1)}_{\mathbf{X}'_{n_{\text{off}}}}\right)\right] = \mathbb{E}\left[\left|X - \text{NN}_{\mathbf{X}'_{n_{\text{off}}}}(X)\right|^q\right] = \mathbb{E}\left[\min_{j\in[\![1,n_{\text{off}}]\!]} |X - X'_j|^q\right].$$

By Assumption 3.1, $X \in \text{supp}(\mu_{X'})$ almost surely, so that we deduce from Lemma 2.2 in [3, Chapter 2] that

$$\min_{j\in[\![1,n_{\text{off}}]\!]} |X - X'_j|^q \xrightarrow[n_{\text{off}}\to+\infty]{a.s} 0.$$

Let m_0 be the integer given by Assumption 3.2, we have

$$\min_{j\in[\![1,n_{\text{off}}]\!]} |X - X'_j|^q \leq 2^{q-1}\left(|X|^q + \min_{j\in[\![1,n_{\text{off}}]\!]} |X'_j|^q\right).$$

The random variable $|X|^q$ is integrable by assumption and for $n_{\text{off}} \geq \lceil q\rceil m_0$, the inequality

$$\begin{aligned}
\mathbb{E}\left[\min_{j\in[\![1,n_{\text{off}}]\!]} \left|X'_j\right|^q\right] &\leq \mathbb{E}\left[\min_{j\in[\![1,n_{\text{off}}]\!]} \left|X'_j\right|^{\lceil q\rceil}\right]^{q/\lceil q\rceil} \\
&\leq \mathbb{E}\left[\min_{j\in[\![1,m_0]\!]} \left|X'_j\right| \min_{j\in[\![m_0+1,2m_0]\!]} \left|X'_j\right| \cdots \min_{j\in[\![(\lceil q\rceil-1)m_0+1,\lceil q\rceil m_0]\!]} \left|X'_j\right|\right]^{q/\lceil q\rceil} \\
&\leq \mathbb{E}\left[\min_{j\in[\![1,m_0]\!]} \left|X'_j\right|\right]^q < +\infty
\end{aligned}$$

holds. Then by the dominated convergence theorem,

$$\mathbb{E}\left[\min_{j=[\![1,n_{\text{off}}]\!]} |X - X'_j|^q\right] \xrightarrow[n_{\text{off}}\to+\infty]{} 0.$$

For the general case $k_n/n \to 0$, we adapt directly the proof of [3, Theorem 2.4] to the context $\mu_X \neq \mu_{X'}$. Let us fix $l \in [\![1, n_{\text{off}}/2]\!]$ and partition the set $\{X'_1, \ldots, X'_{n_{\text{off}}}\}$ into $2l$ sets of size $n_1, \ldots, n_{2l}$ with, for all $j \in [\![1, 2l]\!]$,

$$\lfloor n_{\text{off}}/2l\rfloor \leq n_j \leq \lfloor n_{\text{off}}/2l\rfloor + 1.$$

We denote by $\text{NN}^{(1,j)}_{\mathbf{X}'_{n_{\text{off}}}}$ the 1-NN among the subset j. By the definition of $\text{NN}^{(l)}_{\mathbf{X}'_{n_{\text{off}}}}$, there are at least l subsets j for which

$$|X - \text{NN}^{(l)}_{\mathbf{X}'_{n_{\text{off}}}}(X)| \leq |X - \text{NN}^{(1,j)}_{\mathbf{X}'_{n_{\text{off}}}}(X)|,$$

therefore

$$|X - \text{NN}^{(l)}_{\mathbf{X}'_{n_{\text{off}}}}(X)|^q \leq \frac{1}{l}\sum_{j=1}^{2l} |X - \text{NN}^{(1,j)}_{\mathbf{X}'_{n_{\text{off}}}}(X)|^q,$$

and consequently

$$\mathbb{E}\left[\left|X - \text{NN}^{(l)}_{\mathbf{X}'_{n_{\text{off}}}}(X)\right|^q\right] \leq 2\mathbb{E}\left[\left|X - \text{NN}_{\mathbf{X}'_{\lfloor n_{\text{off}}/2l\rfloor}}(X)\right|^q\right].$$

Finally, we deduce from (13) that, as soon as $k_{n_{\mathrm{off}}} \leq n_{\mathrm{off}}/2$,

$$\begin{aligned}\mathbb{E}\left[W_q^q\left(\widehat{\mu}_{\mathbf{X}_{n_{\mathrm{on}}}}, \widehat{\mu}^{(k_{n_{\mathrm{off}}})}_{\mathbf{X}'_{n_{\mathrm{off}}}}\right)\right] &\leq \frac{1}{k_{n_{\mathrm{off}}}} \sum_{l=1}^{k_{n_{\mathrm{off}}}} \mathbb{E}\left[\left|X - \mathrm{NN}^{(l)}_{\mathbf{X}'_{n_{\mathrm{off}}}}(X)\right|^q\right] \\ &\leq \frac{2}{k_{n_{\mathrm{off}}}} \sum_{l=1}^{k_{n_{\mathrm{off}}}} \mathbb{E}\left[\left|X - \mathrm{NN}_{\mathbf{X}'_{\lfloor n_{\mathrm{off}}/2l \rfloor}}(X)\right|^q\right] \\ &\leq 2\mathbb{E}\left[\left|X - \mathrm{NN}_{\mathbf{X}'_{\lfloor n_{\mathrm{off}}/2k_{n_{\mathrm{off}}} \rfloor}}(X)\right|^q\right],\end{aligned} \tag{15}$$

which goes to 0 as a consequence of the first part of the proof when $n_{\mathrm{off}}/2k_{n_{\mathrm{off}}}$ goes to infinity. □

Proof of Theorem 3.7. By (12), we have

$$\begin{aligned}\mathbb{E}\left[{n_{\mathrm{off}}}^{q/d} W_q^q(\widehat{\mu}_{\mathbf{X}_{n_{\mathrm{on}}}}, \widehat{\mu}^{(1)}_{\mathbf{X}'_{n_{\mathrm{off}}}})\right] &= \mathbb{E}\left[\mathbb{E}\left[{n_{\mathrm{off}}}^{q/d} \min_{j \in [\![1, n_{\mathrm{off}}]\!]} \left|X - X'_j\right|^q \middle| X\right]\right] \\ &= \int_{\mathbb{R}^d} \int_0^{+\infty} \mathbb{P}({n_{\mathrm{off}}}^{q/d} \min_{j \in [\![1, n_{\mathrm{off}}]\!]} |x - X'_j|^q > t) \mathrm{d}t \mathrm{d}\mu_X(x) \\ &= \int_{\mathbb{R}^d} \int_0^{+\infty} \mathbb{P}({n_{\mathrm{off}}}^{q/d} |x - X'|^q > t)^{n_{\mathrm{off}}} \mathrm{d}t \mathrm{d}\mu_X(x),\end{aligned} \tag{16}$$

by independence of the X'_j. The proof consists in computing the pointwise limit of $\mathbb{P}({n_{\mathrm{off}}}^{p/d}|x - Y|^p > t)^{n_{\mathrm{off}}}$ for $(x,t) \in \mathrm{supp}(\mu_X) \times \mathbb{R}^+$ and then establishing the convergence of the integral via the dominated convergence theorem.

Pointwise convergence. We have

$$\begin{aligned}\mathbb{P}({n_{\mathrm{off}}}^{q/d}|x - X'|^q > t)^{n_{\mathrm{off}}} &= \left(1 - \mathbb{P}(|x - X'| \leq t^{1/q}/{n_{\mathrm{off}}}^{1/d})\right)^{n_{\mathrm{off}}} \\ &= \exp\left(n_{\mathrm{off}} \log\left(1 - \mathbb{P}(|x - X'| \leq t^{1/q}/{n_{\mathrm{off}}}^{1/d})\right)\right).\end{aligned}$$

By Assumption 3.5, we have

$$\mathbb{P}(|x - X'| \leq t^{1/q}/{n_{\mathrm{off}}}^{1/d}) = p_{X'}(x) v_d t^{d/q}/n_{\mathrm{off}} + \mathrm{o}(1/n_{\mathrm{off}}),$$

with v_d the volume of the unit sphere. Thus

$$n_{\mathrm{off}} \log\left(1 - \mathbb{P}(|x - X'| \leq t^{1/q}/{n_{\mathrm{off}}}^{1/d})\right) = -p_{X'}(x) v_d t^{d/q} + \mathrm{o}(1),$$

and we conclude that

$$\mathbb{P}({n_{\mathrm{off}}}^{q/d}|x - X'|^q > t)^{n_{\mathrm{off}}} \underset{n_{\mathrm{off}} \to +\infty}{\longrightarrow} \exp\left(-p_{X'}(x) v_d t^{d/q}\right).$$

Dominated convergence. Let $r_\kappa > 0$ be given by Assumption 3.5. We split the integral in the right-hand side of (16) and study each term separately

$$\int_{\mathbb{R}^d} \int_0^{+\infty} \mathbb{P}({n_{\mathrm{off}}}^{q/d}|x - X'|^q > t)^{n_{\mathrm{off}}} \mathrm{d}t \mathrm{d}\mu_X(x) = \mathrm{I} + \mathrm{II}$$

with

$$\begin{aligned}\mathrm{I} &:= \int_{\mathbb{R}^d} \int_0^{r_\kappa^q {n_{\mathrm{off}}}^{q/d}} \mathbb{P}(|x - X'| > t^{1/q}/{n_{\mathrm{off}}}^{1/d})^{n_{\mathrm{off}}} \mathrm{d}t \mathrm{d}\mu_X(x), \\ \mathrm{II} &:= \int_{\mathbb{R}^d} \int_{r_\kappa^q {n_{\mathrm{off}}}^{q/d}}^{+\infty} \mathbb{P}(|x - X'| > t^{1/q}/{n_{\mathrm{off}}}^{1/d})^{n_{\mathrm{off}}} \mathrm{d}t \mathrm{d}\mu_X.\end{aligned}$$

Convergence of I. For $t \in [0, r_\kappa^q {n_{\text{off}}}^{q/d}]$, we have ${t^{1/q}}/{{n_{\text{off}}}^{1/d}} \leq r_\kappa$ and thus

$$\begin{aligned}\mathbb{P}(|x - X'| > {t^{1/q}}/{{n_{\text{off}}}^{1/d}})^{n_{\text{off}}} &= \left(1 - \mathbb{P}(|x - X'| \leq {t^{1/q}}/{{n_{\text{off}}}^{1/d}})\right)^{n_{\text{off}}} \\ &\leq \left(1 - \frac{p_{X'}(x) v_d \kappa t^{d/q}}{n_{\text{off}}}\right)^{n_{\text{off}}}\end{aligned}$$

by Assumption 3.5.

