Nonparametric Estimation of Non-Smooth Divergences

Mina Mahbub Hossain
mahbub.hossain@usu.edu
Utah State University
Logan, Utah, USA

Alan Wisler
alan.wisler@usu.edu
Utah State University
Logan, Utah, USA

Kevin R. Moon
kevin.moon@usu.edu
Utah State University
Logan, Utah, USA

ABSTRACT

Nonparametric estimation of information divergence functionals between two probability densities is an important problem in machine learning. Several estimators exist that guarantee the parametric rate of mean squared error (MSE) of $O(1/N)$ under various assumptions on the smoothness and boundary of the underlying densities, with N being the number of samples. In particular, previous work on ensemble estimation theory derived ensemble estimators of divergence functionals that achieve the parametric rate without requiring knowledge of the densities' support set and are simple to implement. However, these and most other methods all assume some level of differentiability of the divergence functional. This excludes important divergence functionals such as the total variation distance and the Bayes error rate. Here, we show empirically that the ensemble estimation approach for smooth functionals can be applied to less smooth functionals and obtain good convergence rates, suggesting a gap in current theory.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; **Supervised learning by classification**.

KEYWORDS

Machine Learning, Divergence Estimation, Bayes Error Rate, Nonparametric Estimation

ACM Reference Format:
Mina Mahbub Hossain, Alan Wisler, and Kevin R. Moon. 2024. Nonparametric Estimation of Non-Smooth Divergences. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), October 21–25, 2024, Boise, ID, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3627673.3679972

1 INTRODUCTION

Information divergence functionals are integral functionals of two probability distributions. Accurate divergence estimation is of great importance to the fields of machine learning, information theory, and statistics. Some applications of divergences include estimating bounds on the Bayes error for a classification problem [2, 15, 18, 36], extending machine learning algorithms to distributional features [16, 25, 34], feature selection and classification [2, 5, 27], and image segmentation [8, 13]. See [1] for more applications of divergence measures.

CIKM '24, October 21–25, 2024, Boise, ID, USA

ACM ISBN 979-8-4007-0436-9/24/10
https://doi.org/10.1145/3627673.3679972

The family of f-divergences is an important subset of information divergences [7]. This family includes the well-known Kullback-Leibler (KL) divergence [12], the Rényi-α divergence [26], the Hellinger-Bhattacharyya distance [4, 9], the Chernoff-α divergence [6], the Henze-Penrose divergence [2], and the total variation distance. The Bayes error rate (BER) is another important divergence functional that represents the best generalization error that can be achieved on a classification problem with a given feature space. Thus if known, the BER is incredibly useful for model benchmarking as it can help diagnose sub-optimal model performance or cases where inaccurate evaluations may overestimate generalized accuracy.

In many problems, parametric divergence estimators are inaccurate due to a mismatch between the data and the parametric model. Thus many nonparametric estimators of different divergence functionals have been proposed with varying levels of theoretical guarantees [2, 3, 10, 11, 15, 17, 19, 20, 23–25, 28–32, 35, 37]. Some of these estimators are guaranteed to achieve the parametric mean squared error (MSE) of $O(1/N)$ under certain smoothness assumptions on the densities and the divergence functional [3, 10, 11, 15, 17, 19, 20, 23, 24, 29–31, 37]. However, the vast majority of these guarantees require the divergence functional to be differentiable, which excludes some of the most important functionals for machine learning, such as the total variation distance and the BER.

Of particular interest to us are ensemble estimators, which take a weighted average of an ensemble of simple base estimators [3, 20]. The theory of optimally weighted ensemble estimation is a general theory originally presented by Sricharan et al [33] and later extended in [20]. The theory is especially well-suited for problems where the bias of a base estimator is high while the variance is low. In this case, the theory allows us to construct a weighted ensemble estimator where the weights are chosen to greatly reduce the bias in exchange for smaller increases in the variance, thus reducing the overall MSE. The theory has been applied successfully to derive nonparametric estimators of entropy, divergence, and mutual information that achieve the parametric convergence rate.

