Harnessing the Power of Vicinity-Informed Analysis for Classification under Covariate Shift

Mitsuhiro Fujikawa
University of Tsukuba & RIKEN AIP
Tsukuba, Ibaraki 305-8573, Japan
mitsuhiro@mdl.cs.tsukuba.ac.jp

Youhei Akimoto
University of Tsukuba & RIKEN AIP
Tsukuba, Ibaraki 305-8573, Japan
akimoto@cs.tsukuba.ac.jp

Jun Sakuma
Institute of Science Tokyo & RIKEN AIP
Meguro, Tokyo 152-8550 Japan
sakuma@c.titech.ac.jp

Kazuto Fukuchi[*]
University of Tsukuba & RIKEN AIP
Tsukuba, Ibaraki 305-8573, Japan
fukuchi@cs.tsukuba.ac.jp

Abstract

Transfer learning enhances prediction accuracy on a target distribution by leveraging data from a source distribution, demonstrating significant benefits in various applications. This paper introduces a novel dissimilarity measure that utilizes vicinity information, i.e., the local structure of data points, to analyze the excess error in classification under covariate shift, a transfer learning setting where marginal feature distributions differ but conditional label distributions remain the same. We characterize the excess error using the proposed measure and demonstrate faster or competitive convergence rates compared to previous techniques. Notably, our approach is effective in the support non-containment assumption, which often appears in real-world applications, holds. Our theoretical analysis bridges the gap between current theoretical findings and empirical observations in transfer learning, particularly in scenarios with significant differences between source and target distributions.

1 Introduction

Transfer learning is a technique for enhancing prediction accuracy by utilizing a sample from a distribution (source distribution), which is different from the distribution where predictions are actually made (target distribution). Existing empirical studies of transfer learning have shown significant accuracy improvements by leveraging a sample from the source distribution [4, 5, 8, 18, 19, 21, 24]. However, these findings are valid only for the datasets tested, leaving the effectiveness in unexplored scenarios uncertain. Theoretical analysis, on the other hand, offers broader assurances of these enhancements across various situations.

Our paper primarily focuses on theoretical analysis of classification under the covariate-shift [20] environment. Covariate-shift refers to a scenario where, despite the relationships between features and labels remaining consistent across source and target distributions, the marginal distributions of features differ. A key property in characterizing the success of transfer learning under covariate-shift is consistency with respect to the source sample size. A classification algorithm is deemed consistent with respect to source sample size if its error rate decreases to the optimal one as the size of the source sample increases indefinitely, highlighting the achievability of the optimal classifier by utilizing the source sample. The main focus of our theoretical analyses is to validate source sample-size consistency of the constructed classification algorithm under the covariate-shift.

[*] Corresponding author

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Several theoretical techniques have been developed to analyze classification error under the covariate-shift setup; however, most of them lack the capability to validate source sample size consistency. For example, many researchers have derived upper bounds on the generalization error using distance measures between source and target distributions [1, 3, 14, 15, 17]. These techniques are applicable to a broad range of situations since they do not make assumptions about the source and target distributions. However, they might fail to validate source sample size consistency because the distance measures used in these techniques may remain positive even when the source sample size tends to infinity.

Only a few theoretical results can prove the achievability of source sample size consistency. One notable result is the work by Pathak et al. [16], who deal with the nonparametric regression problem under covariate-shift and analyze the regression error using the following dissimilarity measure. Let P and Q be source and target distributions whose marginals for features are denoted as P_X and Q_X, respectively. Let $\mathcal{X}$ be the universe of features equipped with a metric ρ. Given a level $r > 0$, their dissimilarity measure is defined [2] as

$$\Delta_{\mathrm{PMW}}(P, Q; r) = \int_{\mathcal{X}} \frac{1}{P_X(B_\rho(x, r))} Q_X(dx), \tag{1}$$

where $B_\rho(x, r) = \{x' \in \mathcal{X} : \rho(x, x') \leq r\}$ is the closed ball of radius r centered at x. We use the notation $B(x, r)$ when ρ is clear from the context. Pathak et al. [16] demonstrate that a consistent regression algorithm exists if the dissimilarity measure in Eq (1) is less than a polynomial order of r^{-1}. This result can be readily extended to the classification case by utilizing the results of Galbraith et al. [7] and Kpotufe et al. [12] (See Section 3).

One significant limitation of their techniques is the inability to prove source sample-size consistency in situations where the support of the source distribution does not cotain that of the target distribution. In such situations, their dissimilarity measure in Eq (1) becomes infinite because the probability $P_X(B(x, r))$ becomes zero for small r at points x that appear in the target distribution but not in the source distribution. However, these situations are prevalent in real-world applications, and empirical evidence indicates the effectiveness of transfer learning even under a support non-containment environment. For instance, several researchers [10, 23, 25, 26] have demonstrated the success of their methods on the Office-Home dataset [22], in which source and target datasets consist of images from different domains, including artistic depictions, clipart images, images without backgrounds, and real-world images. The appearances of images across different domains are considerably different, suggesting non-containment of supports. Consequently, the current theoretical framework fails to capture the success demonstrated in this example, revealing a gap between existing theoretical results and real-world observations. This discrepancy highlights the need for a theoretical approach that can account for the effectiveness of transfer learning in scenarios where the support containment assumption does not hold.

Our dissimilarity measure and contributions. This study bridges this gap by introducing a novel dissimilarity measure and characterizing the classification error under covariate-shift using the proposed measure. Our dissimilarity measure is defined [3] as follows:

$$\Delta_{\mathcal{V}}(P, Q; r) = \int_{\mathcal{X}} \inf_{x' \in \mathcal{V}(x)} \frac{1}{P_X(B(x', r))} Q_X(dx), \tag{2}$$

where $\mathcal{V}(x)$ denotes the set of the vicinity surrounding the point x, whose rigorous definition will be explored in Section 3. The only difference between Eq (2) and Eq (1) is that Eq (2) takes the infimum over $\mathcal{V}(x)$ when evaluating the inverse probability, whereas Eq (1) evaluates the inverse probability at x. By taking the infimum, we may avoid evaluating the inverse probability at points where the probability $P_X(B(x, r))$ becomes zero. This makes the resultant dissimilarity value finite even when the support of the source distribution does not contain that of the target distribution.

The utility of our dissimilarity measure in Eq (2) is highlighted by the following contributions:

- We derive an upper bound on the excess error under covariate-shift and provide a characterization of it via the dissimilarity measure in Eq (2). A notable insight from this characterization is the existence of a classification algorithm that is consistent for the source sample size, which can validate the source sample-size consistency even under the support non-containment environment:

[2] We interpret as $\Delta_{\mathrm{PMW}}(P, Q; r) = \infty$ if $Q_X(P_X(B(X, r)) > 0) < 1$.

[3] We interpret as $\Delta_{\mathcal{V}}(P, Q; r) = \infty$ if $Q_X(\sup_{x' \in \mathcal{V}(X)} P_X(B(x', r)) > 0) < 1$.

Theorem 1 (Informal)**.** *Under certain conditions, there exists a classification algorithm that is consistent for the source sample size if $\Delta_{\mathcal{V}}(P, Q; r)$ is less than a polynomial order of r^{-1}.*
Theorem 1 provides the same characterization of the source sample size consistency as shown by Pathak et al. [16], except it uses our dissimilarity measure $\Delta_{\mathcal{V}}$ instead of their measure Δ_{PMW}.

- We propose novel notions of Δ-transfer-exponent and Δ-self-exponent for a dissimilarity measure Δ. These notions are a generalization of the concept of α-families provided by Pathak et al. [16]. Our notions of the Δ-transfer-exponent and Δ-self-exponent universally characterize the upper bounds obtained by Pathak et al. [16], Kpotufe et al. [12], and our own work, thereby enabling a fair comparison among these upper bounds. Indeed, we prove that an upper bound on the excess error derived from our dissimilarity measure in Eq (2) always exhibits faster or competitive convergence rates compared to the rates of the upper bounds obtained from the existing measures provided by Pathak et al. [16] and Kpotufe et al. [12]. This improvement in convergence rates highlights the advantage of incorporating vicinity information in the dissimilarity measure.
- We conducted experiments comparing our method with Pathak et al. [16]'s approach on synthetic datasets with support non-containment setups. The results demonstrate the tightness of our derived upper bound and showcase our method's ability to achieve source sample-size consistency in the support non-containment setting, a feat unattained by the existing method.

All the missing proofs can be found in Appendix C.

2 Preliminaries

Notations For a probability measure P and a positive integer k, let P^k denote the k-fold product measure of P. Given a probability measure P and a random variable X, we denote $\mathbb{E}_P[X]$ as the expectation of X under the distribution P. For an event $\mathcal{E}$, we use $\mathbb{1}\{\mathcal{E}\}$ to denote the indicator function. Given a metric space $(\mathcal{X}, \rho)$ and a radius $r > 0$, let denote the closed sphere centered at $x \in \mathcal{X}$ with radius r as $B(x, r) = \{x' \in \mathcal{X} : \rho(x, x') \le r\}$.