Using the elementary inequality $(1 - a/n)^n \leq \exp(-a)$ for $a \leq n$, we can write

$$\mathbb{P}(|x - X'| > {t^{1/q}}/{{n_{\text{off}}}^{1/d}})^{n_{\text{off}}} \leq \exp(-\kappa v_d p_{X'}(x) t^{d/q}).$$

This bound does not depend on n_{off} and the integral

$$\begin{aligned}\int_{\mathbb{R}^d} \int_0^{+\infty} \exp\left(-\kappa v_d p_{X'}(x) t^{d/q}\right) \mathrm{d}t \mathrm{d}\mu_X(x) &= \int_{\mathbb{R}^d} \frac{\Gamma(1 + q/d)}{(\kappa v_d p_{X'}(x))^{q/d}} \mathrm{d}\mu_X(x) \\ &= \frac{\Gamma(1 + q/d)}{(\kappa v_d)^{q/d}} \mathbb{E}\left[\frac{1}{p_{X'}(X)^{q/d}}\right],\end{aligned}$$

is finite by Assumption 3.6. We therefore deduce from the dominated convergence theorem that

$$\mathrm{I} \underset{n_{\text{off}} \to +\infty}{\longrightarrow} \int_{\mathbb{R}^d} \int_0^{+\infty} \exp\left(-p_{X'}(x) v_d t^{d/q}\right) \mathrm{d}t \mu_X(dx) = \frac{\Gamma(1 + q/d)}{v_d^{q/d}} \mathbb{E}\left[\frac{1}{p_{X'}(X)^{q/d}}\right].$$

Convergence of II. Let $n_{\text{off}} \geq 2(q+1)m_0$. Using the change of variable $r^q = t/{n_{\text{off}}}^{q/d}$, we have

$$\begin{aligned}\mathrm{II} &= q \int_{\mathbb{R}^d} \int_{r_\kappa}^{+\infty} {n_{\text{off}}}^{q/d} r^{q-1} \mathbb{P}(|x - X'| > r)^{n_{\text{off}}} \mathrm{d}r \mathrm{d}\mu_X(x) \\ &\leq q \int_{\mathbb{R}^d} \int_{r_\kappa}^{+\infty} V_{n_{\text{off}}}(x, r) \mathrm{d}r \mathrm{d}\mu_X(x),\end{aligned}$$

with

$$V_{n_{\text{off}}}(x, r) := {n_{\text{off}}}^{q/d} \mathbb{P}(|x - X'| > r_\kappa)^{n_{\text{off}} - (q+1)m_0} r^{q-1} \mathbb{P}(|x - X'| > r)^{(q+1)m_0}.$$

As $\mathbb{P}(|x - X'| > r_\kappa) < 1$ for all x in U, by Assumption 3.5, $V_{n_{\text{off}}}(x, r)$ is pointwise convergent to 0 on the support of μ_X. We check that $V_{n_{\text{off}}}(x, r)$ is bounded from above by an integrable function which does not depend on n_{off}. Let us denote ${n_{\text{off}}}' = n_{\text{off}} - (q + 1)m_0 \geq n_{\text{off}}/2$ and rewrite

$$\begin{aligned}{n_{\text{off}}}^{q/d} \mathbb{P}(|x - X'| > r_\kappa)^{n_{\text{off}} - (q+1)m_0} &= \left(\frac{n_{\text{off}}}{{n_{\text{off}}}'}\right)^{q/d} {{n_{\text{off}}}'}^{q/d} \mathbb{P}(|x - X'| > r_\kappa)^{{n_{\text{off}}}'} \\ &\leq 2^{q/d} {{n_{\text{off}}}'}^{q/d} \left(1 - \frac{\mathbb{P}(|x - X'| \leq r_\kappa)}{{n_{\text{off}}}'}\right)^{{n_{\text{off}}}'} \\ &\leq 2^{q/d} {{n_{\text{off}}}'}^{q/d} \exp\left(-{n_{\text{off}}}' \kappa p_{X'}(x) v_d r_\kappa^d\right),\end{aligned}$$

where we have used Assumption 3.5 and the elementary above inequality at the third line. We deduce that

$$ {n_{\text{off}}}^{q/d} \mathbb{P}(|x - X'| > r_\kappa)^{n_{\text{off}} - (q+1)m_0} \leq \frac{C_1}{p_{X'}(x)^{q/d}}, \qquad C_1 := \frac{2^{q/d}}{(\kappa v_d r_\kappa^d)^{q/d}} \sup_{u \geq 0} (u^{q/d} e^{-u}),$$

so that

$$V_{n_{\text{off}}}(x, r) \leq \widetilde{V}(x, r) := \frac{C_1}{p_{X'}(x)^{q/d}} r^{q-1} \mathbb{P}(|x - X'| > r)^{(q+1)m_0}. \tag{17}$$

To complete the proof, we verify that $\widetilde{V}(x,r)$ is integrable on $U \times [r_\kappa, +\infty)$. We first fix $x \in \mathbb{R}^d$ and estimate the integral of $\widetilde{V}(x,r)$ in r. Using the fact that if $|x - X'| > r$ then $|X'| > r - |x|$, we first write

$$\begin{aligned}\int_{r_\kappa}^{+\infty} r^{q-1}\mathbb{P}(|x-X'|>r)^{(q+1)m_0}\mathrm{d}r &\leq \int_0^{+\infty} r^{q-1}\mathbb{P}(|X'|>r-|x|)^{(q+1)m_0}\mathrm{d}r \\ &= \int_{-|x|}^{+\infty}(r+|x|)^{q-1}\mathbb{P}(|X'|>r)^{(q+1)m_0}\mathrm{d}r.\end{aligned}$$

On the interval $[-|x|, 0]$, we have

$$\int_{-|x|}^0 (r+|x|)^{q-1}\mathbb{P}(|X'|>r)^{(q+1)m_0}\mathrm{d}r = \int_{-|x|}^0 (r+|x|)^{q-1}\mathrm{d}r = \frac{|x|^q}{q}.$$

On the interval $[0, +\infty)$, we first rewrite

$$\int_0^{+\infty}(r+|x|)^{q-1}\mathbb{P}(|X'|>r)^{(q+1)m_0}\mathrm{d}r = \int_0^{+\infty}(r+|x|)^{q-1}\mathbb{P}\left(\min_{j\in[\![1,m_0]\!]}|X'_j|>r\right)^{q+1}\mathrm{d}r,$$

and recall from Assumption 3.2 that $C_2 := \mathbb{E}[\min_{j\in[\![1,m_0]\!]}|X'_j|] < \infty$. As a consequence, we deduce from Markov's inequality that the right-hand side in the previous equality is bounded from above by

$$\int_0^{|x|\vee 1}(r+|x|)^{q-1}\mathrm{d}r + C_2^{q+1}\int_{|x|\vee 1}^{+\infty}\frac{(r+|x|)^{q-1}}{r^{q+1}}\mathrm{d}r.$$

If $|x| \leq 1$ then this expression is bounded from above. If $|x| > 1$, then we have

$$\int_0^{|x|}(r+|x|)^{q-1}\mathrm{d}r \leq 2^{q-1}|x|^q$$

on the one hand, and

$$\int_{|x|}^{+\infty}\frac{(r+|x|)^{q-1}}{r^{q+1}}\mathrm{d}r = \frac{1}{|x|}\int_1^{+\infty}\frac{(u+1)^{q-1}}{u^{q+1}}\mathrm{d}u,$$

which is bounded from above, on the other hand. Overall, we conclude that there exists a constant C_3 such that

$$\int_{r_\kappa}^{+\infty} r^{q-1}\mathbb{P}(|x-X'|>r)^{(q+1)m_0}\mathrm{d}r \leq C_3(1+|x|^q). \tag{18}$$

As a consequence, the combination of (17) and (18) yields

$$\int_{\mathbb{R}^d}\int_{r_\kappa}^{+\infty}\widetilde{V}(x,r)\mathrm{d}r\mathrm{d}\mu_X(x) \leq C_1C_3\mathbb{E}\left[\frac{1+|X|^q}{p_{X'}(X)^{q/d}}\right],$$

which by Assumption 3.6 allows to apply the dominated convergence theorem to show that II goes to 0, and thereby completes the proof. □

Proof of Corollary 3.8. We start from the second line of Equation (15) and estimate its right-hand side

$$\begin{aligned}\mathbb{E}\left[W_q^q(\widehat{\mu}_{\mathbf{X}_{n_{\mathrm{on}}}}, \widehat{\mu}_{\mathbf{X}'_{n_{\mathrm{off}}}}^{(k_{n_{\mathrm{off}}})})\right] &\leq \frac{2}{k_{n_{\mathrm{off}}}} \sum_{l=1}^{k_{n_{\mathrm{off}}}} \mathbb{E}\left[\left|X - \mathrm{NN}_{\mathbf{X}'_{\lfloor n_{\mathrm{off}}/2l \rfloor}}(X)\right|^q\right] \\ &= \frac{2}{k_{n_{\mathrm{off}}}} \sum_{l=1}^{k_{n_{\mathrm{off}}}} \left(\frac{2k_{n_{\mathrm{off}}}}{n_{\mathrm{off}}} \frac{l}{k_{n_{\mathrm{off}}}} \frac{n_{\mathrm{off}}}{2l}\right)^{q/d} \mathbb{E}\left[\left|X - \mathrm{NN}_{\mathbf{X}'_{\lfloor n_{\mathrm{off}}/2l \rfloor}}(X)\right|^q\right] \\ &= \left(\frac{k_{n_{\mathrm{off}}}}{n_{\mathrm{off}}}\right)^{q/d} \frac{2^{q/d+1}}{k_{n_{\mathrm{off}}}} \sum_{l=1}^{k_{n_{\mathrm{off}}}} \left(\frac{l}{k_{n_{\mathrm{off}}}}\right)^{q/d} F\left(\frac{n_{\mathrm{off}}}{2l}\right)\end{aligned}$$

with $F(u) = u^{q/d}\mathbb{E}[|X - \mathrm{NN}_{\mathbf{X}'_{\lfloor u \rfloor}}(X)|^q]$. Let $\epsilon > 0$. By Theorem 3.7, there exists $u_\epsilon \geq 0$ such that, for all $u \geq u_\epsilon$,

$$\left|F(u) - \frac{\Gamma(1+q/d)}{v_d^{q/d}} \mathbb{E}\left[\frac{1}{p_{X'}(X)^{q/d}}\right]\right| \leq \epsilon.$$