The base estimators for these ensemble methods typically consist of either kernel density estimator (KDE) or k-nn plug-in estimators. To select the optimal weights, a bound on the variance and an expression for the bias of the plug-in estimators in terms of the ensemble parameter (e.g. the bandwidth for KDE and k for k-nn) must be derived. These ensemble estimators are computationally fast when the base estimators are computed efficiently, simple to implement, achieve the parametric MSE convergence rate when the densities are sufficiently smooth, and do not require direct knowledge of the densities' support set. The latter point is relevant in that many competing non-parametric estimators require complicated calculations at the boundary of the support [10, 11, 29, 30].

While these ensemble estimators are relatively straightforward to implement, the theoretical groundwork required to derive the

bias terms (which are necessary to set up optimization problem and obtain guarantees on the convergence rate) can be cumbersome [17, 20–22]. These derivations can be especially difficult when the density support set contains boundaries. Furthermore, this theory has (thus far) only been applied to differentiable functionals. In this paper we provide empirical evidence that the ensemble estimation approach can be applied to non-differentiable functionals, such as the BER, and still obtain good MSE convergence rates.

2 ENSEMBLE ESTIMATION

Consider an indexed ensemble of estimators $\{\hat{\mathbf{E}}_\ell\}_{\ell\in\bar{\ell}}$ of a parameter E and a set of weights $\{w(\ell)\}_{\ell\in\bar{\ell}}$ with $\sum_{\ell\in\bar{\ell}} w(\ell) = 1$. The weighted ensemble estimator of E is simply $\hat{\mathbf{E}}_w = \sum_{\ell\in\bar{\ell}} w(\ell)\hat{\mathbf{E}}_\ell$. In [20], the authors required that the variance of $\hat{\mathbf{E}}_\ell$ be $O(1/T)$, where T is the sample size, and that the bias of $\hat{\mathbf{E}}_\ell$ can be written where its dependence on T and the index parameter ℓ is known precisely for all bias terms that converge slower than $O(1/\sqrt{T})$. Then an offline convex optimization problem can be derived that chooses optimal weights w_0 such that $\hat{\mathbf{E}}_{w_0}$ achieves the parametric MSE rate. The optimization problem does this by selecting weights w that minimize a term that controls the variance of $\hat{\mathbf{E}}_w$ while forcing the slow bias terms to converge faster to zero. Specifically, consider the following conditions on $\{\hat{\mathbf{E}}\}_{\ell\in\bar{\ell}}$:

- $C.1$ The bias can be written as

$$\mathbf{Bias}(\mathbf{E}_\ell) = \sum_{i\in J} c_i\psi_i(\ell)\phi_{i,d}(T) + O\left(1/\sqrt{T}\right),$$

  where c_i are constants that are independent of T and ℓ, J is a finite index set with $I < L$ values, ψ_i are basis functions that depend only on the parameter ℓ, and $\phi_{i,d}$ depend only on T.
- $C.2$ The variance can be bounded by $O(1/T)$.

Theorem 2.1 (Adapted from [20]). *Assume conditions $C.1$ and $C.2$ hold for an ensemble of estimators $\{\hat{\mathbf{E}}\}_{\ell\in\bar{\ell}}$. Then there exists a weight vector w_0 such that the MSE of the weighted ensemble estimator is $O(1/T)$. The weight vector w_0 is obtained by solving the following convex optimization problem:*

$$\begin{aligned} \min_{w} \quad & \epsilon \\ \textit{subject to} \quad & \sum_{\ell\in\bar{\ell}} w(\ell) = 1 \\ & \mid \phi_{i,d}(T) \sum_{\ell\in\bar{\ell}} w(\ell)\psi_i(\ell) \mid \leq \epsilon T^{-1/2}, \quad \forall i \in J \\ & \|w\|_2^2 \leq \eta\epsilon. \end{aligned} \tag{1}$$

The parameter η is chosen to achieve a tradeoff between bias and variance. Note how the second constraint ensures that the slow bias terms in Eq. 3 have a rate of $O(1/\sqrt{T})$ by controlling the basis functions $\psi_i(\ell)$.