Classification under Covariate-shift Consider a classification problem under the covariate shift setup. Let X be a random variable representing the input to a classifier, equipped with a compact metric space $(\mathcal{X}, \rho)$ of diameter $D_{\mathcal{X}}$, and let Y be a random variable signifying the binary label, i.e., with a universe of $\mathcal{Y} = \{0, 1\}$. The learner has access to a sample composed of labeled data from two distributions: the source distribution P and the target distribution Q. The labeled data from the source and target distributions are denoted as $(\mathbf{X}, \mathbf{Y})_P = \{(X_i, Y_i)\}_{i=1}^{n_P} \sim P^{n_P}$ and $(\mathbf{X}, \mathbf{Y})_Q = \{(X_i, Y_i)\}_{i=n_P+1}^{n_P+n_Q} \sim Q^{n_Q}$, respectively, where n_P and n_Q represent the source and target sample sizes. Given the sample $(\mathbf{X}, \mathbf{Y}) = (\mathbf{X}, \mathbf{Y})_P \cup (\mathbf{X}, \mathbf{Y})_Q$, the learner's objective is to construct a classifier $h : \mathcal{X} \to \mathcal{Y}$ that minimizes its error rate for the target distribution, defined as:

$$err_Q(h) = \mathbb{E}_Q \mathbb{1}\{h(X) \neq Y\}.$$

For convenience, let $\mathcal{X}_P$ be the support of P_X, i.e., $\mathcal{X}_P = \{x \in \mathcal{X} : P_X(B(x, r)) > 0, \forall r > 0\}$. Define $\mathcal{X}_Q$ similarly to $\mathcal{X}_P$.

Covariate-shift is a relationship between the source and target distributions, in which the marginal distributions of the input X can differ between P and Q, whereas the distributions of the label Y conditioned on the input X are identical. Let P_X and Q_X be the marginal source and target distributions of X, respectively. Let $P_{Y|X}$ and $Q_{Y|X}$ be the source and target distributions of Y conditioned on X, respectively. Then, covariate shift is rigorously defined as follows:

Definition 1 (Covariate-shift)**.** *The relationship between distribution P and distribution Q is covariate shift if there exists a measurable function $\eta : \mathcal{X} \to [0, 1]$, called a regression function, such that $P_{Y|X}(Y = 1|X) = Q_{Y|X}(Y = 1|X) = \eta(X)$ P_X- and Q_X-almost surely.*

This definition indicates that, for example, traffic signs appearing in urban and rural areas may differ ($P_X \neq Q_X$), but their instructions are consistent regardless of the location ($P_{Y|X} = Q_{Y|X}$) in the context of sign recognition in an automated driving system.

Excess Error The objective of our theoretical analyses is to elucidate the relationship between the source and target sample sizes (n_P and n_Q) and the *excess error*. The excess error of a classifier h is defined as the difference between the error of h and the error incurred by the *Bayes classifier* h^*. The Bayes classifier, under the error metric $err_Q(h)$, is the classifier that minimizes $err_Q(h)$. The formal definition of the excess error is as follows:

Definition 2 (excess error)**.** *The excess error of the classifier h for the distribution Q is given by:*

$$\mathcal{E}_Q(h) = err_Q(h) - err_Q(h^*).$$

As the excess error approaches 0, the classifier h approaches the performance of the ideal classifier. Under our setup, the Bayes classifier can be expressed as $h^*(x) = \mathbb{1}\{\eta(x) \geq 1/2\}$.

Difficulty in Classification under Distribution Q For the purpose of our analyses, we introduce the following common assumptions that stipulate the difficulty in classification under distribution Q.

Definition 3 (Smoothness)**.** *A regression function η is (C_α, α)-Hölder for $\alpha \in (0, 1]$ and $C_\alpha > 0$ if $\forall x, x' \in \mathcal{X}, |\eta(x) - \eta(x')| \leq C_\alpha \cdot \rho(x, x')^\alpha$.*

Definition 4 (Tsybakov's noise condition)**.** *A distribution Q satisfies Tsybakov's noise condition with parameters $\beta > 0$ and $C_\beta > 0$ if $\forall t \geq 0, Q_X(0 < |\eta(X) - \frac{1}{2}| \leq t) \leq C_\beta t^\beta$.*

The smoothness condition in Definition 3 requires that the labels for similar inputs are likely to be the same. The noise parameters determine the probability of observing a label with a large amount of noise. It is worth noting that Galbraith et al. [7] and Kpotufe et al. [12] conducted their analyses under the same assumptions. Similarly, Pathak et al. [16] employ assumptions regarding smoothness and noise; however, their assumptions differ slightly from ours, as they address a different problem: regression, while we focus on classification.

Our analyses will be conducted under the assumption that the target distribution satisfies both the smoothness and noise conditions.

Definition 5 ($\mathrm{STN}(\alpha, \beta)$)**.** *A distribution Q is $\mathrm{STN}(\alpha, \beta)$ if there exist some constants $C_\alpha > 0$ and $C_\beta > 0$ such that the regression function η is (C_α, α)-Hölder, and Q satisfies Tsybakov's noise condition with parameters β and C_β.*

3 Main Result

Our main result is a characterization of the excess error under the covariate-shift setup via our dissimilarity measure in Eq (2), with an appropriate choice of the vicinity set function $\mathcal{V}(x)$. The choice of the vicinity set function $\mathcal{V}$ is formally given as follows:

$$\mathcal{V}(x) = \left\{ x' \in \mathcal{X} : 2C_\alpha \rho(x, x')^\alpha < \left| \eta(x) - \frac{1}{2} \right| \right\} \cup \{x\}. \tag{3}$$

The vicinity set $\mathcal{V}(x)$ is the (nearly-)largest open ball centered at x such that the labels of the Bayes classifier evaluated at points within the ball are consistent. We can expect that these vicinity points may share the same label information and thus are useful for predicting the label at x.

To characterize the excess error by some quantity of (P, Q), we generalize the notion of the α-family proposed by Pathak et al. [16]. Specifically, we characterize the excess error by the following quantities determined by a dissimilarity measure.

Definition 6 (Δ-transfer-exponent)**.** *Given a dissimilarity measure Δ, a distribution pair (P, Q) has a Δ-transfer-exponent of $\tau \in [0, \infty]$ if there exists a constant $C \geq 1$ such that*

$$\sup_{0 < r \leq D_\mathcal{X}} (r/D_\mathcal{X})^\tau \Delta(P, Q; r) \leq C,$$

where $0 \cdot \Delta(P, Q; 0) = 0$.

Definition 7 (Δ-self-exponent)**.** *Given a dissimilarity measure Δ, a distribution Q has a Δ-self-exponent of $\psi \in (0, \infty]$ if there exists a constant $C \geq 1$ such that*

$$\sup_{0 < r \leq D_\mathcal{X}} (r/D_\mathcal{X})^\psi \Delta(Q, Q; r) \leq C.$$

Definition 6 and Definition 7 imply that the dissimilarities $\Delta(P, Q; r)$ and $\Delta(Q, Q; r)$ decrease at a polynomial rate with respect to r^{-1}, with exponents τ and ψ, respectively. In other words, $\Delta(P, Q; r) = O(r^{-\tau})$ and $\Delta(Q, Q; r) = O(r^{-\psi})$ for a decreasing r. It is worth noting that our definitions of Δ-transfer-exponent and Δ-self-exponent are universal in the sense that we can exactly reproduce the quantities used in existing characterizations by choosing an appropriate dissimilarity measure Δ, which will be discussed later.

As our characterization, we provide an upper bound on the excess error composed of the source and target sample sizes as well as transfer- and self-exponents.

Theorem 2. *Given $\alpha \in (0,1]$, $\beta > 0$, and $\psi \in (0,\infty]$, suppose the target distribution Q is $\mathrm{STN}(\alpha,\beta)$ and has $\Delta_{\mathcal{V}}$-self-exponent of ψ. Also, suppose (P,Q) has $\Delta_{\mathcal{V}}$-transfer-exponent of τ for some $\tau \in (0,\infty]$. Then, there exists a classification algorithm which produces a classifier $\hat{h}$ such that for all $n_P > 0$ and $n_Q > 0$,*

$$\mathbb{E}\left[\mathcal{E}_Q(\hat{h})\right] \le C \begin{cases} \log(n_P + n_Q)\left(n_p^{\frac{1+\beta}{2+\beta+\max\{1,\tau/\alpha\}}} + n_Q^{\frac{1+\beta}{2+\beta+\max\{1,\psi/\alpha\}}}\right)^{-1} & \text{if } \alpha = \tau \text{ or } \alpha = \psi, \\ \left(n_p^{\frac{1+\beta}{2+\beta+\max\{1,\tau/\alpha\}}} + n_Q^{\frac{1+\beta}{2+\beta+\max\{1,\psi/\alpha\}})}\right)^{-1} & \text{otherwise} , \end{cases}$$

where $C > 0$ is some constant independent of n_P and n_Q.