We can remark that for $n_{\mathrm{off}} \in \mathbb{N}^*$ and $l \in [\![1, k_{n_{\mathrm{off}}}]\!]$,

$$\frac{n_{\mathrm{off}}}{2l} \geq \frac{n_{\mathrm{off}}}{2k_{n_{\mathrm{off}}}} \xrightarrow[n_{\mathrm{off}} \to +\infty]{} +\infty.$$

Thus, if we take n_ϵ such that for all $n \geq n_\epsilon$, $\left\lfloor \frac{n}{2k_n} \right\rfloor \geq u_\epsilon$, we have

$$\left|F\left(\frac{n_{\mathrm{off}}}{2l}\right) - \frac{\Gamma(1+q/d)}{v_d^{q/d}} \mathbb{E}\left[\frac{1}{p_{X'}(X)^{q/d}}\right]\right| \leq \epsilon$$

for any $n_{\mathrm{off}} \geq n_\epsilon$ and $l \leq k_{n_{\mathrm{off}}}$. Consequently,

$$\begin{aligned}&\left|\frac{1}{k_{n_{\mathrm{off}}}} \sum_{l=1}^{k_{n_{\mathrm{off}}}} \left(\frac{l}{k_{n_{\mathrm{off}}}}\right)^{q/d} \left(F\left(\frac{n_{\mathrm{off}}}{2l}\right) - \frac{\Gamma(1+q/d)}{v_d^{q/d}} \mathbb{E}\left[\frac{1}{p_{X'}(X)^{q/d}}\right]\right)\right| \\ &\leq \epsilon \left|\frac{1}{k_{n_{\mathrm{off}}}} \sum_{l=1}^{k_{n_{\mathrm{off}}}} \left(\frac{l}{k_{n_{\mathrm{off}}}}\right)^{q/d}\right| \\ &\leq \epsilon,\end{aligned}$$

so that

$$\lim_{n_{\mathrm{off}} \to +\infty} \frac{2^{q/d+1}}{k_{n_{\mathrm{off}}}} \sum_{l=1}^{k_{n_{\mathrm{off}}}} \left(\frac{l}{k_{n_{\mathrm{off}}}}\right)^{q/d} F\left(\frac{n_{\mathrm{off}}}{2l}\right) = c_{d,q} \frac{\Gamma(1+q/d)}{v_d^{q/d}} \mathbb{E}\left[\frac{1}{p_{X'}(X)^{q/d}}\right],$$

where

$$\begin{aligned}c_{d,q} &:= \lim_{n \to +\infty} \frac{2^{q/d+1}}{k_n} \sum_{l=1}^{k_n} \left(\frac{l}{k_n}\right)^{q/d} \\ &= \begin{cases} \dfrac{2^{q/d+1}}{k} \displaystyle\sum_{l=1}^{k} \left(\frac{l}{k}\right)^{q/d} & \text{if } \sup_{n \geq 1} k_n = k < +\infty, \\ 2^{q/d+1} \displaystyle\int_0^1 u^{q/d} \mathrm{d}u = \frac{2^{q/d+1}}{q/d+1} & \text{if } \sup_{n \geq 1} k_n = +\infty, \end{cases}\end{aligned}$$

concluding the proof. □

Proof of Lemma 3.9. Obviously, it suffices to check that $p_{X'}$ satisfies the point (iii) of Assumption 3.5. Let us denote by $\langle \cdot, \cdot \rangle$ and $\mathcal{B}(x,r)$ respectively the inner product and the ball of center x and radius r associated to $\|\cdot\|$. We set $x_0 = 0$ without loss of generality. As h is positive and nonincreasing, we may fix $r_0 > 0$ and define

$$\overline{\kappa} := \frac{h(r_0)}{h(0)} \in (0,1].$$

If $\|x\| \leq r_0/2$, then for all $y \in \mathcal{B}(x, r_0/2)$, the monotonicity of h ensures that $p_{X'}(x+y) \geq \overline{\kappa} p_{X'}(x)$. By the equivalence of the norms, there exist $C \geq c > 0$ such that for any $x \in \mathbb{R}^d$ and any $r \geq 0$, $\mathcal{B}(x, cr) \subset B(x,r) \subset \mathcal{B}(x, Cr)$. Thus

$$\forall r \leq r_0/2c, \qquad \mathbb{P}\left(X' \in B(x,r)\right) \geq \mathbb{P}\left(X' \in \mathcal{B}(x,cr)\right) \geq (c/C)^d v_d \overline{\kappa} p_{X'}(x) r^d.$$

If $\|x\| > r_0/2$, let us introduce the half-cone

$$\mathcal{C}_x = \left\{ x' \in \mathbb{R}^d : \langle x' - x, -x \rangle \geq \frac{\|x' - x\| \|x\|}{2} \right\},$$

and notice that for all $r \leq r_0/2$ and $x' \in \mathcal{C}_x \cap \mathcal{B}(x,r)$,

$$\begin{aligned} \|x'\|^2 &= \|x\|^2 + \|x' - x\|^2 + 2\langle x' - x, x \rangle \\ &\leq \|x\|^2 + \|x' - x\|^2 - \|x' - x\| \|x\| \\ &\leq \|x\|^2 + \|x' - x\|^2 - \|x' - x\|^2 \\ &= \|x\|^2. \end{aligned}$$

Thus, for all $x' \in \mathcal{C}_x \cap \mathcal{B}(x,r)$, $p_{X'}(x') \geq p_{X'}(x)$. For a given r, the sets $\mathcal{C}_x \cap \mathcal{B}(x,r)$ have the same volume for all x, which we denote by $\alpha v_d r^d$ for some $\alpha \in (0, 1/C^d)$. Finally, we have

$$\forall r \leq r_0/2c, \qquad \mathbb{P}(X' \in B(x,r)) \geq \mathbb{P}(X' \in \mathcal{B}(x,cr) \cap \mathcal{C}_x) \geq \alpha c^d v_d p_{X'}(x) r^d.$$

If we take $\kappa = (c/C)^d \min(\alpha C^d, \overline{\kappa})$ and $r_\kappa = r_0/2c$, we obtain the point (iii) of Assumption 3.5. □

4. **Discussion**

Going back to our initial problem, we are now able to compute L^q rates of convergence of the weighted estimator

$$\widehat{\mathrm{QI}}^{(k_{n_{\mathrm{off}}})}_{n_{\mathrm{off}}, n_{\mathrm{on}}} = \frac{1}{n_{\mathrm{off}}} \sum_{j=1}^{n_{\mathrm{off}}} w_j^{(k_{n_{\mathrm{off}}})} \phi(Y_j')$$

to $\mathrm{QI} = \mathbb{E}[\phi(Y)]$. First, in Section 4.1, we derive the convergence rates of $\widehat{\mu}^{(k_{n_{\mathrm{off}}})}_{\mathbf{X}'_{n_{\mathrm{off}}}}$ to μ_X in terms of Wasserstein distance. Then in Section 4.2, we study the case in which $Y = f(X)$ and there is no external source of uncertainty, that we call the noiseless case, using the terminology from statistical Machine Learning regression. The noisy case $Y = f(X, \Theta)$ is treated in Section 4.3.

Finally, in Section 4.4, we reinterpret Theorem 3.7 under the prism of the Nearest Neighbor literature.

4.1. **Convergence to μ_X.** Let us focus on the speed of convergence of $\widehat{\mu}_{\mathbf{X}'_{n_{\text{off}}}}^{(k_{n_{\text{off}}})}$ to μ_X. Provided that X has enough moments, namely that there exists $s > 2q$ such that $\mathbb{E}[|X|^s] < +\infty$, we have from [9, Theorem 1]

$$\mathbb{E}\left[W_q^q\left(\mu_X, \widehat{\mu}_{\mathbf{X}_{n_{\text{on}}}}\right)\right] = \begin{cases} \mathrm{O}\left({n_{\text{on}}}^{-1/2}\right) & \text{if } q > d/2, \\ \mathrm{O}\left({n_{\text{on}}}^{-1/2}\log(1+n_{\text{on}})\right) & \text{if } q = d/2, \\ \mathrm{O}\left({n_{\text{on}}}^{-q/d}\right) & \text{if } q < d/2. \end{cases}$$

In dimension $d = 1$, we deduce from (11) that

$$\mathbb{E}\left[W_q^q\left(\mu_X, \widehat{\mu}_{\mathbf{X}'_{n_{\text{off}}}}^{(k_{n_{\text{off}}})}\right)\right]^{1/q} = \mathrm{O}\left(\left(\frac{1}{n_{\text{on}}}\right)^{1/2q} + \frac{k_{n_{\text{off}}}}{n_{\text{off}}}\right)$$

for any value of $q \geq 1$. The right-hand side is minimized for the choice $q = 1$, in which case both error terms have the same order of magnitude if the sizes of the offline and online samples satisfy

$$n_{\text{on}} \propto \left(\frac{n_{\text{off}}}{k_{n_{\text{off}}}}\right)^2.$$