We will now consider specifically the problem of divergence functional estimation. Let f_1 and f_2 be d-dimensional probability densities with common support. The f-divergence between f_1 and f_2 has the following form [7]:

$$\mathcal{D}_\phi(f_1, f_2) = \int \phi\left(\frac{f_1(x)}{f_2(x)}\right) f_2(x)dx.$$

For $\mathcal{D}_\phi$ to be considered a true divergence, the function ϕ must be convex and $\phi(1) = 0$. Ensemble estimation can be extended to other divergence functionals, although we focus on f-divergences for simplicity.

We will assume that the densities f_1 and f_2 have a common bounded support set $\mathcal{S}$ and f_1 and f_2 are strictly lower bounded. Assume that $T = N + M$ independent and identically distributed (i.i.d.) realizations $\mathcal{X}_T = \{X_1, X_2, \ldots, X_N, X_{N+1}, \ldots, X_{N+M}\}$ are available from the density f_2 and M i.i.d. realizations $\mathcal{Y}_M = \{Y_1, Y_2, \ldots, Y_M\}$ are available from the density f_1, where M is proportional to T.

The ensemble theory was first applied to k-nearest neighbor (nn) density plug-in estimators. Let $k \leq M$ and let $\rho_{2,k}(i)$ be the distance of the k^{th} nearest neighbor of X_i in $\{X_{N+1}, X_{N+2}, \cdots ..., X_{N+M}\}$. Similarly, define $\rho_{1,k}(i)$ be the distance of the k^{th} nearest neighbor of X_i in $\{Y_1, Y_2, \cdots ..., Y_M\}$. Then the k-nn density estimator [14] at the point X_i is

$$\hat{f}_{j,k}(X_i) = \frac{k}{M\bar{c}\rho_{j,k}^d(i)},$$

where $\bar{c}$ is the volume of a d-dimensional unit ball. The functional $\mathcal{D}_\phi$ is then approximated as

$$\hat{\mathcal{D}}_{\phi,k} = \frac{1}{N}\sum_{i=1}^{N} \phi\left(\frac{\hat{f}_{1,k}(X_i)}{\hat{f}_{2,k}(X_i)}\right). \tag{2}$$

Choose an ensemble of positive numbers $\bar{\ell} = \{\ell_1, \ell_2, \cdots ..., \ell_L\}$ where $L > d - 1$ and let $k(\ell) = \ell\sqrt{M}$. It was shown that if 1. the density support set is bounded without any boundaries (e.g. the surface of a torus), 2. the functional ϕ has a sufficient number of derivatives, and 3. the densities have at least d derivatives, then the bias and variance of $\hat{\mathcal{D}}_{\phi,k(\ell)}$ are [17]

$$\mathbf{Bias}(\hat{\mathcal{D}}_{\phi,k(\ell)}) = \sum_{j=1}^{d} c_j\left(\frac{\ell}{\sqrt{M}}\right)^{\frac{j}{d}} + O\left(\frac{1}{\sqrt{M}}\right) \tag{3}$$

$$\mathbf{Var}(\hat{\mathcal{D}}_{\phi,k(\ell)}) = O\left(\frac{1}{N} + \frac{1}{M}\right), \tag{4}$$

where the constants are independent of M and ℓ. Then given a weight vector w with length L, define $\hat{\mathcal{D}}_{\phi,w} = \sum_{\ell\in\bar{\ell}} w(\ell)\hat{\mathcal{D}}_{\phi,k(\ell)}$. Then the optimization problem in Eq. 1 with $\psi_i(\ell) = \ell^{\frac{i}{d}}$ and $\phi_{i,d}(N) = N^{-\frac{i}{2d}}$ for $i \in \{1, \ldots, d\}$ returns a weight vector w_0 such that the weighted ensemble estimator $\hat{\mathcal{D}}_{\phi,w_0}$ achieves the parametric MSE rate under the same assumptions given above.

By a similar procedure, ensemble estimators were obtained for KDE plug-in estimators where the bandwidth h of the KDE is chosen to depend on the parameter ℓ, resulting in basis functions that can be similarly controlled by the weight vector w [20]. The resulting ensemble estimator can achieve the parametric rate when the densities have more than $d/2$ derivatives and can be applied to densities with boundaries on their support set as long as the boundaries are sufficiently smooth. However, deriving the bias results in both scenarios is tedious and difficult, especially when the density support set contains boundaries, and it has not been performed for the case when ϕ is not differentiable. Our work suggests empirically that assuming the bias and variance for less smooth

divergences have the same form as in Eqs. 3 and 4 results in an effective ensemble estimator.