The implications of Theorem 2 are as follows:

1. Theorem 2 directly establishes that the necessary condition for the existence of a source sample size consistent classification algorithm is $\tau < \infty$. In this case, the exponent of n_P is non-zero, indicating the algorithm's consistency with respect to the source sample size.
2. In the non-transfer setting, the excess error decreases as the sample size increases, with an exponent of $-\frac{1+\beta}{2+\beta+d/\alpha}$ for d-dimensional input, i.e., $\mathcal{X} \subset \mathbb{R}^d$ [2]. Our bound exhibits the same characterization, except that the dimensionality d is replaced by the $\Delta_{\mathcal{V}}$-transfer- or $\Delta_{\mathcal{V}}$-self-exponent, corresponding to n_P or n_Q, respectively. Indeed, the $\Delta_{\mathcal{V}}$-self-exponent plays a role similar to the dimensionality d, as it is smaller than d for $\mathcal{X} \subset \mathbb{R}^d$ [4].
3. The $\Delta_{\mathcal{V}}$-transfer- and $\Delta_{\mathcal{V}}$-self-exponents characterize the dependency of the excess error on the source and target sample sizes, respectively. Indeed, the convergence rate of the excess error for n_P (resp., n_Q) becomes faster as the $\Delta_{\mathcal{V}}$-transfer-exponent (resp., $\Delta_{\mathcal{V}}$-self-exponent) decreases.

Comparisons with Pathak et al. [16] and Kpotufe et al. [12]. We explore the comparison with the excess error upper bounds shown by Pathak et al. [16] and Kpotufe et al. [12]. As mentioned in the introduction, Pathak et al. [16] provide a characterization of the excess error through Δ_{PMW} in Eq (1). We can reproduce the results of Kpotufe et al. [12] via the transfer- and self-exponents of the following measures:

$$\Delta_{\mathrm{DM}}(Q,Q;r) = \sup_{x \in \mathcal{X}_Q} \frac{1}{Q_X(B(x,r))}, \Delta_{\mathrm{BCN}}(Q,Q;r) = \mathcal{N}(\mathcal{X}_Q,\rho,r),$$
$$\Delta_{\mathrm{KM}}(P,Q;r) = \sup_{x \in \mathcal{X}_Q} \frac{Q_X(B(x,r))}{P_X(B(x,r))},$$

where $\mathcal{N}(\mathcal{X}_Q,\rho,r)$ denotes the r-covering number of the set $\mathcal{X}_Q$. Building upon the measures Δ_{PMW}, Δ_{DM}, Δ_{BCN}, and Δ_{KM}, their upper bounds are reproduced as follows:

Proposition 1 (Kpotufe et al. [12] and Pathak et al. [16])**.** *Given $\alpha \in (0,1]$ and $\beta > 0$, suppose the target distribution Q is $\mathrm{STN}(\alpha,\beta)$. For $\psi \in (0,\infty]$ and $\tau \in (0,\infty]$, we suppose that the one of the following conditions holds:*

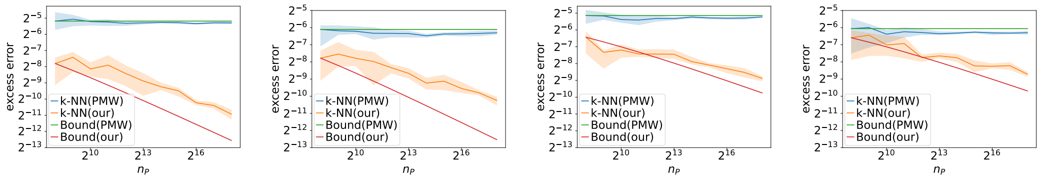
1. *Q has the Δ_{PMW}-self-exponent of ψ, and (P,Q) has Δ_{PMW}-transfer-exponent of τ.*
2. *Q has the Δ_{DM}- or Δ_{BCN}-self-exponent of ψ, (P,Q) has Δ_{KM}-transfer-exponent of $\tau - \psi$, and $\tau \ge \psi$.*

Then, there exists an algorithm that exhibits the excess error upper bound obtained by Theorem 2.

Proposition 1 indicates that our bound in Theorem 2 coincides with theirs, except for using the self- and transfer-exponents with their measures.

Next, we compare the self- and transfer-exponents between our and their measures.

[4] This discussion is valid only when $\mathcal{X}$ is bounded. Exploring the unbounded case is one of our future directions.



(a) $\alpha = \frac{1}{2}, \tau = 1$ (b) $\alpha = \frac{1}{4}, \tau = 1$ (c) $\alpha = \frac{1}{2}, \tau = 2$ (d) $\alpha = \frac{1}{4}, \tau = 2$

Figure 1: Excess error $\mathcal{E}_Q$ vs. source sample size n_P. The horizontal axis represents the source sample size n_P, and the vertical axis represents the excess error. Both axes are plotted on a logarithmic scale, resulting in the slopes of the lines signifying the exponents of n_P for each method. ”k-NN (our)“ and ”k-NN (PMW)“ denote the experimental results of k-NN classifiers with our and Pathak et al. [16]'s parameter settings, respectively. For guidance, we also plot the lines, ”Bound (our)“ and ”Bound (PMW)“, representing the upper bounds obtained by this paper and Pathak et al. [16], respectively.

Proposition 2. *For any pair of distributions* (P, Q), *we have*

$$\begin{aligned} \tau_{\Delta_{\mathcal{V}}} \leq & \tau_{\Delta_{\mathrm{PMW}}} \leq \tau_{\Delta_{\mathrm{KM}}} + \min\{\psi_{\Delta_{\mathrm{DM}}}, \psi_{\Delta_{\mathrm{BCN}}}\}, \\ \psi_{\Delta_{\mathcal{V}}} \leq & \psi_{\Delta_{\mathrm{PMW}}} \leq \qquad\quad \min\{\psi_{\Delta_{\mathrm{DM}}}, \psi_{\Delta_{\mathrm{BCN}}}\}, \end{aligned}$$

where τ_Δ *and* ψ_Δ *denotes the minimum* Δ*-transfer- and* Δ*-self-exponents* (P, Q) *has.*

Proposition 2 showcases that $\Delta_{\mathcal{V}}$ achieves the smallest transfer- and self-exponents, indicating that our measure can provide faster rates than those obtained by the existing measures.

4 Experiment

To confirm the tightness of Theorem 2 and the source sample-size consistency under the support non-containment environment, we carried out experiments on a synthetic dataset.

Data distribution. Let $\mathcal{X} = \mathbb{R}$. For $\tau > 0$, P_X has a density function proportional to $(1 - x^2)^{-\tau/2}$ supported on $[-\frac{8^{\frac{1}{\alpha}}\cdot 2-1}{8^{\frac{1}{\alpha}}\cdot 2}, \frac{8^{\frac{1}{\alpha}}\cdot 2-1}{8^{\frac{1}{\alpha}}\cdot 2}]$. Q_X is the uniform distribution over $[-1, 1]$, indicating that $\mathcal{X}_Q \not\subseteq \mathcal{X}_P$. Given $\alpha > 0$, the regression function η is $\eta(x) = \frac{1}{2} + \frac{1}{2}\,\mathrm{sgn}(x)|x|^\alpha$. With this setup, Q is $\mathrm{STN}(\alpha, \beta)$ with $\beta = 1/\alpha$. The self-exponents are equivalent, i.e., $\psi_{\Delta_{\mathcal{V}}} = \psi_{\Delta_{\mathrm{PMW}}} = 1$. Due to the support non-containment, the Δ_{PMW}-transfer-exponent is ∞. On the other hand, the $\Delta_{\mathcal{V}}$-transfer-exponent is τ.

Setup. We evaluated the dependency of the source sample size n_P on the excess error for k-NN classifiers with our and Pathak et al. [16]'s parameter settings. Specifically, we fixed $n_Q = 10$ and evaluated the excess errors with $n_P \in \{2^8, 2^9, ..., 2^{18}\}$. We varied the parameters α and τ as $\alpha \in \{\frac{1}{2}, \frac{1}{4}\}$ and $\tau \in \{1, 2\}$. For each parameter, we report the average and the first and third quartiles of the excess error over 10 runs. The detailed setup can be found in Appendix D.

Results. Figure 1 shows the log-log plots of the excess errors corresponding to the source sample sizes for each α and τ. For both our and Pathak et al. [16]'s k-NN, the slopes of the excess errors match the corresponding theoretical upper bounds for any parameter of α and τ, demonstrating the tightness of the upper bounds across various situations. In Figure 1, the line representing Pathak et al. [16]'s k-NN does not decrease, indicating a failure to achieve source sample-size consistency. In contrast, our k-NN exhibits a decreasing excess error, signifying the successful achievement of source sample-size consistency.

5 Conclusion

In this paper, we provide a novel analysis of excess error under the covariate-shift setup, demonstrating the usefulness of our new dissimilarity measure that utilizes vicinity information. Unlike existing analyses, our results can validate the consistency of the source sample size under certain situations where the support of the source sample does not contain that of the target distribution. We also demonstrate that our dissimilarity measure can provide faster rates than those provided by existing techniques, including [12, 16].

Border imacts and limitations There might be no additional societal impacts from those of the standard classification, as our focus is to leverage the multiple samples following different distributions to improve the classification accuracy.

Acknowledgments and Disclosure of Funding

This work was partly supported by JSPS KAKENHI Grant Numbers JP23K13011, JST CREST JPMJCR21D3, and JSPS Grants-in-Aid for Scientific Research 23H00483.