In dimension $d \geq 2$, the minimal upper bound for $\mathbb{E}[W_q^q(\mu_X, \widehat{\mu}_{\mathbf{X}_{n_{\text{on}}}})]^{1/q}$ is achieved for $q \leq d/2$, in which case, up to the logarithmic correction,

$$\mathbb{E}\left[W_q^q\left(\mu_X, \widehat{\mu}_{\mathbf{X}'_{n_{\text{off}}}}^{(k_{n_{\text{off}}})}\right)\right]^{1/q} = \mathrm{O}\left(\left(\frac{1}{n_{\text{on}}}\right)^{1/d} + \left(\frac{k_{n_{\text{off}}}}{n_{\text{off}}}\right)^{1/d}\right),$$

and both error terms have the same order of magnitude if the sizes of the offline and online samples satisfy

$$n_{\text{on}} \propto \frac{n_{\text{off}}}{k_{n_{\text{off}}}}.$$

4.2. **Rate of convergence of $\widehat{\mathrm{QI}}_{n_{\text{off}},n_{\text{on}}}^{(k_{n_{\text{off}}})}$ in the noiseless case.** We assume that $Y = f(X)$ and study the rate of convergence of $\widehat{\mathrm{QI}}_{n_{\text{off}},n_{\text{on}}}^{(k_{n_{\text{off}}})}$ to QI. When $\phi \circ f$ is L-Lipschitz continuous, we can derive the result using the duality formula of the W_1 Wasserstein distance [19, Remark 6.5]

$$W_1\left(\mu_X, \widehat{\mu}_{\mathbf{X}'_{n_{\text{off}}}}^{(k_{n_{\text{off}}})}\right) = \sup_{|\varphi|_{\text{Lip}} \leq 1} \left\{\int_{\mathbb{R}^d} \varphi(x)\mathrm{d}\mu_X(x) - \int_{\mathbb{R}^d} \varphi(x)\mathrm{d}\widehat{\mu}_{\mathbf{X}'_{n_{\text{off}}}}^{(k_{n_{\text{off}}})}(x)\right\} \tag{19}$$

and bound

$$\begin{aligned} \left|\mathrm{QI} - \widehat{\mathrm{QI}}_{n_{\text{off}},n_{\text{on}}}^{(k_{n_{\text{off}}})}\right|^q &= \left|\int_{\mathbb{R}^d} \phi\circ f(x)\mathrm{d}\mu_X(x) - \int_{\mathbb{R}^d} \phi\circ f(x)\mathrm{d}\widehat{\mu}_{\mathbf{X}'_{n_{\text{off}}}}^{(k_{n_{\text{off}}})}(x)\right|^q \\ &\leq L W_1^q\left(\mu_X, \widehat{\mu}_{\mathbf{X}'_{n_{\text{off}}}}^{(k_{n_{\text{off}}})}\right) \\ &\leq L W_q^q\left(\mu_X, \widehat{\mu}_{\mathbf{X}'_{n_{\text{off}}}}^{(k_{n_{\text{off}}})}\right). \end{aligned}$$

We can conclude from Section 4.1.

Proposition 4.1 (Rates of convergence in the noiseless case)**.** *Assume that:*

(i) the function f does not depend on Θ,

(ii) the function $\phi \circ f$ is globally Lipschitz continuous,

and let the assumptions of Corollary 3.8 hold. We have, as soon as $q \neq d/2$ and there exists $s > 2q$ such that $\mathbb{E}[|X|^s] < +\infty$,

$$\mathbb{E}\left[\left|\mathrm{QI} - \widehat{\mathrm{QI}}^{(k_{n_{\mathrm{off}}})}_{n_{\mathrm{off}},n_{\mathrm{on}}}\right|^q\right]^{1/q} = \mathrm{O}\left(n_{\mathrm{on}}^{-\min(1/2q,1/d)}\right) + \mathrm{O}\left(\left(\frac{k_{n_{\mathrm{off}}}}{n_{\mathrm{off}}}\right)^{1/d}\right). \tag{20}$$

There is no need for $k_{n_{\mathrm{off}}}$ to go to infinity and thus $k_{n_{\mathrm{off}}} = 1$ seems a reasonable choice.

These computations can be adapted to cases other than $\phi \circ f$ Lipschitz continuous. For instance, if $A \subset \mathbb{R}^e$, $\phi(y) = \mathbb{1}_{\{y \in A\}}$ and f is globally Lipschitz continuous, it is possible to use the margin assumption of [17] to deduce theoretical rates of convergence in the estimation of $\mathrm{QI} = \mathbb{P}(Y \in A)$.

4.3. **Noisy case.** We now study the convergence of $\widehat{\mathrm{QI}}^{(k_{n_{\mathrm{off}}})}_{n_{\mathrm{off}},n_{\mathrm{on}}}$ to QI when $Y = f(X,\Theta)$. A first striking result is then that even under the assumptions of Theorem 3.3, the estimator $\widehat{\mathrm{QI}}^{(1)}_{n_{\mathrm{off}},n_{\mathrm{on}}}$ need not be consistent. Indeed, consider the case where X is actually deterministic and always equal to some $x_0 \in \mathbb{R}^d$. Then we have

$$\widehat{\mathrm{QI}}^{(1)}_{n_{\mathrm{off}},n_{\mathrm{on}}} = \frac{1}{n_{\mathrm{on}}} \sum_{i=1}^{n_{\mathrm{on}}} \phi(Y'_{j_i^{(1)}}),$$

where $j_i^{(1)}$ is the index of the closest X'_j to X_i. But since $X_i = x_0$ for all i, all indices $j_i^{(1)}$ are equal to some $j^{(1)}$ and the estimator rewrites

$$\widehat{\mathrm{QI}}^{(1)}_{n_{\mathrm{off}},n_{\mathrm{on}}} = \phi(Y'_{j^{(1)}}) = \phi(f(X'_{j^{(1)}}, \Theta_{j^{(1)}})).$$

While Assumption 3.1 ensures that $X'_{j^{(1)}}$ converges to x_0 when $n_{\mathrm{off}} \to +\infty$, in general the corresponding sequence of $\Theta_{j^{(1)}}$ does not converge.

As is evidenced on this example, the presence of an atom in the law of X makes the estimator $\widehat{\mathrm{QI}}^{(1)}_{n_{\mathrm{off}},n_{\mathrm{on}}}$ depend on a single realization of Θ and therefore prevents this estimator to display an averaging behavior with respect to the law of Θ. In Proposition 4.2, we clarify this point by exhibiting a necessary and sufficient condition for the estimator $\widehat{\mathrm{QI}}^{(1)}_{n_{\mathrm{off}},n_{\mathrm{on}}}$ to be consistent, while in Proposition 4.3, we show that replacing $\widehat{\mathrm{QI}}^{(1)}_{n_{\mathrm{off}},n_{\mathrm{on}}}$ with $\widehat{\mathrm{QI}}^{(k_{n_{\mathrm{off}}})}_{n_{\mathrm{off}},n_{\mathrm{on}}}$ with $k_{n_{\mathrm{off}}} \to +\infty$ allows to recover such an averaging behavior and make the estimator consistent, even when μ_X has atoms. In the latter case, we also provide rates of convergence.

We recall that $\psi(x) = \mathbb{E}[\phi(f(x,\Theta))]$ is defined in Equation (9). In the next statement, we denote by $\mathcal{A}_X$ the set of atoms of μ_X, that is to say the set of $x \in \mathbb{R}^d$ such that $\mathbb{P}(X = x) > 0$, and introduce the notation $\vartheta(x) := \mathrm{Var}(\phi(f(x,\Theta)))$.

Proposition 4.2 (Consistency of the 1-NN in the noisy case)**.** *Assume that:*

(i) the function ϕ is bounded,

(ii) the function ψ is globally Lipschitz continuous,

(iii) the function ϑ is continuous,

and let the assumptions of Theorem 3.3 hold. We have

$$\mathbb{E}\left[\left|\widehat{\mathrm{QI}}^{(1)}_{n_{\mathrm{off}},n_{\mathrm{on}}} - \mathrm{QI}\right|\right] \xrightarrow[n_{\mathrm{off}},n_{\mathrm{on}} \to +\infty]{} 0$$

if and only if,

$$\forall x \in \mathcal{A}_X, \quad \mathrm{Var}(\phi(f(x,\Theta))) = 0.$$

In particular, under the above assumptions, if the law of X has no atom, *i.e.* $\mathcal{A}_X = \varnothing$, then $\widehat{\mathrm{QI}}^{(1)}_{n_{\mathrm{off}},n_{\mathrm{on}}}$ converges to QI.

Proof. Let us write

$$\widehat{\mathrm{QI}}^{(1)}_{n_{\mathrm{off}},n_{\mathrm{on}}} - \mathrm{QI} = \left(\widehat{\mathrm{QI}}^{(1)}_{n_{\mathrm{off}},n_{\mathrm{on}}} - \widetilde{\mathrm{QI}}^{(1)}_{n_{\mathrm{off}},n_{\mathrm{on}}}\right) + \left(\widetilde{\mathrm{QI}}^{(1)}_{n_{\mathrm{off}},n_{\mathrm{on}}} - \mathrm{QI}\right),$$

with

$$\widetilde{\mathrm{QI}}^{(1)}_{n_{\mathrm{off}},n_{\mathrm{on}}} = \frac{1}{n_{\mathrm{off}}} \sum_{j=1}^{n_{\mathrm{off}}} w_j^{(1)} \psi(X'_j).$$

Using the Lipschitz continuity of ψ, the duality formula (19) and Theorem 3.3, we get that $\widetilde{\mathrm{QI}}^{(1)}_{n_{\mathrm{off}},n_{\mathrm{on}}} - \mathrm{QI}$ converges to 0 when $n_{\mathrm{off}}, n_{\mathrm{on}} \to +\infty$, in L^1. Therefore, $\widehat{\mathrm{QI}}^{(1)}_{n_{\mathrm{off}},n_{\mathrm{on}}} - \mathrm{QI}$ converges to 0 if and only if $\widehat{\mathrm{QI}}^{(1)}_{n_{\mathrm{off}},n_{\mathrm{on}}} - \widetilde{\mathrm{QI}}^{(1)}_{n_{\mathrm{off}},n_{\mathrm{on}}}$ converges to 0.