3 BIAS CONSIDERATIONS

The bias expression in Eq. 3 is a polynomial function of $\left(\frac{\ell}{\sqrt{M}}\right)^{\frac{1}{d}}$. If ϕ is not differentiable everywhere, the bias expression for $\hat{\mathcal{D}}_{\phi,k}$ could include polynomial terms of the form $\left(\frac{\ell}{\sqrt{M}}\right)^{\frac{\lambda}{d}}$ with $\lambda > 0$. In that case, the constraints in the optimization problem in Eq. 1 should include terms $\psi_i(\ell) = \ell^{\frac{\lambda i}{d}}$ and $\phi_{i,d}(N)$ should have terms of the form $N^{-\frac{\lambda i}{2d}}$. Thus one way to apply the ensemble estimation approach to estimate less smooth functionals is to include these extra terms in the optimization. On the other hand, including extra terms in the constraints that aren't present in the bias could potentially hurt estimation performance. Thus in our experiments, we assess the potential benefit or harm of including additional constraint terms by estimating both the BER (unknown bias terms) and the Renyi-α divergence integral (known bias terms) with different values of λ.

We do not need any constraints with $\lambda < 1$ if the densities are bounded above and below and the function ϕ is Lipschitz continuous (this includes most functionals of interest). In that case, it can be shown that the bias of $\hat{\mathcal{D}}_{\phi,k}$ is bounded above by the bias of the density estimators. Finally, the slowest term of the density estimators is $O\left(\frac{\ell}{\sqrt{M}}\right)^{\frac{1}{d}}$ when the densities' support set contains boundaries [22]. So in our experiments, we will only consider $\lambda \geq 1$.

We analyze the ensemble estimator bias to determine the potential effects of including extra terms in the constraints. Assume that w_0 and ϵ_0 are the solutions to Eq. 1. Define

$$r(\lambda) = \left| \sum_{\ell \in \bar{\ell}} w_0(\ell) \ell^{\lambda/d} N^{-\lambda/2d} N^{1/2} \right|, \tag{5}$$

$$r_{\max} = \max_{\lambda \in [1,d]} \left| \sum_{\ell \in \bar{\ell}} w_0(\ell) \ell^{\lambda/d} N^{-\lambda/2d} N^{1/2} \right|. \tag{6}$$

Suppose that there exists a term of the form $\ell^{\frac{\lambda}{d}}$ for some λ in the bias that is not included in the optimization problem in Eq. 1. Then the bias of $\hat{\mathcal{D}}_{\phi,\mathbf{w}_0}$ will have a term of the form:

$$\begin{aligned}
\left| \sum_i c_i \sum_{\ell \in L\bar{\ell}} w(\ell) \ell^{\lambda/d} N^{-\lambda/2d} \right| &\leq \left| \sum_i c_i \right| \left| \sum_{\ell \in \bar{\ell}} w(\ell) \ell^{\lambda/d} N^{-\lambda/2d} \right| \\
&\leq \sum_i |c_i| \left| \sum_{\ell \in \bar{\ell}} w(\ell) \ell^{\lambda/d} N^{-\lambda/2d} \right| \\
&\leq \sum_i |c_i| \left(r_{\max} N^{-(1/2)} \right) \\
&\leq \|\mathbf{c}\|_1 r_{\max} N^{-(1/2)}.
\end{aligned} \tag{7}$$

Thus in the worst case scenario (a term is neglected in the bias), the bias of the ensemble estimator is asymptotically bounded above by $r_{\max} N^{-(1/2)}$. In our experiments, we will compare the MSE of the estimators to the asymptotic squared bias terms $b_0(N) = \epsilon_0^2 N^{-1}$ and $b_{\max}(N) = r_{\max}^2 N^{-1}$.