References

[1] 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*, pages 7433–7449. PMLR, 2022.

[2] Jean-Yves Audibert and Alexandre B. Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633, 2007. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/009053606000001217.

[3] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1):151–175, 2010. ISSN: 1573-0565. DOI: 10.1007/s10994-009-5152-4.

[4] Wenyuan Dai, Qiang Yang, Gui-Rong Xue, and Yong Yu. Boosting for transfer learning. In *Proceedings of the 24th International Conference on Machine learning*, pages 193–200, New York, NY, USA. Association for Computing Machinery, 2007. ISBN: 978-1-59593-793-3. DOI: 10.1145/1273496.1273521.

[5] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. DeCAF: a deep convolutional activation feature for generic visual recognition. In *Proceedings of the 31st International Conference on Machine Learning*, pages 647–655. PMLR, 2014.

[6] Xingdong Feng, Xin He, Caixing Wang, Chao Wang, and Jingnan Zhang. Towards a unified analysis of kernel-based methods under covariate shift. In *Advances in Neural Information Processing Systems*, volume 36, pages 73839–73851, 2023.

[7] Nicholas R. Galbraith and Samory Kpotufe. Classification tree pruning under covariate shift. *IEEE Transactions on Information Theory*, 70(1):456–481, 2024. ISSN: 1557-9654. DOI: 10.1109/TIT.2023.3308914.

[8] Tom Ginsberg, Zhongyuan Liang, and Rahul G. Krishnan. A learning based hypothesis test for harmful covariate shift. In *The Eleventh International Conference on Learning Representations*, 2022.

[9] László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer Series in Statistics. Springer, New York, NY, 2002. ISBN: 978-0-387-95441-7. DOI: 10.1007/b97848.

[10] Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool. MIC: masked image consistency for context-enhanced domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11721–11732, 2023.

[11] Samory Kpotufe. Lipschitz density-ratios, structured data, and data-driven tuning. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 1320–1328. PMLR, 2017.

[12] Samory Kpotufe and Guillaume Martinet. Marginal singularity and the benefits of labels in covariate-shift. *The Annals of Statistics*, 49(6):3299–3323, 2021. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/21-AOS2084.

[13] 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. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/23-AOS2268.

[14] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: learning bounds and algorithms, 2023. arXiv: 0902.3430.

[15] Sangdon Park, Osbert Bastani, James Weimer, and Insup Lee. Calibrated prediction with covariate shift via unsupervised domain adaptation. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pages 3219–3229. PMLR, 2020.

[16] Reese Pathak, Cong Ma, and Martin Wainwright. A new similarity measure for covariate shift with applications to nonparametric regression. In *Proceedings of the 39th International Conference on Machine Learning*, pages 17517–17530. PMLR, 2022.

[17] Yangjun Ruan, Yann Dubois, and Chris J. Maddison. Optimal representations for covariate shift. In *International Conference on Learning Representations*, 2021.

[18] Nicolas Schreuder and Evgenii Chzhen. Classification with abstention but without disparities. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 1227–1236. PMLR, 2021.

[19] Xiaoxiao Shi, Wei Fan, and Jiangtao Ren. Actively transfer domain knowledge. In Walter Daelemans, Bart Goethals, and Katharina Morik, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 342–357, Berlin, Heidelberg. Springer, 2008. ISBN: 978-3-540-87481-2. DOI: 10.1007/978-3-540-87481-2_23.

[20] 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. ISSN: 0378-3758. DOI: 10.1016/S0378-3758(00)00115-4.

[21] Yongduo Sui, Qitian Wu, Jiancan Wu, Qing Cui, Longfei Li, Jun Zhou, Xiang Wang, and Xiangnan He. Unleashing the power of graph data augmentation on covariate distribution shift. In *Advances in Neural Information Processing Systems*, volume 36, pages 18109–18131, 2023.

[22] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5385–5394, 2017. DOI: 10.1109/CVPR.2017.572.

[23] Thomas Westfechtel, Dexuan Zhang, and Tatsuya Harada. Combining inherent knowledge of vision-language models with unsupervised domain adaptation through self-knowledge distillation, 2023. arXiv: 2312.04066.

[24] Yu-Jie Zhang, Zhen-Yu Zhang, Peng Zhao, and Masashi Sugiyama. Adapting to continuous covariate shift via online density ratio estimation. In *Advances in Neural Information Processing Systems*, volume 36, pages 29074–29113, 2023.

[25] Wenlve Zhou and Zhiheng Zhou. Unsupervised domain adaption harnessing vision-language pre-training. *IEEE Transactions on Circuits and Systems for Video Technology*:1–1, 2024. ISSN: 1558-2205. DOI: 10.1109/TCSVT.2024.3391304.

[26] Jinjing Zhu, Haotian Bai, and Lin Wang. Patch-mix transformer for unsupervised domain adaptation: a game perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3561–3571, 2023.

A Example

We demonstrate that, unlike existing measures, our dissimilarity measure can validate the source sample size consistency even when $\mathcal{X}_Q \not\subseteq \mathcal{X}_P$. We provide a concrete example of P and Q to illustrate this property. Consider the case where $\mathcal{X} = \mathbb{R}$. Suppose P_X and Q_X are uniform distributions over $[-\frac{7}{8}, \frac{7}{8}]$ and $[-1, 1]$, respectively, and the regression function is $\eta(x) = \frac{1}{2}x + \frac{1}{2}$. It is clear that $\mathcal{X}_P \subset \mathcal{X}_Q$, and hence $\mathcal{X}_Q \not\subseteq \mathcal{X}_P$. The self-exponents are equivalent, i.e., $\psi_{\Delta_{\mathcal{V}}} = \psi_{\Delta_{\mathrm{PMW}}} = \psi_{\Delta_{\mathrm{DM}}} = \psi_{\Delta_{\mathrm{BCN}}} = 1$. However, the probability $P_X(B(x, r))$ takes zero at $x \in [-1, -\frac{7}{8} - r) \cup (\frac{7}{8} + r, 1]$ for a small r, causing the existing transfer-exponents to become infinite, i.e., $\tau_{\Delta_{\mathrm{PMW}}} = \tau_{\Delta_{\mathrm{KM}}} = \infty$. In contrast, our measure satisfies $\tau_{\Delta_{\mathcal{V}}} = 1$ because $\mathcal{V}(x) \cap [-\frac{7}{8}, \frac{7}{8}]$ is non-empty for any $x \in [-1, 1]$, and the probability $P_X(B(x, r))$ is non-zero for any $x \in [-\frac{7}{8}, \frac{7}{8}]$. Consequently, our bound exhibits the rate of $\ln(n_P + n_Q)(n_P^{1/2} + n_Q^{1/2})^{-1}$, as $\alpha = 1$ and $\beta = 1$ in this case, achieving the source sample size consistency.

B Analyses

To prove Theorem 2, we provide an upper bound on the excess error of a specific classification algorithm, the k-nearest neighbor (k-NN) classifier proposed by Kpotufe et al. [12]. Given a point X distributed according to Q_X for which the label will be predicted, the k-NN classifier first estimates the regression function's output at X, denoted as $\eta(X)$, by computing the average of labels over the k nearest neighbor points in $(\mathbf{X}, \mathbf{Y})$. The predicted label is then determined to be 1 if the estimated value of $\eta(X)$ is greater than $1/2$ and 0 otherwise. Formally, let $(X_{(1)}, Y_{(1)}), ..., (X_{(k)}, Y_{(k)})$ be the k nearest neighbors of X and their corresponding labels. The estimated regression function is given by $\hat{\eta}_k(X) = \frac{1}{k}\sum_{i=1}^{k} Y_{(i)}$, and the predicted label $\hat{h}_k(X)$ is determined as $\hat{h}_k(X) = \mathbb{1}\{\hat{\eta}_k(X) \geq 1/2\}$.

The goal of this section is to demonstrate that the upper bound shown in Theorem 2 is achievable by the k-NN classifier with an appropriate choice of k.

Theorem 3. *Under the same assumptions as in Theorem 2, the k-NN classifier with $k = \left\lfloor (n_P^{\frac{1}{2+\beta+\max\{1,\tau/\alpha\}}} + n_Q^{\frac{1}{2+\beta+\max\{1,\psi/\alpha\}}})^2 \right\rfloor$ achieves the excess error upper bound shown in Theorem 2.*

The main challenge in proving Theorem 3 is linking the excess error of the k-NN classifier to the minimum inverse probability that appears in our dissimilarity measure in Eq (2). To achieve this, we derive an upper bound on the excess error of the k-NN classifier using the *vicinity distance*, defined as, for $z, x \in \mathcal{X}$,

$$\rho_{\mathcal{V}}(z, x) = \inf_{x' \in \mathcal{V}(x)} \rho(z, x').$$

The vicinity distance $\rho_{\mathcal{V}}$ characterizes the minimum inverse probability, as it can be rewritten as $\inf_{x' \in \mathcal{V}(x)} P_X^{-1}(B(x', r)) = P_X^{-1}(\inf_{x' \in \mathcal{V}(x)} \rho(X, x') \leq r) = P_X^{-1}(\rho_{\mathcal{V}}(X, x) \leq r)$. Therefore, characterizing the excess error using $\rho_{\mathcal{V}}$ is crucial for revealing its connection to our dissimilarity measure $\Delta_{\mathcal{V}}$. The details of this analysis will be explored in the subsequent subsection.