Let us rewrite

$$\begin{aligned}
\widehat{\mathrm{QI}}^{(1)}_{n_{\mathrm{off}},n_{\mathrm{on}}} - \widetilde{\mathrm{QI}}^{(1)}_{n_{\mathrm{off}},n_{\mathrm{on}}} &= \frac{1}{n_{\mathrm{off}}} \sum_{j=1}^{n_{\mathrm{off}}} w_j^{(1)} \left(\phi\left((X'_j,\Theta_j)\right) - \psi\left(X'_j\right)\right) \\
&= \frac{1}{n_{\mathrm{on}}} \sum_{i=1}^{n_{\mathrm{on}}} \left(\phi\left(f\left(X'_{j_i^{(1)}},\Theta_{j_i^{(1)}}\right)\right) - \psi\left(X'_{j_i^{(1)}}\right)\right),
\end{aligned}$$

introduce the notation

$$\mathcal{A}_X^+ := \{x \in \mathcal{A}_X : \vartheta(X) > 0\},$$

and denote

$$\begin{aligned}
e_1 &:= \frac{1}{n_{\mathrm{on}}} \sum_{i=1}^{n_{\mathrm{on}}} \left(\phi(f(X'_{j_i^{(1)}},\Theta_{j_i^{(1)}})) - \psi(X'_{j_i^{(1)}})\right) \mathbb{1}_{\{X_i \notin \mathcal{A}_X^+\}}, \\
e_2 &:= \frac{1}{n_{\mathrm{on}}} \sum_{i=1}^{n_{\mathrm{on}}} \left(\phi(f(X'_{j_i^{(1)}},\Theta_{j_i^{(1)}})) - \psi(X'_{j_i^{(1)}})\right) \mathbb{1}_{\{X_i \in \mathcal{A}_X^+\}}.
\end{aligned}$$

In Step 1 below, we prove that

$$\mathbb{E}\left[|e_1|\right] \xrightarrow[n_{\mathrm{on}},n_{\mathrm{off}}\to+\infty]{} 0,$$

while in Step 2, we show that if $\mathcal{A}_X^+ \neq \varnothing$ then $\mathbb{E}[|e_2|]$ does not converge to 0, which implies that in this case, $\widehat{\mathrm{QI}}^{(1)}_{n_{\mathrm{off}},n_{\mathrm{on}}} - \widetilde{\mathrm{QI}}^{(1)}_{n_{\mathrm{off}},n_{\mathrm{on}}}$ does not converge to 0 in L^1.

In both steps, we shall use the following preliminary remark: given a measurable subset $\mathcal{A}$ of $\mathbb{R}^d$, taking the conditional expectation with respect to $(\mathbf{X}_{n_{\mathrm{on}}}, \mathbf{X}'_{n_{\mathrm{off}}})$ it is easy to see that for $i \in [\![1, n_{\mathrm{on}}]\!]$,

$$\mathbb{E}\left[(\phi(f(X'_{j_i^{(1)}},\Theta_{j_i^{(1)}})) - \psi(X'_{j_i^{(1)}}))\mathbb{1}_{\{X_i \in \mathcal{A}\}}\right] = 0,$$

and for $(i_1, i_2) \in [\![1, n_{\mathrm{on}}]\!]^2$,

$$\begin{aligned}
&\mathbb{E}\left[(\phi(f(X'_{j_{i_1}^{(1)}},\Theta_{j_{i_1}^{(1)}})) - \psi(X'_{j_i^{(1)}{}_1}))\mathbb{1}_{\{X_{i_1} \in \mathcal{A}\}}(\phi(f(X'_{j_{i_2}^{(1)}},\Theta_{j_{i_2}^{(1)}})) - \psi(X'_{j_{i_2}^{(1)}}))\mathbb{1}_{\{X_{i_2} \in \mathcal{A}\}}\right] \\
&= \mathbb{E}\left[\mathbb{1}_{\{j_{i_1}^{(1)} = j_{i_2}^{(1)}\}} \vartheta(X'_{j_{i_1}^{(1)}}) \mathbb{1}_{\{X_{i_1} \in \mathcal{A}, X_{i_2} \in \mathcal{A}\}}\right].
\end{aligned}$$

Therefore,

$$\mathbb{E}[|e_1|^2] = \frac{1}{{n_{\mathrm{on}}}^2}\left(\sum_{i=1}^{n_{\mathrm{on}}}\mathbb{E}\left[\vartheta(X'_{j_i^{(1)}})\mathbb{1}_{\{X_i\notin\mathcal{A}_X^+\}}\right] + \sum_{i_1\neq i_2}\mathbb{E}\left[\mathbb{1}_{\{j_{i_1}^{(1)}=j_{i_2}^{(1)}\}}\vartheta(X'_{j_{i_1}^{(1)}})\mathbb{1}_{\{X_{i_1}\notin\mathcal{A}_X^+,X_{i_2}\notin\mathcal{A}_X^+\}}\right]\right)$$
$$= \frac{1}{n_{\mathrm{on}}}\mathbb{E}\left[\vartheta(X'_{j_1^{(1)}})\mathbb{1}_{\{X_1\in\mathcal{A}_X^+\}}\right] + \frac{n_{\mathrm{on}}-1}{n_{\mathrm{on}}}\mathbb{E}\left[\mathbb{1}_{\{j_1^{(1)}=j_2^{(1)}\}}\vartheta(X'_{j_1^{(1)}})\mathbb{1}_{\{X_1\notin\mathcal{A}_X^+,X_2\notin\mathcal{A}_X^+\}}\right],$$

and a similar expression holds for $\mathbb{E}[|e_2|^2]$.

Step 1. Thanks to the boundedness of ϕ, and thus of ϑ, it is immediate that $\frac{1}{n_{\mathrm{on}}}\mathbb{E}[\vartheta(X'_{j_1^{(1)}})\mathbb{1}_{\{X_1\notin\mathcal{A}_X^+\}}]$ converges to 0 when $n_{\mathrm{on}}\to+\infty$, uniformly in n_{off}. Therefore, to show that $\mathbb{E}[|e_1|^2]$ converges to 0, it suffices to prove that

$$\mathbb{E}\left[\mathbb{1}_{\{j_1^{(1)}=j_2^{(1)}\}}\vartheta(X'_{j_1^{(1)}})\mathbb{1}_{\{X_1\notin\mathcal{A}_X^+,X_2\notin\mathcal{A}_X^+\}}\right] \xrightarrow[n_{\mathrm{off}}\to+\infty]{} 0.$$

In this purpose, let us first write

$$\mathbb{E}\left[\mathbb{1}_{\{j_1^{(1)}=j_2^{(1)}\}}\vartheta(X'_{j_1^{(1)}})\mathbb{1}_{\{X_1\notin\mathcal{A}_X^+,X_2\notin\mathcal{A}_X^+\}}\right] \leq \mathbb{E}\left[\mathbb{1}_{\{\mathrm{NN}_{\mathbf{X}'_{n_{\mathrm{off}}}}(X_1)=\mathrm{NN}_{\mathbf{X}'_{n_{\mathrm{off}}}}(X_2)\}}\vartheta(\mathrm{NN}_{\mathbf{X}'_{n_{\mathrm{off}}}}(X_1))\mathbb{1}_{\{X_1\notin\mathcal{A}_X^+\}}\right],$$

and recall that, by Assumption 3.1 and Lemma 2.2 in [3, Chapter 2], $\mathrm{NN}_{\mathbf{X}'_{n_{\mathrm{off}}}}(X_1)$ converges to X_1 and $\mathrm{NN}_{\mathbf{X}'_{n_{\mathrm{off}}}}(X_2)$ converges to X_2, almost surely. As a consequence, if $X_1\in\mathcal{A}_X\setminus\mathcal{A}_X^+$ then $\vartheta(X_1)=0$ and by the continuity of ϑ and the boundedness of ϕ, the dominated convergence theorem shows that

$$\mathbb{E}\left[\mathbb{1}_{\{X_1\in\mathcal{A}_X\setminus\mathcal{A}_X^+\}}\mathbb{1}_{\{\mathrm{NN}_{\mathbf{X}'_{n_{\mathrm{off}}}}(X_1)=\mathrm{NN}_{\mathbf{X}'_{n_{\mathrm{off}}}}(X_2)\}}\vartheta(\mathrm{NN}_{\mathbf{X}'_{n_{\mathrm{off}}}}(X_1))\right] \xrightarrow[n_{\mathrm{off}}\to+\infty]{} 0.$$

On the other hand, if $X_1\notin\mathcal{A}_X$, then almost surely $X_1\neq X_2$, and therefore $\mathbb{1}_{\{\mathrm{NN}_{\mathbf{X}'_{n_{\mathrm{off}}}}(X_1)=\mathrm{NN}_{\mathbf{X}'_{n_{\mathrm{off}}}}(X_2)\}}$ converges to 0 almost surely. Using the boundedness of ϕ and the dominated convergence theorem again, we deduce that

$$\mathbb{E}\left[\mathbb{1}_{\{X_1\notin\mathcal{A}_X\}}\mathbb{1}_{\{\mathrm{NN}_{\mathbf{X}'_{n_{\mathrm{off}}}}(X_1)=\mathrm{NN}_{\mathbf{X}'_{n_{\mathrm{off}}}}(X_2)\}}\vartheta(\mathrm{NN}_{\mathbf{X}'_{n_{\mathrm{off}}}}(X_1))\right] \xrightarrow[n_{\mathrm{off}}\to+\infty]{} 0,$$

which completes the proof of the fact that $\mathbb{E}[|e_1|^2]$, and thus $\mathbb{E}[|e_1|]$, converge to 0.