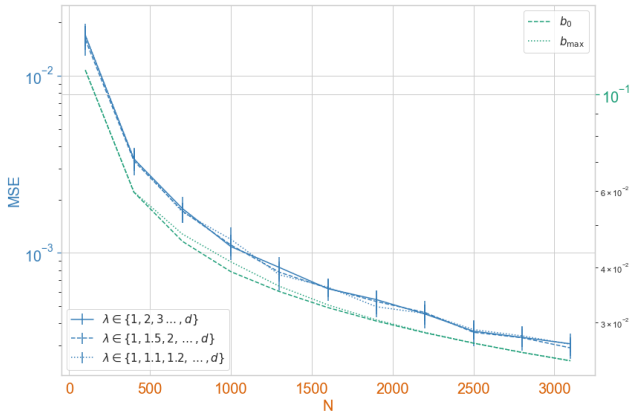

Figure 1: MSE as a function of sample size N when estimating the Rényi-α divergence integral when including different sets of constraints in the optimization problem in Eq. 1. Error bars on the MSE reflect the standard deviation from 20 trials. The asymptotic squared bias bounds b_0 and $b_{\max}$ are also included.

4 EXPERIMENTS

To compare the performance of the ensemble estimator $\hat{\mathcal{D}}_{\phi,\mathbf{w}_0}$ on smooth versus non smooth divergence functionals, we estimated the Rényi-α divergence integral and the BER between two truncated normal densities restricted to the unit cube with varying dimension and sample size. Note that the true divergences will change as d changes and can be computed analytically. Assuming the prior class probabilities are equal, the respective functionals $\phi(t)$ are then t^α and $\frac{1}{2}\min(t, 1)$. The densities have means $\bar{\mu}_1 = 0.7 * \bar{1}_d$, $\bar{\mu}_2 = 0.3 * \bar{1}_d$ and covariance matrices $\sigma_i * \mathcal{I}_d$ where $\sigma_1 = 0.1$, $\sigma_2 = 0.3$, $\bar{1}_d$ is a d-dimensional vector of ones, and $\mathcal{I}_d$ is a d-dimensional identity matrix. We used $\alpha = 0.5$ and computed w_0 by solving the convex optimization problem in Eq. 1.

For each value of k, we used a leave-one-out estimator instead of the data-splitting approach given in Eq. 2. Thus N is effectively equal to T. The default simulation parameters were chosen as follows: sample size $N = 1000$, dimension of data $d = 7$, trade-off parameter between bias and variance $\eta = 0.3$, the minimum and maximum values of $\bar{\ell}$ are respectively 0.3 and 3.0, and the number of values in this range is $L = |\bar{\ell}| = 50$. When we varied N, we chose $N \in \{100, 400, 700, \ldots, 3100\}$. When varying the dimension, we chose $d \in \{2, 5, 8, \ldots, 25, 30\}$. Experiments were repeated 100 times for each setting to estimate the MSE and this was repeated 20 times to obtain error bars on the MSE.

Based on Eqs. 1 and 3, the standard terms to include in the optimization problem for smooth functionals include integer values of λ between 1 and d. In our experiments, we considered the effects of including additional values of λ in the optimization problem, specifically all terms with $\lambda \in \{1, 1.5, 2, \ldots, d\}$ and with $\lambda \in \{1, 1.1, 1.2, \ldots, d\}$.

Figure 1 shows the MSE of the ensemble estimator under each of these scenarios as well as the asymptotic bounds b_0 and $b_{\max}$ when estimating the Rényi-α divergence integral. From these results, it is

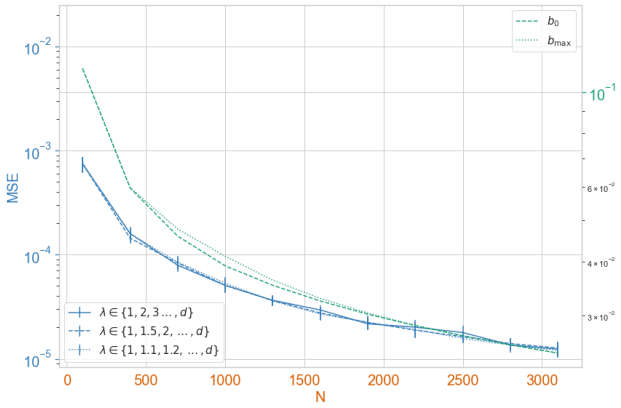

Figure 2: MSE as a function of sample size N when estimating the BER when including different sets of constraints in the optimization problem in Eq. 1. The asymptotic squared bias bounds b_0 and $b_{\max}$ are also included.