B.1 Bounding Excess Error by Vicinity Distance

This subsection aims to derive an upper bound on the excess error of the k-NN classifier using the vicinity distance $\rho_{\mathcal{V}}$. We first employ two existing techniques from Kpotufe et al. [12]: an upper bound on the excess error by the approximation error of $\hat{\eta}_k$ and the concept of implicit 1-NNs. Then, we show that the approximation error of $\hat{\eta}_k$ is bounded above by the expected vicinity distance between an implicit 1-NN and a point to be predicted.

Bounding via the approximation error of $\hat{\eta}_k$. We construct an upper bound on the excess error of the k-NN classifier using the approximation error of the estimated regression function $\hat{\eta}_k$. Define $g(X) = |\eta(X) - \frac{1}{2}|$. For a random variable Z (possibly) depending on X and $(\mathbf{X}, \mathbf{Y})$, define $\Phi(Z) = 2\mathbf{E}[g(X)\mathbb{1}\{Z \geq g(X)\}]$. Then, we bound the excess error of $\hat{h}_k$ as

$$\mathbf{E}\left[\mathcal{E}_Q(\hat{h}_k)\right] \leq \Phi(|\hat{\eta}_k(X) - \eta(X)|). \tag{4}$$

Eq (4) indicates that a smaller approximation error results in a smaller excess error.

Implicit 1-NNs and implicit vicinity 1-NNs. Implicit 1-NNs, introduced by Györfi et al. [9], are a crucial technique for analyzing the k-NN classifier. Given a point X to be predicted, the implicit 1-NNs in the transfer learning setup are the 1-NNs of X within k disjoint batches consisting of subsamples of $(\mathbf{X}, \mathbf{Y})_P$ and $(\mathbf{X}, \mathbf{Y})_Q$ with sizes $\lfloor \frac{n_P}{k} \rfloor$ and $\lfloor \frac{n_Q}{k} \rfloor$, respectively. Let $B_1, ..., B_k$ be the sets of X_i appearing in the ith batch. The ith implicit 1-NN is defined as $X_i^* = \arg\min_{X^* \in B_i} \rho(X^*, X)$ for $i = 1, ..., k$. Implicit 1-NNs behave similarly to k-NNs but are mutually independent, allowing, e.g., the use of concentration inequalities for independent random variables.

We also leverage the concept of implicit 1-NNs, but we employ the vicinity distance $\rho_{\mathcal{V}}$ instead of using the standard distance ρ, which we refer to as *implicit vicinity 1-NNs*. With the same definition of batches $B_1, ..., B_k$, the implicit vicinity 1-NNs are defined as $\tilde{X}^i = \arg\min_{X^* \in B_i} \rho_{\mathcal{V}}(X^*, X)$ for $i = 1, ..., k$.

Bounding via $\rho_{\mathcal{V}}$. We show an upper bound on the approximation error of $\hat{\eta}_k$ using the vicinity distance with the implicit vicinity 1-NNs, providing an upper bound on the excess error due to Eq (4).

Theorem 4. *Given $\alpha \in (0, 1]$ and $\beta > 0$, suppose the target distribution Q is $\mathrm{STN}(\alpha, \beta)$ with constants C_α and C_β. Then, there exist constant $C > 0$ and $c > 0$ (possibly) depending on α and β such that for all $k > 1$ and for all $t > 0$, with probability, taken over the randomness of $(\mathbf{X}, \mathbf{Y})$, at least $1 - Ce^{-ckt^2}$,*

$$|\hat{\eta}_k(X) - \eta(X)| \leq C_\alpha \mathbf{E}\left[\rho_{\mathcal{V}}^\alpha\left(\tilde{X}^1, X\right)\middle| X\right] + \frac{1}{2}g(X) + t,$$

almost surely for the randomness of X.

Kpotufe et al. [12] provide a similar bound to Theorem 4, but with the expected distance $C_\alpha \mathbf{E}[\rho^\alpha(X_1^*, X)|X]$ instead of $C_\alpha \mathbf{E}[\rho_{\mathcal{V}}^\alpha(\tilde{X}_1^*, X)|X] + \frac{1}{2}g(X)$. To achieve source sample-size consistency, the expected distance needs to vanish as the source sample size tends to infinity. However, under source and target distributions without support containment, the distance $\rho(X_1^*, X)$ is larger than a non-zero positive constant. In contrast, the vicinity distance $\rho_{\mathcal{V}}(\tilde{X}_1^*, X)$ can vanish as it takes the infimum over the vicinity set.

B.2 Bounding via Dissimilarity Measure

We now derive a high probability upper bound on the distance between the implicit (vicinity) 1-NN and the point to be predicted. To obtain this upper bound, we utilize a part of the analysis conducted by Pathak et al. [16].

Theorem 5. *Given a distance ρ over $\mathcal{X}$, define $\Delta(P, Q; r) = \int_{\mathcal{X}} \frac{1}{P_X(B_\rho(x,r))} Q(dx)$. Then, for $t > 0$,*

$$\mathbf{E}\left[\mathbb{1}\left\{\min_{X^* \in B_1} \rho(X^*, X) > t\right\}\right] \leq \min\left\{\frac{\Delta(P, Q; t)}{\lfloor n_P/k \rfloor}, \frac{\Delta(Q, Q; t)}{\lfloor n_Q/k \rfloor}\right\},$$

where the expectation is taken over the randomness of $(\mathbf{X}, \mathbf{Y})$ and X.

By applying Theorem 5 with $\rho = \rho_{\mathcal{V}}$, we obtain a high probability upper bound on $\rho_{\mathcal{V}}\left(\tilde{X}_1^*, X\right)$ using our dissimilarity measure $\Delta_{\mathcal{V}}$. This result is essential for establishing the connection between the excess error of the k-NN classifier and our dissimilarity measure.

B.3 Sketch Proof of Theorem 3

Theorem 3 is validated by combining Theorem 4, Theorem 5, and Eq (4). For simplicity, we only prove the case where $\tau, \psi > \alpha$ and left the other cases to Appendix C. For a constant $\epsilon > 0$, define $A(\epsilon, X) = \int_\epsilon^{D_{\mathcal{X}}} \mathbf{E}[\mathbb{1}\{C_\alpha \rho_{\mathcal{V}}^\alpha\left(\tilde{X}_1^*, X\right) \geq t\}|X]dt$. Then, $C_\alpha \mathbf{E}\left[\rho_{\mathcal{V}}^\alpha(\tilde{X}_1^*, X)\middle| X\right] \leq \epsilon + A(\epsilon, X)$. Hence, from Eq (4) and Theorem 4, there exists a random variable $\xi \geq 0$ depending on $(\mathbf{X}, \mathbf{Y})$ and X such that conditioned on X, $\xi \leq t$ with probability at least $1 - Ce^{-ckt^2}$, and

$$\mathbf{E}\left[\mathcal{E}_Q(\hat{h}_k)\right] \leq \Phi(2(\epsilon + A(\epsilon, X) + \xi)). \tag{5}$$

Under the assumptions of $\Delta_{\mathcal{V}}$-transfer- and $\Delta_{\mathcal{V}}$-self-exponents, applying Theorem 5 and adopting an approach similar to Kpotufe et al. [12], we obtain that for some constant $C > 0$,

$$\mathbf{E}\left[\mathcal{E}_Q(\hat{h}_k)\right] \leq C\left(\epsilon^{1+\beta} + \min\left\{\frac{\epsilon^{-\frac{\tau}{\alpha}+1}}{\lfloor n_P/k \rfloor}, \frac{\epsilon^{-\frac{\psi}{\alpha}+1}}{\lfloor n_Q/k \rfloor}\right\} + k^{-\frac{1+\beta}{2}}\right), \tag{6}$$

where the three terms in Eq (6) are bounds for the three terms in Eq (5), respectively. To achieve the rate in Theorem 2, we set $\epsilon = c\min\{\lfloor \frac{n_P}{k} \rfloor^{-\frac{1}{\beta + \frac{\tau}{\alpha}}}, \lfloor \frac{n_Q}{k} \rfloor^{-\frac{1}{\beta + \frac{\psi}{\alpha}}}\}$ for some constant $c > 0$ and assign k as specified in the theorem statement.

C Missing Proofs

C.1 Proof of Proposition 1 and Theorem 3

With the choice of k shown in the statement of Theorem 3, there is a universal constant $c > 0$ such that for $n_P \geq c$ and $n_Q \geq c$, $n_P \geq 2k$ and $n_Q \geq 2k$. We can verify the upper bound in Theorem 2 holds when $n_P < c$ or $n_Q < c$ by adjusting the multiplicative constant, as the upper bound in Theorem 2 is decreasing in n_P and n_Q. Therefore, we assume $n_P \geq 2k$ and $n_Q \geq 2k$ in the subsequent analyses.