Step 2. Let us now assume that $\mathcal{A}_X^+$ is nonempty and show that e_2 does not converge to 0 in L^1. We shall actually prove that e_2 does not converge to 0 in L^2: since e_2 is bounded then this prevents the convergence from occuring in L^1. From the preliminary remark, we write

$$\mathbb{E}[|e_2|^2] = \frac{1}{n_{\mathrm{on}}}\mathbb{E}\left[\vartheta(X'_{j_1^{(1)}})\mathbb{1}_{\{X_1\in\mathcal{A}_X^+\}}\right] + \frac{n_{\mathrm{on}}-1}{n_{\mathrm{on}}}\mathbb{E}\left[\mathbb{1}_{\{j_1^{(1)}=j_2^{(1)}\}}\vartheta(X'_{j_1^{(1)}})\mathbb{1}_{\{X_1\in\mathcal{A}_X^+,X_2\in\mathcal{A}_X^+\}}\right],$$

and we prove that

$$\liminf_{n_{\mathrm{off}}\to+\infty}\mathbb{E}\left[\mathbb{1}_{\{j_1^{(1)}=j_2^{(1)}\}}\vartheta(X'_{j_1^{(1)}})\mathbb{1}_{\{X_1\in\mathcal{A}_X^+,X_2\in\mathcal{A}_X^+\}}\right] > 0.$$

Let $x\in\mathcal{A}_X^+$. Obviously,

$$\mathbb{E}\left[\mathbb{1}_{\{j_1^{(1)}=j_2^{(1)}\}}\vartheta(X'_{j_1^{(1)}})\mathbb{1}_{\{X_1\in\mathcal{A}_X^+,X_2\in\mathcal{A}_X^+\}}\right] \geq \mathbb{E}\left[\mathbb{1}_{\{j_1^{(1)}=j_2^{(1)}\}}\vartheta(X'_{j_1^{(1)}})\mathbb{1}_{\{X_1=X_2=x\}}\right]$$
$$= \mathbb{E}\left[\vartheta(\mathrm{NN}_{\mathbf{X}'_{n_{\mathrm{off}}}}(x))\mathbb{1}_{\{X_1=X_2=x\}}\right].$$

By Assumption 3.1 and Lemma 2.2 in [3, Chapter 2] again, $\mathrm{NN}_{\mathbf{X}'_{n_{\mathrm{off}}}}(x)$ converges to x almost surely, therefore using the continuity and boundedness assumptions on ϑ, the dominated convergence theorem shows that

$$\mathbb{E}\left[\vartheta(\mathrm{NN}_{\mathbf{X}'_{n_{\mathrm{off}}}}(x))\mathbb{1}_{\{X_1=X_2=x\}}\right] \xrightarrow[n_{\mathrm{off}}\to+\infty]{} \vartheta(x)\mu_X(\{x\})^2 > 0,$$

which completes the proof.

□

We now study the estimator $\widehat{\mathrm{QI}}^{(k_{n_{\mathrm{off}}})}_{n_{\mathrm{off}},n_{\mathrm{on}}}$ and show that it is unconditionnally consistent as soon as $k_n \to +\infty$. Besides, we provide L^2 convergence rates.

Proposition 4.3 (Rates of convergence in the noisy case)**.** *Assume that:*

(i) the function ϕ is bounded,
(ii) the function ψ is globally Lipschitz continuous,
(iii) there exists $s > 4$ such that $\mathbb{E}\left[|X|^s\right] < +\infty$,

and the assumptions of Corollary 3.8 hold.

The estimator $\widehat{\mathrm{QI}}^{(k_{n_{\mathrm{off}}})}_{n_{\mathrm{off}},n_{\mathrm{on}}}$ is consistent as soon as $k_{n_{\mathrm{off}}}$ goes to infinity. Besides, the L^2 rate of convergence is optimal when $k_{n_{\mathrm{off}}} \sim {n_{\mathrm{off}}}^{2/(d+2)}$ and is, when $d \neq 4$,

$$\mathbb{E}\left[\left|\widehat{\mathrm{QI}}^{(k_{n_{\mathrm{off}}})}_{n_{\mathrm{off}},n_{\mathrm{on}}} - \mathrm{QI}\right|^2\right]^{1/2} = \mathrm{O}\left({n_{\mathrm{on}}}^{-\min(1/4,1/d)}\right) + \mathrm{O}\left({n_{\mathrm{off}}}^{-1/(d+2)}\right).$$

When $d = 4$, the first term is replaced with ${n_{\mathrm{on}}}^{-1/4}\log(1+n_{\mathrm{on}})^{1/2}$.

The loss of convergence order with respect to Proposition 4.1 is similar to the NNR, in which it deteriorates from the rate $1/d$ in the noiseless case to the rate of $1/(d+2)$ in the noisy case [3, Section 14.6 and Section 15.3].

Proof. We decompose the error as

$$\widehat{\mathrm{QI}}^{(k_{n_{\mathrm{off}}})}_{n_{\mathrm{off}},n_{\mathrm{on}}} - \mathrm{QI} = \left(\widehat{\mathrm{QI}}^{(k_{n_{\mathrm{off}}})}_{n_{\mathrm{off}},n_{\mathrm{on}}} - \widetilde{\mathrm{QI}}^{(k_{n_{\mathrm{off}}})}_{n_{\mathrm{off}},n_{\mathrm{on}}}\right) + \left(\widetilde{\mathrm{QI}}^{(k_{n_{\mathrm{off}}})}_{n_{\mathrm{off}},n_{\mathrm{on}}} - \mathrm{QI}\right),$$

with

$$\widetilde{\mathrm{QI}}^{(k_{n_{\mathrm{off}}})}_{n_{\mathrm{off}},n_{\mathrm{on}}} = \frac{1}{n_{\mathrm{off}}}\sum_{j=1}^{n_{\mathrm{off}}} w_j^{(k_{n_{\mathrm{off}}})}\psi(X'_j).$$

As ψ is globally Lipschitz continuous and does not depend on Θ, we can deduce from Proposition 4.1 that

$$\mathbb{E}\left[\left(\widetilde{\mathrm{QI}}^{(k_{n_{\mathrm{off}}})}_{n_{\mathrm{off}},n_{\mathrm{on}}} - \mathrm{QI}\right)^2\right]^{1/2} = \mathrm{O}\left({n_{\mathrm{on}}}^{-\min(1/4,1/d)}\right) + \mathrm{O}\left(\left(\frac{k_{n_{\mathrm{off}}}}{n_{\mathrm{off}}}\right)^{1/d}\right)$$

when $d \neq 4$ and has an additional logarithmic term when $d = 4$. We write the quadratic error for the first term

$$\mathbb{E}\left[\left|\widehat{\mathrm{QI}}^{(k_{n_{\mathrm{off}}})}_{n_{\mathrm{off}},n_{\mathrm{on}}} - \widetilde{\mathrm{QI}}^{(k_{n_{\mathrm{off}}})}_{n_{\mathrm{off}},n_{\mathrm{on}}}\right|^2\right]$$
$$= \mathbb{E}\left[\left(\frac{1}{n_{\mathrm{off}}}\sum_{j=1}^{n_{\mathrm{off}}} w_j^{(k_{n_{\mathrm{off}}})}(\psi(X'_j) - \phi(f(X'_j,\Theta_j)))\right)^2\right]$$
$$= \mathbb{E}\left[\frac{1}{{n_{\mathrm{off}}}^2}\sum_{j=1}^{n_{\mathrm{off}}} w_j^{(k_{n_{\mathrm{off}}})2}\left(\psi(X'_j) - \phi(f(X'_j,\Theta_j))\right)^2\right]$$
$$+ \mathbb{E}\left[\frac{n_{\mathrm{off}}-1}{{n_{\mathrm{off}}}^2}\sum_{j,l=1,j\neq l}^{n_{\mathrm{off}}} w_j^{(k_{n_{\mathrm{off}}})} w_l^{(k_{n_{\mathrm{off}}})}\left(\psi(X'_j) - \phi(f(X'_j,\Theta_j))\right)\left(\psi(X'_l) - \phi(f(X'_l,\Theta_l))\right)\right].$$

Using the fact that $\mathbb{E}[w_j^{(k_{n_{\mathrm{off}}})} f(X'_j,\Theta_j)|\mathbf{X}_{n_{\mathrm{on}}},\mathbf{X}_{n_{\mathrm{off}}}] = w_j^{(k_{n_{\mathrm{off}}})}\psi(X'_j)$ by definition and the independence of the Θ_j, the cross terms vanish. The remaining quadratic term is

(21)
$$\mathbb{E}\left[\frac{1}{{n_{\mathrm{off}}}^2}\sum_{j=1}^{n_{\mathrm{off}}} w_j^{(k_{n_{\mathrm{off}}})2}\left(\psi(X'_j) - \phi(f(X'_j,\Theta_j))\right)^2\right] = \frac{1}{{n_{\mathrm{off}}}^2}\sum_{j=1}^{n_{\mathrm{off}}}\mathbb{E}\left[w_j^{(k_{n_{\mathrm{off}}})2}(\psi(X'_j) - \phi(f(X'_j,\Theta)))^2\right]$$
$$\leq \frac{4}{{n_{\mathrm{off}}}^2}\sum_{j=1}^{n_{\mathrm{off}}}\mathbb{E}\left[\left(w_j^{(k_{n_{\mathrm{off}}})}\right)^2\right]\|\phi\|^2_{L^\infty}.$$

We remark that

$$\sum_{j=1}^{n_{\mathrm{off}}}\left(w_j^{(k_{n_{\mathrm{off}}})}\right)^2 = \frac{{n_{\mathrm{off}}}^2}{{n_{\mathrm{on}}}^2 k^2_{n_{\mathrm{off}}}}\sum_{i_1,i_2=1}^{n_{\mathrm{on}}}\sum_{l_1,l_2=1}^{k_{n_{\mathrm{off}}}}\mathbb{1}_{\{j_{i_1}^{(l_1)}=j_{i_2}^{(l_2)}\}}$$

and that for some fixed i_1,i_2 and l_1, there exists exactly one $l_2 \in [\![1, n_{\mathrm{off}}]\!]$ such that $j_{i_1}^{(l_1)} = j_{i_2}^{(l_2)}$ as $(j_{i_2}^{(l)})_{1\leq l\leq n_{\mathrm{off}}}$ is a permutation of $[\![1, n_{\mathrm{off}}]\!]$. Therefore, there exists at most one $l_2 \in [\![1, k_{n_{\mathrm{off}}}]\!]$ verifying this property and, consequently,

$$\sum_{j=1}^{n_{\mathrm{off}}}\left(w_j^{(k_{n_{\mathrm{off}}})}\right)^2 \leq \frac{{n_{\mathrm{off}}}^2}{{n_{\mathrm{on}}}^2 k^2_{n_{\mathrm{off}}}}\sum_{i_1,i_2=1}^{n_{\mathrm{on}}}\sum_{l_1=1}^{k_{n_{\mathrm{off}}}} 1 = \frac{{n_{\mathrm{off}}}^2}{k_{n_{\mathrm{off}}}}.$$