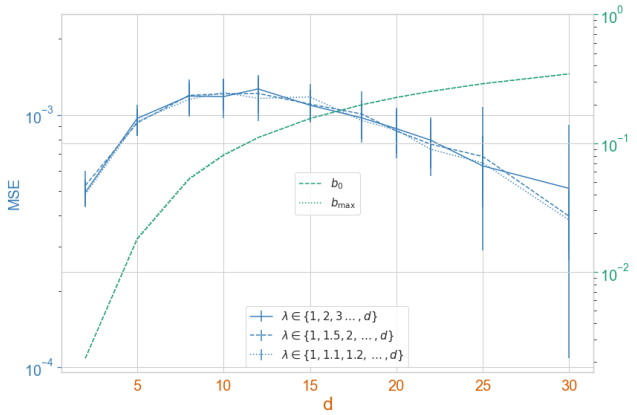

Figure 3: MSE as a function of dimension d when estimating the Rényi-α divergence integral when including different sets of constraints in the optimization problem in Eq. 1. The asymptotic squared bias bounds b_0 and $b_{\max}$ are also included.

clear that the ensemble estimators' MSE decreases as N increases and the choice of terms in the optimization problem has a little effect. This is corroborated by the behavior of the bounds b_0 and $b_{\max}$, which show a similar trend. Note that since the constants c_i are unknown, b_0 and $b_{\max}$ may not be true upper bounds on the squared bias and thus have a different scale from the MSE, although they do reflect the asymptotic behavior. These results suggest that including extra terms in the constraints of the optimization problem in Eq. 1 beyond those corresponding to $\lambda \in \{1, 2, \ldots, d\}$ is unlikely to hurt the estimation performance when those terms are not actually present in the bias. We hypothesize that this is because controlling the behavior of the $\psi_\lambda(\ell) = \ell^{\frac{\lambda}{d}}$ for integer values of λ is sufficient to control any terms with noninteger values of λ.

Figure 2 shows the same results when estimating the BER. We observe similar trends as before where all configurations of the ensemble estimators perform well, even though the exact bias for the BER estimator is unknown. These results suggest that the ensemble estimation approach can be applied to this less smooth divergence functional without modification.

Figures 3 and 4 show similar plots when varying d instead. Here the asymptotic bounds b_0 and $b_{\max}$ generally increase as the dimension increases. However, for the Rényi-α ensemble estimators, the MSE initially increases but then steadily decreases as d increases while the MSE for the estimators for the BER steadily decreases. The reason for this is that as the dimension increases, the value of both the Rényi-α divergence integral and the BER gets closer to zero. Since neither value can be negative, the estimation problem becomes somewhat simpler. However, the estimators are not simply defaulting to zero as the error bars are nonzero. As before, the differences in performance between the different optimization configurations are small, suggesting that including extra constraints in Eq. 1 is unnecessary for both smooth and nonsmooth functionals.

5 CONCLUSION

We applied the theory of optimally weighted ensemble estimation to an ensemble of k-nn plug-in estimators to estimate a nonsmooth

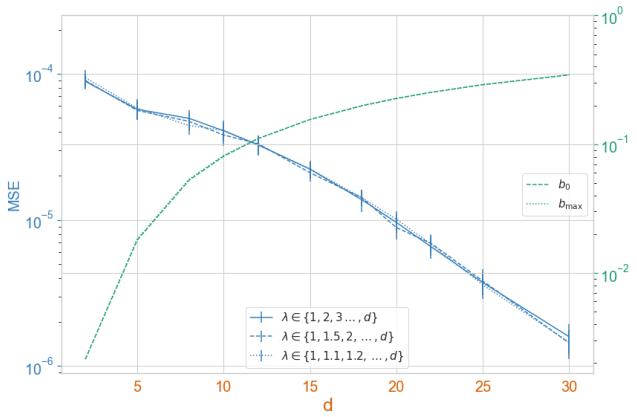

Figure 4: MSE as a function of dimension d when estimating the BER when including different sets of constraints in the optimization problem in Eq. 1. The asymptotic squared bias bounds b_0 and $b_{\max}$ are also included.