The most parts of the proofs of Proposition 1 and Theorem 3 are overlapped. As a non-overlapped part, we first demonstrate that in both cases of Proposition 1 and Theorem 3, we can validate that for a distance $\bar{\rho}$, which is either $\rho_{\mathcal{V}}$ or ρ, there exists a random variable $\xi \geq 0$ depending on $(\mathbf{X}, \mathbf{Y})$ and X such that conditioned on X, $\xi \leq t$ with probability at least $1 - e^{-ckt^2}$, and

$$\mathbf{E}\left[\mathcal{E}_Q(\hat{h}_k)\right] \leq \mathbf{E}\left[2g(X)\mathbb{1}\left\{g(X) \leq C\left(C_\alpha \mathbf{E}\left[\min_{X^* \in B_i} \bar{\rho}^\alpha(X^*, X)\middle|X\right] + \xi\right)\right\}\right], \tag{7}$$

where $C > 0$ is a universal constant.

To prove Eq (7) in the case of Proposition 1, we utilize the result by Kpotufe et al. [12]. Kpotufe et al. [12] reveal that the approximation error of $\hat{\eta}_k$ can be bounded above by the expected distance between an implicit 1-NN and X.

Theorem 6 (Kpotufe et al. [12])**.** *Given $\alpha \in (0, 1]$ and $\beta > 0$, suppose the target distribution Q is* $\mathrm{STN}(\alpha, \beta)$ *with constants C_α and C_β. Then, there exist constants $C > 0$ and $c > 0$ (possibly) depending on α and β such that with probability, taken over the randomness of* $(\mathbf{X}, \mathbf{Y})$*, at least* $1 - Ce^{-ckt^2}$*,*

$$|\hat{\eta}_k(X) - \eta(X)| \leq C_\alpha \mathbf{E}[\rho^\alpha(X_1^*, X)|X] + t, \tag{8}$$

almost surely for the randomness of X.

The expected distance $\mathbf{E}[\rho^\alpha(X_1^*, X)|X]$ in Eq (8) corresponds to the bias incurred by the k-NN estimator $\hat{\eta}_k$.

From Eq (4) and Theorem 4, there exists a random variable $\xi \geq 0$ depending on $(\mathbf{X}, \mathbf{Y})$ and X such that conditioned on X, $\xi \leq t$ with probability at least $1 - e^{-ckt^2}$, and

$$\mathbf{E}\left[\mathcal{E}_Q(\hat{h}_k)\right] \leq \mathbf{E}\left[2g(X)\mathbb{1}\left\{\frac{1}{2}g(X) \leq C_\alpha \mathbf{E}\left[\rho_{\mathcal{V}}^\alpha\left(\tilde{X}_1^*, X\right)\middle|X\right] + \xi\right\}\right].$$

Similarly, from Eq (4) and Theorem 6, there exists a random variable $\xi \geq 0$ depending on $(\mathbf{X}, \mathbf{Y})$ and X such that conditioned on X, $\xi \leq t$ with probability at least $1 - e^{-ckt^2}$, and

$$\mathbf{E}\left[\mathcal{E}_Q(\hat{h}_k)\right] \leq \mathbf{E}[2g(X)\mathbb{1}\{g(X) \leq C_\alpha \mathbf{E}[\rho^\alpha(X_1^*, X)|X] + \xi\}].$$

Consequently, Eq (7) is verified in both cases.

Universal analyses for proving Proposition 1 and Theorem 3. For a constant $\epsilon > 0$, define

$$A(\epsilon, X) = \int_\epsilon^\infty \mathbf{E}\left[\mathbb{1}\left\{C_\alpha \min_{X^* \in B_i} \bar{\rho}^\alpha(X^*, X) \geq t\right\}\middle|X\right] dt.$$

Then, we have

$$\mathbf{E}\left[C_\alpha \min_{X^* \in B_i} \bar{\rho}(X^*, X)\middle|X\right] = \int_0^\infty \mathbf{E}\left[\mathbb{1}\left\{C_\alpha \min_{X^* \in B_i} \bar{\rho}^\alpha(X^*, X) > t\right\}\middle|X\right] dt \leq \epsilon + A(\epsilon, X).$$

Hence,

$$\mathbf{E}\left[\mathcal{E}_Q(\hat{h}_k)\right] \leq \mathbf{E}[2g(X)\mathbb{1}\{g(X) \leq C(\epsilon + A(\epsilon, X) + \xi)\}]. \tag{9}$$

For any positive reals a and $b_1, ..., b_m$, we have

$$\mathbb{1}\left\{a \leq \sum_{i=1}^m b_i\right\} \leq \mathbb{1}\left\{a \leq m \max_{i=1,\dots,m} b_i\right\} \leq \sum_{i=1}^m \mathbb{1}\{a \leq mb_i\}. \tag{10}$$

Applying Eq (10) to Eq (9) yields

$$\mathbf{E}\left[\mathcal{E}_Q(\hat{h}_k)\right] \leq \mathbf{E}[2g(X)\mathbb{1}\{g(X) \leq 3C\epsilon\}] \\ + \mathbf{E}[2g(X)\mathbb{1}\{g(X) \leq 3C\xi\}] + \mathbf{E}[2g(X)\mathbb{1}\{g(X) \leq 3CA(\epsilon, X)\}]. \quad (11)$$

We will provide upper bounds on the three terms in Eq (11).

First term in Eq (11). From Definition 4, we have

$$\mathbf{E}[2g(X)\mathbb{1}\{g(X) \leq 3C\epsilon\}] \leq 6C\epsilon Q_X(g(X) \leq 3C\epsilon) \leq 2C_\beta(3C\epsilon)^{1+\beta}. \quad (12)$$

Second term in Eq (11). We utilize Lemma 4 of Kpotufe et al. [12].

Lemma 1 (Kpotufe et al. [12])**.** *Let Z be a random variable depending on $(\mathbf{X}, \mathbf{Y})$ and X such that for $t > 0$,*

$$\mathbf{E}[\mathbb{1}\{Z \geq t\}] \leq Ce^{-ckt^2},$$

for some constants $C > 0$ and $c > 0$. Then, we have

$$\mathbf{E}[g(X)\mathbb{1}\{g(X) \leq Z\}] \leq 3CC_\beta\left(\frac{1+\beta}{ck}\right)^{\frac{1+\beta}{2}}.$$

Applying Lemma 1 yields

$$\mathbf{E}[g(X)\mathbb{1}\{g(X) \leq 3C\xi\}] \leq Ck^{-\frac{1+\beta}{2}}, \quad (13)$$

for some constant $C > 0$.

Third term in Eq (11). Let $\bar{D}_\mathcal{X}$ be the diameter of $\mathcal{X}$ with respect to $\bar{\rho}$. Applying Theorem 5 to the third term in Eq (11) yields

$$\mathbf{E}[g(X)\mathbb{1}\{g(X) \leq 3CA(\epsilon, X)\}] \leq 3C\int_\epsilon^{\bar{D}_\mathcal{X}} \min\left\{\frac{\Delta_\mathcal{V}(P, Q; (t/C_\alpha)^{1/\alpha})}{\lfloor n_P/k \rfloor}, \frac{\Delta_\mathcal{V}(Q, Q; (t/C_\alpha)^{1/\alpha})}{\lfloor n_Q/k \rfloor}\right\}dt. \quad (14)$$

Under the assumptions of $\Delta_\mathcal{V}$-transfer-exponent and $\Delta_\mathcal{V}$-self-exponent, Eq (14) is bounded above by

$$C\int_\epsilon^{\bar{D}_\mathcal{X}} \min\left\{\frac{t^{-\frac{\tau}{\alpha}}}{\lfloor n_P/k \rfloor}, \frac{t^{-\frac{\psi}{\alpha}}}{\lfloor n_Q/k \rfloor}\right\}dt.$$

We analyze the integration of $t^{-\gamma}$ for $\gamma > 0$, as exchanging the minimum and integral will give an upper bound. Some elementary calculations give

$$\int_\epsilon^{\bar{D}_\mathcal{X}} t^{-\gamma}dt = \begin{cases} \frac{1}{1-\gamma}\left(\bar{D}_\mathcal{X}^{1-\gamma} - \epsilon^{1-\gamma}\right) & \text{if } \gamma \neq 1, \\ \ln\left(\frac{\bar{D}_\mathcal{X}}{\epsilon}\right) & \text{if } \gamma = 1. \end{cases}$$

Hence,

$$\int_\epsilon^{\bar{D}_\mathcal{X}} t^{-\gamma}dt \leq \begin{cases} C\left(\frac{\bar{D}_\mathcal{X}}{\epsilon}\right)^{\gamma-1} & \text{if } \gamma > 1, \\ \log\left(\frac{\bar{D}_\mathcal{X}}{\epsilon}\right) & \text{if } \gamma = 1, \\ C & \text{if } \gamma < 1. \end{cases}$$

for some constant $C > 0$. Consequentially, letting

$$u_\gamma(x) = \begin{cases} x^{-(\gamma-1)} & \text{if } \gamma > 1, \\ \log(1/x) & \text{if } \gamma = 1, , \\ 1 & \text{if } \gamma < 1, \end{cases}$$

for $x > 0$, Eq (14) is bounded above by

$$C\min\left\{\frac{u_{\frac{\tau}{\alpha}}\left(\frac{\epsilon}{\bar{D}_\mathcal{X}}\right)}{\lfloor n_P/k \rfloor}, \frac{u_{\frac{\psi}{\alpha}}\left(\frac{\epsilon}{\bar{D}_\mathcal{X}}\right)}{\lfloor n_Q/k \rfloor}\right\}. \quad (15)$$