We can then bound the second term

$$\mathbb{E}\left[\left(\frac{1}{n_{\mathrm{off}}}\sum_{j=1}^{n_{\mathrm{off}}} w_j^{(k_{n_{\mathrm{off}}})}(\psi(X'_j) - f(X'_j,\Theta_j))\right)^2\right]^{1/2} \leq \frac{4}{k^{1/2}_{n_{\mathrm{off}}}}\|\phi\|_\infty$$

and the triangle inequality gives

$$\mathbb{E}\left[\left|\mathrm{QI} - \widehat{\mathrm{QI}}^{(k_{n_{\mathrm{off}}})}_{n_{\mathrm{off}},n_{\mathrm{on}}}\right|^2\right]^{1/2} = \mathrm{O}\left(\left(\frac{1}{n_{\mathrm{on}}}\right)^{\min(1/4,1/d)} + \left(\frac{k_{n_{\mathrm{off}}}}{n_{\mathrm{off}}}\right)^{1/d} + \frac{1}{k^{1/2}_{n_{\mathrm{off}}}}\right).$$

This estimator is consistent as soon as $k_{n_{\mathrm{off}}}$ goes to infinity and $k_{n_{\mathrm{off}}}/n_{\mathrm{off}}$ goes to 0, even when μ_X has atoms. The optimal rate of growth is reached at $k_{n_{\mathrm{off}}} \sim {n_{\mathrm{off}}}^{2/(d+2)}$, leading

to

$$\mathbb{E}\left[\left|\mathrm{QI}-\widehat{\mathrm{QI}}^{(k_{n_{\mathrm{off}}})}_{n_{\mathrm{off}},n_{\mathrm{on}}}\right|^2\right]^{1/2} \leq \mathrm{O}\left({n_{\mathrm{on}}}^{-\min(1/4,1/d)}+{n_{\mathrm{off}}}^{-1/(d+2)}\right). \qquad \square$$

4.4. **Reformulation of our results in terms of Nearest Neighbors.** In this section, we do not refer to an offline and an online phase. Instead, we consider a sample $(X'_j)_{j\in[\![1,n]\!]}$ of n iid observation of law $\mu_{X'}$ and a random variable $X\sim\mu_X$ independent of the sample. We do not distinguish anymore $\phi\circ f$ from f in the regression function defined in Equation (9), that now writes

$$\psi(X)=\mathbb{E}\left[f(X,\Theta)|X\right]$$

and its Nearest Neighbor approximation of Equation (10) is

$$\widehat{\psi}^{(k)}_n(x)=\frac{1}{k}\sum_{l=1}^{k} f(X_{j^{(l)}(x)},\Theta_{j^{(l)}(x)}).$$

In Section 4.4.1, we study the case $\mu_X=\mu_{X'}$ and in Section 4.4.2 the case $\mu_X\neq\mu_{X'}$.

4.4.1. *Convergence of the Nearest Neighbor distance for non compact support.* By rewriting

$$\mathbb{E}[W_q^q(\widehat{\mu}_{\mathbf{X}_{n_{\mathrm{on}}}},\widehat{\mu}^{(1)}_{\mathbf{X}'_{n_{\mathrm{off}}}})]=\mathbb{E}[|X-\mathrm{NN}_{\mathbf{X}'_{n_{\mathrm{off}}}}(X)|^q]$$

in Theorem 3.7, and choosing $\mu_{X'}=\mu_X$, we get some asymptotic properties on the Nearest Neighbor distance

$$\mathbb{E}[|X-\mathrm{NN}_{\mathbf{X}_n}(X)|^q], \tag{22}$$

which has some application in the theoretical study of the Nearest Neighbor regressors and classifiers [3, Chapter 2]. The previous works on the topic focus mainly on the convergence when $q=2$ and assume that X have a bounded support [3, 7, 12, 14].

Some works [5, 11] consider some random variables $X$with unbounded support of in the context of the k-NN regression, *i.e.* they study the convergence of

$$\mathbb{E}\left[\left|\psi(X)-\widehat{\psi}^{(k)}_n(X)\right|\right].$$

However, they make the assumption of a bounded regression function ψ whereas, in Equation (22), we would like to take $\psi(X)=X$ and thus these results do not apply. A direct corollary from Theorem 3.7 is

Corollary 4.4. *Let X have a density p_X for which the strong minimal mass assumption 3.5 (iii) and Assumption 3.6 hold. We have*

$$\mathbb{E}[|X-\mathrm{NN}_{\mathbf{X}_n}(X)|^q]\underset{n\to+\infty}{\sim}\frac{\Gamma(1+q/d)}{v_d^{q/d}n^{q/d}}\int_{\mathbb{R}^d}p_X(x)^{1-q/d}\mathrm{d}x.$$

This extends the results of the literature by ensuring the asymptotic equivalence for some random variables with unbounded support.

4.4.2. *L^2 convergence rates of the generalization error under covariate shift.* The case $\mu_X \neq \mu_{X'}$ is also of interest in the framework of the Nearest Neighbor regression. The law of the training sample is $\mu_{X'}$ and is different from the law of the test sample μ_X, leading to the so-called covariate shift.

Theorem 4.5 (L^2 generalization error of the k-NN regression under covariate shift)**.** *Let μ_X (the law of the test sample) and $\mu_{X'}$ (the law of the training sample) verify the assumptions of Theorem 3.7, f be Lipschitz continuous in x of constant $L > 0$ uniform in Θ, and $\mathrm{Var}(f(x,\Theta)) = \mathbb{E}\left[|f(x,\Theta) - \mathbb{E}\left[f(X,\Theta)\right]|^2\right] \leq \sigma^2 < +\infty$ for all x in the support of $\mu_{X'}$. When $k_n \sim n^{2/(d+2)}$, we have*

$$\limsup_{n\to+\infty} n^{1/(2+d)}\mathbb{E}\left[\left|\psi(X) - \widehat{\psi}_n^{(k_n)}(X)\right|^2\right]^{1/2} \leq \sigma + C\mathbb{E}\left[\frac{1}{p_{X'}(X)^{2/d}}\right]^{1/2}$$

with C a positive constant.

We retrieve essentially the same orders of convergence as in the case without covariate shift. The quantity $\mathbb{E}\left[1/p_{X'}(X)^{2/d}\right]^{1/2}$ seems to be the relevant bound of the loss due to the use of $\mu_{X'}$ instead of μ_X and we expect that the greater this quantity is, the slower the convergence will be.

Proof. The proof is an adaptation of [3, Theorem 14.5], using the result of Corollary 3.8. We can decompose the L^2 error

$$\mathbb{E}\left[\left|\psi(X) - \widehat{\psi}_n^{(k)}(X)\right|^2\right]^{1/2} \leq \mathbb{E}\left[\left|\psi(X) - \widetilde{\psi}_n^{(k_n)}(X)\right|^2\right]^{1/2} + \mathbb{E}\left[\left|\widetilde{\psi}_n^{(k_n)}(X) - \widehat{\psi}_n^{(k_n)}(X)\right|^2\right]^{1/2}$$

with $\widetilde{\psi}_n^{(k_n)}(x) = \frac{1}{k_n}\sum_{i=1}^{k_n}\mathbb{E}[f(\mathrm{NN}_{\mathbf{X}'_n}^{(i)}(x),\Theta)]$. The first term can be bounded by

$$\mathbb{E}\left[\left|\psi(X) - \widetilde{\psi}_n^{(k_n)}(X)\right|^2\right]^{1/2} \leq L\mathbb{E}\left[\left|X - \mathrm{NN}_{\mathbf{X}'_n}^{(k_n)}(X)\right|^2\right]^{1/2}$$

and then

$$\limsup_{n\to+\infty}\left(\frac{n}{k_n}\right)^{1/d}\mathbb{E}\left[\left|\psi(X) - \widetilde{\psi}_n^{(k_n)}(X)\right|^2\right]^{1/2} \leq Lc_{d,2}\mathbb{E}\left[\frac{1}{p_{X'}(X)^{2/d}}\right]^{1/2}$$

by Corollary 3.8. The second term is bounded by

$$\begin{aligned}\mathbb{E}\left[\left|\widetilde{\psi}_n^{(k_n)}(X) - \widehat{\psi}_n^{(k_n)}(X)\right|^2\right]^{1/2} &= \frac{1}{k_n}\mathbb{E}\left[\sum_{i=1}^{k_n}\left(f(\mathrm{NN}_{\mathbf{X}'_n}^{(i)}(X),\Theta_{l_i}) - \mathbb{E}[f(\mathrm{NN}_{\mathbf{X}'_n}^{(i)}(X),\Theta)|X]\right)^2\right]^{1/2} \\ &\leq \frac{1}{k_n^{1/2}}\sigma\end{aligned}$$

The optimal rate is $k_n \sim n^{2/(2+d)}$, leading to

$$\limsup_{n\to+\infty} n^{1/(2+d)}\mathbb{E}\left[\left|\psi(X) - \widehat{\psi}_n^{(k_n)}(X)\right|^2\right]^{1/2} \leq \sigma + C\mathbb{E}\left[\frac{1}{p_{X'}(X)^{2/d}}\right]^{1/2},$$

with $C = Lc_{d,2}$ □

5. Numerical application

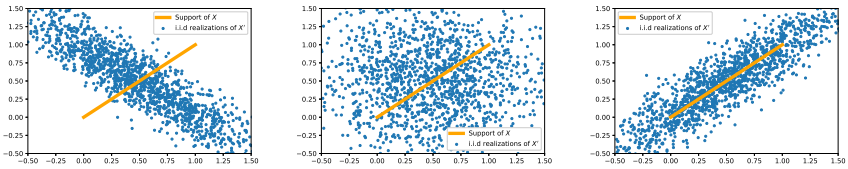
5.1. **Influence of $\mu_{X'}$ on the convergence of $\widehat{\mu}_X^{(1)}$.** We investigate how the relationship between μ_X and $\mu_{X'}$ impacts the convergence of $\widehat{\mu}^{(1)}_{\mathbf{X}'_{n_{\text{off}}}}$ presented Section 4.1 . In this numerical experiment, we set the dimension $d = 2$, choose

$$X = (U, U), \quad U \sim \mathcal{U}([0, 1]),$$

and

$$X' \sim \mathcal{N}\left(\begin{pmatrix}\mu\\ \mu\end{pmatrix}, \sigma^2 \begin{pmatrix}1 & s_{\text{corr}}\\ s_{\text{corr}} & 1\end{pmatrix}\right),$$

with $\mu = 0.5$, $\sigma = 0.3$ and various s_{corr} in $(-1, 1)$. Intuitively, the closer s_{corr} is from 1, the closer $\mu_{X'}$ is from μ_X, as illustrated in Figure 2.