divergence functional (the BER) when assuming the same bias expression that was obtained in the literature for smooth divergence functionals. The ensemble estimator performed well in terms of MSE convergence. There are several possibilities that could explain this performance. One is that the currently unknown bias expression for k-nn plug-in estimators of the BER matches that of estimators for smooth functionals. Another is that the true bias expression for the BER estimator contains additional terms, but they are controlled by controlling the terms from the smooth case. Future work involves deriving the theoretical bounds for the nonsmooth case so that guarantees can be obtained for settings and distributions beyond those considered for our experiments.

ACKNOWLEDGMENTS

This work was supported in part by the NSF under Grant 2212325.

REFERENCES

[1] Michéle Basseville. 2013. Divergence measures for statistical data processing—An annotated bibliography. *Signal Processing* 93, 4 (2013), 621–633.

[2] Visar Berisha, Alan Wisler, Alfred O Hero, and Andreas Spanias. 2015. Empirically estimable classification bounds based on a nonparametric divergence measure. *IEEE Transactions on Signal Processing* 64, 3 (2015), 580–591.

[3] Thomas B Berrett and Richard J Samworth. 2023. Efficient functional estimation and the super-oracle phenomenon. *The Annals of Statistics* 51, 2 (2023), 668–690.

[4] Anil Bhattacharyya. 1946. On a measure of divergence between two multinomial populations. *Sankhyā: the indian journal of statistics* (1946), 401–406.

[5] Lorenzo Bruzzone, Fabio Roli, and Sebastiano B Serpico. 1995. An extension of the Jeffreys-Matusita distance to multiclass cases for feature selection. *IEEE Transactions on Geoscience and Remote Sensing* 33, 6 (1995), 1318–1321.

[6] Herman Chernoff. 1952. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics* (1952), 493–507.

[7] Imre Csiszár. 1967. On information-type measure of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.* 2 (1967), 299–318.

[8] A Ben Hamza and Hamid Krim. 2003. Image registration and segmentation by maximizing the Jensen-Rényi divergence. In *Energy Minimization Methods in Computer Vision and Pattern Recognition: 4th International Workshop, EMMCVPR 2003, Lisbon, Portugal, July 7-9, 2003. Proceedings 4*. Springer, 147–163.

[9] Ernst Hellinger. 1909. Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *Journal für die reine und angewandte Mathematik* 1909, 136 (1909), 210–271.

[10] Kirthevasan Kandasamy, Akshay Krishnamurthy, Barnabas Poczos, Larry Wasserman, and James Robins. 2015. Nonparametric von mises estimators for entropies, divergences and mutual informations. *Advances in Neural Information Processing Systems* 28 (2015).

[11] Akshay Krishnamurthy, Kirthevasan Kandasamy, Barnabas Poczos, and Larry Wasserman. 2014. Nonparametric estimation of Renyi divergence and friends. In *International Conference on Machine Learning*. PMLR, 919–927.

[12] Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics* 22, 1 (1951), 79–86.

[13] Gang Liu, Gui-Song Xia, Wen Yang, and Nan Xue. 2014. SAR image segmentation via non-local active contours. In *2014 IEEE Geoscience and Remote Sensing Symposium*. IEEE, 3730–3733.

[14] Don O Loftsgaarden and Charles P Quesenberry. 1965. A nonparametric estimate of a multivariate density function. *The Annals of Mathematical Statistics* 36, 3 (1965), 1049–1051.

[15] Kevin Moon and Alfred Hero. 2014. Multivariate f-divergence estimation with confidence. *Advances in neural information processing systems* 27 (2014).

[16] Kevin R Moon, Véronique Delouille, Jimmy J Li, Ruben De Visscher, Fraser Watson, and Alfred O Hero. 2016. Image patch analysis of sunspots and active regions-II. Clustering via matrix factorization. *Journal of Space Weather and Space Climate* 6 (2016), A3.