Rest of the proof. By combining Eqs (12), (13) and (15), we have

$$\mathbf{E}\left[\mathcal{E}_Q(\hat{h}_k)\right] \leq C\left(\epsilon^{1+\beta} + k^{-\frac{1+\beta}{2}} + \min\left\{\frac{u_{\frac{\tau}{\alpha}}\left(\frac{\epsilon}{\bar{D}_{\mathcal{X}}}\right)}{\lfloor n_P/k \rfloor}, \frac{u_{\frac{\psi}{\alpha}}\left(\frac{\epsilon}{\bar{D}_{\mathcal{X}}}\right)}{\lfloor n_Q/k \rfloor}\right\}\right).$$

We can obtain the desired rate in Theorem 2 by setting

$$\frac{\epsilon}{\bar{D}_{\mathcal{X}}} = c \min\left\{\left\lfloor\frac{n_P}{k}\right\rfloor^{-\frac{1}{\beta+\max\left\{1,\frac{\tau}{\alpha}\right\}}}, \left\lfloor\frac{n_Q}{k}\right\rfloor^{-\frac{1}{\beta+\max\left\{1,\frac{\psi}{\alpha}\right\}}}\right\},$$

for some constant $c > 0$ so that $\epsilon \leq \bar{D}_{\mathcal{X}}$ and assigning k as shown in the statement. Note that with k shown in the statement, we have

$$\max\left\{n_P^{\frac{2}{2+\beta+\max\{1,\frac{\tau}{\alpha}\}}}, n_Q^{\frac{2}{2+\beta+\max\{1,\frac{\psi}{\alpha}\}}}\right\} \leq k \leq 2\max\left\{n_P^{\frac{2}{2+\beta+\max\{1,\frac{\tau}{\alpha}\}}}, n_Q^{\frac{2}{2+\beta+\max\{1,\frac{\psi}{\alpha}\}}}\right\}. \tag{16}$$

Assume $\tau \neq \alpha$ and $\psi \neq \alpha$. Then, assigning ϵ yields

$$\begin{aligned}\mathbf{E}\left[\mathcal{E}_Q(\hat{h}_k)\right] \leq C\Bigg(&\min\left\{\left\lfloor\frac{n_P}{k}\right\rfloor^{-\frac{1+\beta}{\beta+\max\left\{1,\frac{\tau}{\alpha}\right\}}}, \left\lfloor\frac{n_Q}{k}\right\rfloor^{-\frac{1+\beta}{\beta+\max\left\{1,\frac{\psi}{\alpha}\right\}}}\right\} + k^{-\frac{1+\beta}{2}}\\ &+ \min\left\{\left\lfloor\frac{n_P}{k}\right\rfloor^{-\frac{1+\beta}{\beta+\max\left\{1,\frac{\tau}{\alpha}\right\}}}, \left\lfloor\frac{n_Q}{k}\right\rfloor^{-\frac{1+\beta}{\beta+\max\left\{1,\frac{\psi}{\alpha}\right\}}}\right\}\Bigg),\end{aligned} \tag{17}$$

where we use $\min\{\min\{a^{\alpha_1}, b^{\beta_1}\}/a^{\alpha_2}, \min\{a^{\alpha_1}, b^{\beta_1}\}/b^{\beta_2}\} \leq \min\{a^{\alpha_1-\alpha_2}, b^{\beta_1-\beta_2}\}$ for $a, b, \alpha_1, \alpha_2, \beta_1, \beta_2 > 0$ to obtain the third term. From Eq (16) and the assumption of $n_P \geq 2k$ and $n_Q \geq 2k$, we have

$$\begin{aligned}\left\lfloor\frac{n_P}{k}\right\rfloor \geq& c n_P^{\frac{\beta+\max\{1,\frac{\tau}{\alpha}\}}{2+\beta+\max\{1,\frac{\tau}{\alpha}\}}}\\ \left\lfloor\frac{n_Q}{k}\right\rfloor \geq& c n_Q^{\frac{\beta+\max\{1,\frac{\psi}{\alpha}\}}{2+\beta+\max\{1,\frac{\psi}{\alpha}\}}},\end{aligned}$$

for some constant $c > 0$. Substituting this and Eq (16) into Eq (17) yields the claim.

C.2 Proof of Proposition 2

Proof of Proposition 2. By definitions, for all P, Q, and $r > 0$,

$$\Delta_{\mathcal{V}}(P, Q; r) \leq \Delta_{\mathrm{PMW}}(P, Q; r),$$

which verifies the statement about the relationship between $\Delta_{\mathcal{V}}$ and Δ_{PMW}. Also, by definitions, for all Q and $r > 0$,

$$\Delta_{\mathrm{PMW}}(Q, Q; r) \leq \Delta_{\mathrm{DM}}(Q, Q; r),$$

by which we can verify the relationship between $\psi_{\Delta_{\mathrm{PMW}}}$ and $\psi_{\Delta_{\mathrm{DM}}}$.

Consider a cover of $\mathcal{X}_Q$ with $\mathcal{N}(\mathcal{X}_Q, \rho, r)$ balls of radius $\frac{r}{2}$ whose centers are $x_1, \ldots, x_{\mathcal{N}(\mathcal{X}_Q, \rho, \frac{r}{2})}$. Then, we have

$$\begin{aligned}\Delta_{\mathrm{PMW}}(Q, Q; r) \leq& \sum_i \int_{B(x_i, \frac{r}{2})} \frac{1}{Q_X(B(x, r))} Q_X(dx)\\ \leq& \sum_i \int_{B(x_i, \frac{r}{2})} \frac{1}{Q_X(B(x_i, \frac{r}{2}))} Q_X(dx)\\ \leq& \mathcal{N}\left(\mathcal{X}_Q, \rho, \frac{r}{2}\right),\end{aligned}$$

which yields $\psi_{\Delta_{\mathrm{PMW}}} \leq \psi_{\Delta_{\mathrm{BCN}}}$.

Lastly, we prove the inequality $\tau_{\Delta_{\mathrm{PMW}}} \le \tau_{\Delta_{\mathrm{KM}}} + \min\{\psi_{\Delta_{\mathrm{DM}}}, \psi_{\Delta_{\mathrm{BCN}}}\}$. Suppose $\Delta_{\mathrm{KM}}(P, Q; r) \le Cr^{-\tau}$. Then, we have

$$\begin{aligned}\Delta_{\mathrm{PMW}}(Q, Q; r) =& \int \frac{1}{P_X(B(x,r))} Q_X(dx) \\ \le& Cr^{-\tau} \int \frac{1}{Q_X(B(x,r))} Q_X(dx) = Cr^{-\tau} \Delta_{\mathrm{PMW}}(Q, Q; r),\end{aligned}$$

which gives the desired inequality. □

C.3 Proof of Theorem 4

Proof of Theorem 4. From the α-Hölder continuity assumption, we have, conditioned on X,

$$\begin{aligned}|\hat{\eta}(X) - \eta(X)| =& \left| \frac{1}{k} \sum_{i=1}^{k} Y_{(i)} - \eta(X) \right| \\ \le& \left| \frac{1}{k} \sum_{i=1}^{k} \left(Y_{(i)} - \eta(X_{(i)})\right) \right| + \left| \frac{1}{k} \sum_{i=1}^{k} \eta(X_{(i)}) - \eta(X) \right| \\ \le& \left| \frac{1}{k} \sum_{i=1}^{k} \left(Y_{(i)} - \eta(X_{(i)})\right) \right| + \frac{C_\alpha}{k} \sum_{i=1}^{k} \rho^\alpha\left(X_{(i)}, X\right). \end{aligned} \tag{18}$$

Applying the Hoeffding inequality into the first term in Eq (18) with conditioned on $X_1, ..., X_{n_P + n_Q}$ yields that the first term in Eq (18) is bounded above by $\frac{t}{2}$ with probability at least $1 - 2e^{-kt^2/2}$. Let us focus on the second term in Eq (18). For any distinct indices $j_1, ..., j_k \in \{1, ..., n_P + n_Q\}$, we have

$$\frac{1}{k} \sum_{i=1}^{k} \rho^\alpha\left(X_{(i)}, X\right) \le \frac{1}{k} \sum_{i=1}^{k} \rho^\alpha(X_{j_i}, X),$$

as $X_{(1)}, ..., X_{(k)}$ are the k-NNs of X. We set these indices as the indices of k-NNs in terms of the vicinity distance $\rho_{\mathcal{V}}$. Letting $\tilde{X}_{(1)}, ..., \tilde{X}_{(k)}$ be such k-NNs, we have

$$\frac{1}{k} \sum_{i=1}^{k} \rho^\alpha\left(X_{(i)}, X\right) \le \frac{1}{k} \sum_{i=1}^{k} \rho^\alpha\left(\tilde{X}_{(i)}, X\right). \tag{19}$$