(A) $s_{\text{corr}} = -0.9$ (B) $s_{\text{corr}} = 0$ (C) $s_{\text{corr}} = 0.9$

FIGURE 2. Plot of the support of X and 1500 iid realizations of X' for different values of s_{corr}.

In order to quantify the quality of the reconstruction, we estimate the measure $\widehat{\mu}^{(1)}_{\mathbf{X}'_{n_{\text{off}}}}$ by Gaussian kernel density estimation [18], *i.e.*

$$\widehat{\rho}(x_1, x_2) = K_h * \widehat{\mu}^{(1)}_{\mathbf{X}'_{n_{\text{off}}}}(x_1, x_2) = \frac{1}{n_{\text{off}}} \sum_{j=1}^{n_{\text{off}}} w_j^{(1)} K_h((x_1, x_2) - X'_j)$$

with $K_h(x_1, x_2) = \frac{1}{2\pi h^2} \exp(-(x_1^2 + x_2^2)/2h^2)$. Then, we estimate the density of the first marginal of the conditional distribution of $\widehat{\rho}$ on the support of X

$$\widehat{\rho}_{[0,1]}(x) = \frac{\widehat{\rho}(x, x)}{\int_0^1 \widehat{\rho}(u, u)\mathrm{d}u}, \quad x \in [0, 1]$$

and we compute the integrated L^2 error of this estimation with respect to the theoretical measure $\rho_{[0,1]}(x) = 1, x \in [0, 1]$

$$e_2 = \left(\int_0^1 \left(\widehat{\rho}_{[0,1]}(x) - \rho_{[0,1]}(x)\right)^2 \mathrm{d}x\right)^{1/2} = \left(\int_0^1 \left(\widehat{\rho}_{[0,1]}(x) - 1\right)^2 \mathrm{d}x\right)^{1/2}.$$

As this quantity depends on $\mathbf{X}_{n_{\text{on}}}$ and $\mathbf{X}'_{n_{\text{off}}}$, we estimate its expectation $\mathbb{E}[e_2]$.

We can see in Figure 3 that the greater s_{corr} is, the better the reconstruction looks like. This observation is confirmed in Figure 4, illustrating that $\mathbb{E}[e_2]$ decreases when s_{corr} increases, *i.e.* when the $\mu_{X'}$ gets closer to μ_X. The important amount error that is done for negative values of s_{corr} can be explained by Figures 2a and 3. Indeed, when s_{corr} is low, an observation of X' has a low probability to be drawn close to the segments $[(0, 0), (0.25, 0.25)]$ and $[(0.75, 0.75), (1, 1)]$, and thus, some values are "missed". This effect is mitigated for greater values of s_{corr} in which some observations are closer to the segments.

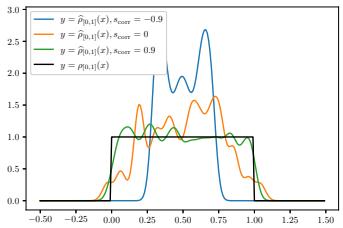


FIGURE 3. Kernel density estimation $\widehat{\rho}_{[0,1]}(x)$ for different values of s_{corr}.

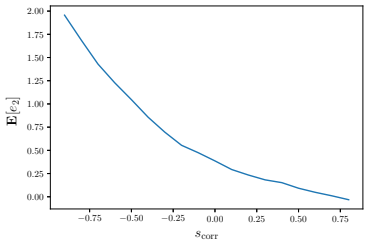


FIGURE 4. Estimation of $\mathbb{E}[e_2]$ with respect to s_{corr} by a Monte Carlo estimatio of size 500 with $n_{\mathrm{off}} = n_{\mathrm{on}} = 600$.

5.2. **Influence of $\mu_{X'}$ on the convergence if $\widehat{\mathrm{QI}}^{(k_{n_{\mathrm{off}}})}_{n_{\mathrm{off}},n_{\mathrm{on}}}$.** We now concentrate on the impact on the efficiency $\widehat{\mathrm{QI}}^{(k_{n_{\mathrm{off}}})}_{n_{\mathrm{off}},n_{\mathrm{on}}}$. We keep the framework of Section 5.1, and we try to estimate the quantity of interest

$$\mathrm{QI} = \mathbb{E}[\phi(f(X,\Theta))], \quad f((x_1,x_2),\theta) = \sin(2\pi x_1)\sin(2\pi x_2)(1+\theta)$$

with $\Theta \sim \mathcal{U}([-1,1])$ and $\phi(y) = y$. The L^2 error

$$\mathbb{E}\left[\left|\widehat{\mathrm{QI}}^{(k_{n_{\mathrm{off}}})}_{n_{\mathrm{off}},n_{\mathrm{on}}} - \mathrm{QI}\right|^2\right]^{1/2} = \mathbb{E}\left[\left|\widehat{\mathrm{QI}}^{(k_{n_{\mathrm{off}}})}_{n_{\mathrm{off}},n_{\mathrm{on}}} - 0.5\right|^2\right]^{1/2}$$

is computed by Monte Carlo estimation. As highlighted in Figure 5, the closeness of μ_X to $\mu_{X'}$ is an important factor for the efficiency of the estimator.

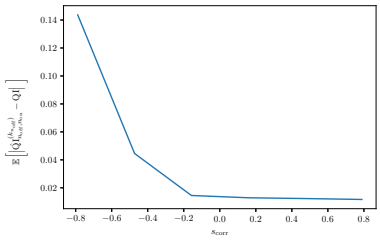


FIGURE 5. Estimation of the L^2 error with respect to s_{corr} for $n_{\mathrm{on}} = n_{\mathrm{off}} = 900$ and $k_{n_{\mathrm{off}}} = 4$ by a Monte Carlo experiment of size 2000.

Acknowledgements

The authors are thankful to Arnaud Guyader and Gérard Biau for their useful comments.

References

[1] Sergio Amaral, Douglas Allaire, and Karen Willcox. A decomposition-based approach to uncertainty analysis of feed-forward multicomponent systems. *International Journal for Numerical Methods in Engineering*, 100(13):982–1005, 2014.

[2] Sergio Amaral, Douglas Allaire, and Karen Willcox. Optimal l_2 -norm empirical importance weights for the change of probability measure. *Statistics and Computing*, 27(3):625–643, 2017.

[3] Gérard Biau and Luc Devroye. *Lectures on the nearest neighbor method*. Springer, 2015.

[4] Robert D Braun and Ilan M Kroo. Development and application of the collaborative optimization architecture in a multidisciplinary design environment. 1995.
[5] George H Chen, Devavrat Shah, et al. Explaining the success of nearest neighbor methods in prediction. *Foundations and Trends® in Machine Learning*, 10(5-6):337–588, 2018.
[6] Etienne De Rocquigny, Nicolas Devictor, and Stefano Tarantola. *Uncertainty in industrial practice: a guide to quantitative uncertainty management*. John Wiley & Sons, 2008.
[7] Dafydd Evans, Antonia J Jones, and Wolfgang M Schmidt. Asymptotic moments of near–neighbour distance distributions. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 458(2028):2839–2849, 2002.
[8] Kai-Tai Fang, Runze Li, and Agus Sudjianto. *Design and modeling for computer experiments*. CRC press, 2005.
[9] Nicolas Fournier and Arnaud Guillin. On the rate of convergence in wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738, 2015.
[10] Sébastien Gadat, Thierry Klein, Clément Marteau, et al. Classification in general finite dimensional spaces with the k-nearest neighbor rule. *The Annals of Statistics*, 44(3):982–1009, 2016.
[11] Michael Kohler, Adam Krzyżak, and Harro Walk. Rates of convergence for partitioning and nearest neighbor regression estimates with unbounded data. *Journal of Multivariate Analysis*, 97(2):311–323, 2006.
[12] Elia Liitiäinen, Amaury Lendasse, and Francesco Corona. Bounds on the mean power-weighted nearest neighbour distance. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 464(2097):2293–2301, 2008.
[13] Marco Loog. Nearest neighbor-based importance weighting. In *2012 IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6. IEEE, 2012.
[14] Mathew D Penrose and JE Yukich. Laws of large numbers and nearest neighbor distances. In *Advances in directional and linear statistics*, pages 189–199. Springer, 2011.
[15] Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. The MIT Press, 2009.
[16] Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.
[17] Alexander B Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.
[18] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.
[19] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.
[20] Wen Yao, Xiaoqian Chen, Wencai Luo, Michel van Tooren, and Jian Guo. Review of uncertainty-based multidisciplinary design optimization methods for aerospace vehicles. *Progress in Aerospace Sciences*, 47(6):450–479, 2011.

Julien Reygner
CERMICS, Ecole des Ponts, Marne-la-Vallée, France
E-mail address: julien.reygner@enpc.fr

Adrien Touboul
CERMICS, Ecole des Ponts, Marne-la-Vallée, France
IRT SystemX, Paris-Saclay, France
E-mail address: adrien.touboul@enpc.fr