[17] Kevin R Moon and Alfred O Hero. 2014. Ensemble estimation of multivariate f-divergence. In *2014 IEEE International Symposium on Information Theory*. IEEE, 356–360.

[18] Kevin R Moon, Alfred O Hero, and B Véronique Delouille. 2015. Meta learning of bounds on the Bayes classifier error. In *2015 IEEE SignalPprocessing and Signal Processing Education Workshop*. IEEE, 13–18.

[19] Kevin R Moon, Kumar Sricharan, Kristjan Greenewald, and Alfred O Hero. 2016. Improving convergence of divergence functional ensemble estimators. In *2016 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 1133–1137.

[20] Kevin R Moon, Kumar Sricharan, Kristjan Greenewald, and Alfred O Hero III. 2018. Ensemble estimation of information divergence. *Entropy* 20, 8 (2018), 560.

[21] Kevin R Moon, Kumar Sricharan, and Alfred O Hero. 2021. Ensemble estimation of generalized mutual information with applications to genomics. *IEEE Transactions on Information Theory* 67, 9 (2021), 5963–5996.

[22] Kevin R Moon, Kumar Sricharan, and Alfred O Hero III. 2017. Ensemble estimation of distributional functionals via k-nearest neighbors. *arXiv preprint arXiv:1707.03083* (2017).

[23] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. 2010. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory* 56, 11 (2010), 5847–5861.

[24] Morteza Noshad, Kevin R Moon, Salimeh Yasaei Sekeh, and Alfred O Hero. 2017. Direct estimation of information divergence using nearest neighbor ratios. In *2017 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 903–907.

[25] Barnabás Póczos and Jeff Schneider. 2011. On the Estimation of \alpha -Divergences. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings, 609–617.

[26] Alfréd Rényi. 1961. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, Vol. 4. University of California Press, 547–562.

[27] DM Sakate and DN Kashid. 2014. Variable selection via penalized minimum φ-divergence estimation in logistic regression. *Journal of Applied Statistics* 41, 6 (2014), 1233–1246.

[28] Salimeh Yasaei Sekeh, Morteza Noshad, Kevin R Moon, and Alfred O Hero. 2019. Convergence rates for empirical estimation of binary classification bounds. *Entropy* 21, 12 (2019), 1144.

[29] Shashank Singh and Barnabás Póczos. 2014. Exponential concentration of a density functional estimator. *Advances in Neural Information Processing Systems* 27 (2014).

[30] Shashank Singh and Barnabás Póczos. 2014. Generalized exponential concentration inequality for Rényi divergence estimation. In *International Conference on Machine Learning*. PMLR, 333–341.

[31] Shashank Singh and Barnabás Póczos. 2016. Finite-sample analysis of fixed-k nearest neighbor density functional estimators. *Advances in Neural Information Processing Systems* 29 (2016).

[32] Sreejith Sreekumar and Ziv Goldfeld. 2022. Neural estimation of statistical divergences. *Journal of machine learning research* 23, 126 (2022).

[33] Kumar Sricharan, Dennis Wei, and Alfred O Hero. 2013. Ensemble estimators for multivariate entropy estimation. *IEEE transactions on information theory* 59, 7 (2013), 4374–4388.

[34] Zoltán Szabó, Arthur Gretton, Barnabás Póczos, and Bharath Sriperumbudur. 2015. Two-stage sampled learning theory on distributions. In *Artificial Intelligence and Statistics*. PMLR, 948–957.

[35] Qing Wang, Sanjeev R Kulkarni, and Sergio Verdú. 2009. Divergence estimation for multidimensional densities via k-Nearest-Neighbor distances. *IEEE Transactions on Information Theory* 55, 5 (2009), 2392–2405.

[36] Alan Wisler, Visar Berisha, Dennis Wei, Karthikeyan Ramamurthy, and Andreas Spanias. 2016. Empirically-estimable multi-class classification bounds. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2594–2598.

[37] Alan Wisler, Kevin Moon, and Visar Berisha. 2018. Direct ensemble estimation of density functionals. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2866–2870.