In the same manner, Eq (19) is bounded above by the average distance of the implicit vicinity 1-NNs, i.e.,

$$\frac{1}{k} \sum_{i=1}^{k} \rho^\alpha\left(\tilde{X}_{(i)}, X\right) \le \frac{1}{k} \sum_{i=1}^{k} \rho^\alpha\left(\tilde{X}_i^*, X\right).$$

From the triangle inequality and the definition of the vicinity set in Eq (3), for any points $X_i' \in \mathcal{V}(X)$, we have

$$\begin{aligned}\frac{1}{k} \sum_{i=1}^{k} \rho^\alpha\left(\tilde{X}_i^*, X\right) \le& \frac{1}{k} \sum_{i=1}^{k} \left(\rho^\alpha\left(\tilde{X}_i^*, X_i'\right) + \rho(X_i', X)\right) \\ \le& \frac{1}{k} \sum_{i=1}^{k} \left(\rho^\alpha\left(\tilde{X}_i^*, X_i'\right) + \frac{1}{2} g(X)\right).\end{aligned}$$

By the definition of the infimum, for any $\epsilon > 0$, there exist X_i' such that

$$\frac{1}{k} \sum_{i=1}^{k} \left(\rho^\alpha\left(\tilde{X}_i^*, X_i'\right) + \frac{1}{2} g(X)\right) \le \frac{1}{k} \sum_{i=1}^{k} \rho_{\mathcal{V}}^\alpha\left(\tilde{X}_i^*, X\right) + \epsilon + \frac{1}{2} g(X).$$

From the arbitrariness of $\epsilon > 0$, we have

$$\frac{1}{k}\sum_{i=1}^{k}\rho^{\alpha}\left(\tilde{X}_i^*, X\right) \leq \frac{1}{k}\sum_{i=1}^{k}\rho_{\mathcal{V}}^{\alpha}\left(\tilde{X}_i^*, X\right) + \frac{1}{2}g(X). \tag{20}$$

We now apply the Hoeffding inequality into the first term of Eq (20). Then, the first term of Eq (20) is bounded above as

$$\frac{C_\alpha}{k}\sum_{i=1}^{k}\rho_{\mathcal{V}}^{\alpha}\left(\tilde{X}_i^*, X\right) \leq C_\alpha \mathbf{E}[\rho^\alpha(X_1^*, X)|X] + \frac{t}{2}, \tag{21}$$

with probability at least $1 - 2e^{-kt^2/2C_\alpha D_{\mathcal{X}}\alpha}$. The claim is verified by combining Eqs (18), (20) and (21) and the union bound. □

C.4 Proof of Theorem 5

Proof of Theorem 5. Remark that B_1 contains the subsamples from $(\mathbf{X}, \mathbf{Y})_P$ with the size $\lfloor n_P/k \rfloor$ and $(\mathbf{X}, \mathbf{Y})_Q$ with the size $\lfloor n_Q/k \rfloor$. By the mutual independence among $X_1, ..., X_{n_P+n_Q}$, we have

$$\begin{aligned}
&\mathbf{E}\left[\mathbb{1}\left\{\min_{X^*\in B_1}\rho(X^*, X) > t\right\}|X\right] \\
=&\mathbf{E}[\mathbb{1}\{\forall X^* \in B_1, \rho(X^*, X) > t\}|X] \\
=&\prod_{X^*\in B_1}\mathbf{E}[\mathbb{1}\{\rho(X^*, X) > t\}|X] \\
=&(1 - P_X(B(X,t)))^{\lfloor \frac{n_P}{k} \rfloor}(1 - Q_X(B(X,t)))^{\lfloor \frac{n_Q}{k} \rfloor} \\
\leq&\left(\left\lfloor\frac{n_P}{k}\right\rfloor P_X(B(X,t)) + \left\lfloor\frac{n_Q}{k}\right\rfloor Q_X(B(X,t))\right)^{-1},
\end{aligned}$$

where the last inequality follows from $(1-p)^n(1-q)^m \leq \exp(-(np+mq)) \leq (np+mq)^{-1}$. Taking the expectation over X yields

$$\begin{aligned}
&\mathbf{E}\left[\mathbb{1}\left\{\min_{X^*\in B_1}\rho(X^*, X) > t\right\}\right] \\
\leq&\mathbf{E}\left[\left(\left\lfloor\frac{n_P}{k}\right\rfloor P_X(B(X,t)) + \left\lfloor\frac{n_Q}{k}\right\rfloor Q_X(B(X,t))\right)^{-1}\right] \\
\leq&\mathbf{E}\left[\min\left\{\frac{1}{\lfloor n_P/k \rfloor P_X(B(X,t))}, \frac{1}{\lfloor n_Q/k \rfloor Q_X(B(X,t))}\right\}\right] \\
\leq&\min\left\{\mathbf{E}\left[\frac{1}{\lfloor n_P/k \rfloor P_X(B(X,t))}\right], \mathbf{E}\left[\frac{1}{\lfloor n_Q/k \rfloor Q_X(B(X,t))}\right]\right\},
\end{aligned}$$

which concludes the claim. □

D Experiment Details

Detailed setup. We investigated the relationship between the source sample size n_P and the excess error for k-NN classifiers using our parameter settings and those of Pathak et al. [16]. The training dataset was constructed by combining a sample from P with size n_P and a sample from Q with size n_Q. We varied n_P as $n_P \in \{2^8, 2^9, ..., 2^{18}\}$ while fixing $n_Q = 10$. The test dataset, denoted as $(X_1', Y_1'), ..., (X_m', Y_m')$, was sampled from Q with size $m = 5000$. The empirical excess error was calculated using the following formula:

$$\mathcal{E}_{\text{test},Q}(\hat{h}_k) = \frac{1}{m}\sum_{i=1}^{m} 2g(X_i')\mathbb{1}\left\{\hat{h}_k(X_i') \neq h(X_i')\right\}.$$

We explored different parameter settings for α and τ, with $\alpha \in \{\frac{1}{2}, \frac{1}{4}\}$ and $\tau \in \{1, 2\}$. For each parameter combination, we reported the average, first quartile, and third quartile of the excess error over 10 runs.

Computational environment. All experiments were conducted on a machine equipped with an Intel Core i7-1065G7 CPU @ 1.30GHz, 16GB RAM. The implementation was done using Python 3.8.10 and the scikit-learn library (version 0.0.post11) for the k-NN classifier.

Remark on results. For the guidelines representing the upper bounds shown in Figure 1, we adjusted the multiplicative constant of the lines so that the point at $n_P = 2^8$ matches the experimental result of the corresponding k-NN at $n_P = 2^8$. This adjustment helps to provide a clear visual comparison between the theoretical upper bounds and the empirical results.

E Detailed Related Work

Theoretical works for covariate-shift typically provide upper bounds on the generalization error, which is the difference between the empirical average and expected losses. These works establish a connection between some divergence measure of the source and target distributions and the generalization error. For instance, Ben-David et al. [3]'s analyses yield a generalization error bound that includes the $\mathcal{H}\Delta\mathcal{H}$-divergence, a measure of the discrepancy between the source and target distributions. Similarly, Park et al. [15] introduce the *source-discrimination error*, which can be interpreted as a divergence between the source and target distributions, and provide a generalization error bound that incorporates this term. Aminian et al. [1] employ the Kullback-Leibler (KL) divergence between the source and target distributions to derive a generalization error bound under covariate-shift. Their techniques are applicable to a broader range of situations, as they do not make any assumptions about the source and target distributions. However, their approach may not be capable of confirming the consistency of the source sample size because their divergence measures remain positive even as the source sample size approaches infinity.

Several researchers leverage the likelihood ratio between the source and target distributions to derive upper bounds on the excess error under the covariate-shift setup [6, 11, 13]. Their techniques can confirm the source sample-size consistency of their algorithms, but under the assumption that the learner has access to the likelihood ratio function. However, in real-world scenarios, the likelihood ratio function needs to be estimated using the training sample, which may introduce an estimation error. It is not certain that their methods exhibit the source sample-size consistency when employing the empirical estimation of the likelihood ratio.

Several techniques can confirm the existence of the source sample-size consistent algorithm [7, 12, 16]. We explored the comparison between our results and those obtained by Kpotufe et al. [12] and Pathak et al. [16] in Section 3 and demonstrated that our analysis always gives an upper bound with faster or competitive rates in Proposition 2. Galbraith et al. [7] introduced an ”average” discrepancy to more tightly capture the behavior of the excess error under covariate-shift in classification. However, their technique does not account for the vicinity information and has the same limitations as the techniques by Kpotufe et al. [12] and Pathak et al. [16], such as the inability to confirm the existence of the source sample-size consistent algorithm under support non-containment situations.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main contributions of the paper are clearly stated and discussed in both the abstract and introduction.

   Guidelines:
   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5

   Guidelines:
   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All the assumptions are presented in the statement of each theorem, and the proofs are provided in Appendix C.

   Guidelines:
   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.

- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides detailed information about the experimental setup, datasets, and methodology in Appendix D.

   Guidelines:
   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We only use the synthetic data. The code will also be provided as supplementary materials.

   Guidelines:
   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Appendix D

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Figure 1

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix D

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Section 5

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification:

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.