Scalable Mutual Information Estimation using Dependence Graphs

Morteza Noshad and Alfred O. Hero III

University of Michigan, Electrical Engineering and Computer Science, Ann Arbor, Michigan, U.S.A

***Abstract*—We propose a unified method for empirical non-parametric estimation of general Mutual Information (MI) function between the random vectors in $\mathbb{R}^d$ based on N i.i.d. samples. The proposed low complexity estimator is based on a bipartite graph, referred to as dependence graph. The data points are mapped to the vertices of this graph using randomized Locality Sensitive Hashing (LSH). The vertex and edge weights are defined in terms of marginal and joint hash collisions. For a given set of hash parameters $\epsilon(1), \ldots, \epsilon(k)$, a base estimator is defined as a weighted average of the transformed edge weights. The proposed estimator, called the ensemble dependency graph estimator (EDGE), is obtained as a weighted average of the base estimators, where the weights are computed offline as the solution of a linear programming problem. EDGE achieves optimal computational complexity $O(N)$, and can achieve the optimal parametric MSE rate of $O(1/N)$ if the density is d times differentiable. To the best of our knowledge EDGE is the first non-parametric MI estimator that can achieve parametric MSE rates with linear time complexity.**

I. Introduction

The Mutual Information (MI) is an often used measure of dependency between two random variables or vectors [1]. MI quantifies the amount of shared information between the two random variables. It is a measure of similarity between the joint distribution and the product of its marginal distributions. In information theory, maximizing MI over different distributions gives the capacity of a communication channel [1]. In machine learning and data science applications, MI plays a significant role in classification [2], clustering [3], feature selection [4], learning graphical models [5], image registration [6], computational biology [7], among other areas.

Non-parametric MI estimation methods have been studied that use estimation strategies including: such as KSG [8]; KDE [9]; Parzen window density estimation [10] and adaptive partitioning [11]. The performance of these estimators has been evaluated and compared based on both empirical studies [12] and asymptotic analysis [13]. Recently several MI estimators have been proposed that can achieve parametric MSE rate of convergence. For example, in [14] a KDE plug-in estimator for Rényi divergence and mutual information achieves the MSE rate of $O(1/N)$ when the densities are at least d times differentiable. Another KDE based mutual information estimator was proposed in [13] that can achieve the MSE rate of $O(1/N)$ when the densities are $d/2$ times differentiable. Both of these estimators use mirroring technique to compensate the boundary bias. However, these estimators have high computational cost and require knowledge of the density support boundary. Furthermore, the derivatives need to vanish at these boundaries.

Most of previously proposed estimators assume that the densities are either purely continuous or discrete. However, in many applications random variables may consist of a mixture of continuous and discrete components. For example, for supervised learning the MI has been used as a feature selection criterion, for which one of the variables denotes a feature vector (input covariates) which the other variable denotes a label associated with the feature vector. This was the setting for which Moon et al proposed a KDE plug-in MI estimator that uses weighted ensemble technique achieving parametric rates when the densities are at least $d/2$ times differentiable [15]. Gao et al [16] proposed a k-NN based estimator of MI that works on mixtures on continuous-discrete random variables. Both of these estimators have computational complexity that is super-linear in N.

In this paper we propose a reduced complexity MI estimator called the ensemble dependency graph estimator (EDGE). The estimator combines randomized locality sensitive hashing (LSH), dependency graphs, and ensemble bias-reduction methods. A dependence graph is a bipartite directed graph consisting of two sets of nodes V and U. The data points are mapped to the sets V and U using a randomized LSH function H that depends on a hash parameter ϵ. Each node is assigned a weight that is proportional to the number of hash collisions. Likewise, each edge between the vertices v_i and u_j has a weight proportional to the number of (X_k, Y_k) pairs mapped to the node pairs (v_i, u_j). For a given value of the hash parameter ϵ, a base estimator of MI is proposed as a weighted average of non-linearly transformed of the edge weights. The proposed EDGE estimator of MI is obtained by applying the method of weighted ensemble bias reduction [15], [17] to a set of base estimators with different hash parameters. This estimator is a non-trivial extension of the LSH divergence estimator defined in [18]. LSH-based methods have previously been used for KNN search and graph constructions problems [19], [20], and they result in fast and low complexity algorithms.

In the following we summarize our contributions:

- To the best of our knowledge this is first MI estimator that has linear complexity and can achieve the optimal MSE rate of $O(1/N)$.
- The proposed MI estimator provides a simplified and unified treatment of mixed continuous-discrete variables.

*This research was partially supported by ARO grant W911NF-15-1-0479.

This is due to the hash function approach that is adopted.

- The proposed dependence graph provides an intuitive way of understanding interdependencies in the data; e.g. sparsity of the graph implies a strong dependency between the covariates, while an equally weighted dense graph implies that the covariates are close to independent.
- An online MI estimation update rule is proposed that can achieve the optimal MSE rate $O(1/N)$ in amortized complexity of $O(1)$ per each update.

The rest of the paper is organized as follows. In SectionII, we introduce the general definition of mutual information and define the dependence graph. In Section III, we introduce the hash based MI estimator and give theory for the bias and variance. In section IV we introduce the ensemble dependence graph MI estimator (EDGE) and show how the ensemble estimation method can be used to improve the convergence rates. In Section V, we discuss the online version of the proposed MI estimator. Finally, in Section VII we validate our theoretical results using numerical experiments.

II. Mutual Information

In this section, we introduce the general mutual information function based on the f-divergence measure. Then, we define a consistent estimator for the mutual information function. Consider the probability measures P and Q on a Euclidean space $\mathcal{X}$. Let $g:(0,\infty)\to\mathbb{R}$ be a convex function with $g(1)=0$. The f-divergence between P and Q can be defined as follows [21], [22].

$$D(P\|Q) := \mathbb{E}_Q\left[g\left(\frac{dP}{dQ}\right)\right]. \tag{1}$$

Similar to [16], [21] we give a general definition of the mutual information function.

Mutual Information: Let $\mathcal{X}$ and $\mathcal{Y}$ be Euclidean spaces and let P_{XY} be a probability measure on the space $\mathcal{X}\times\mathcal{Y}$. For any measurable sets $A\subseteq\mathcal{X}$ and $B\subseteq\mathcal{Y}$, we define the marginal probability measures $P_X(A):=P_{XY}(A\times\mathcal{Y})$ and $P_Y(B):=P_{XY}(\mathcal{X}\times B)$. The general mutual information functional is defined as

$$\begin{aligned} I(X,Y) &:= D(P_{XY}\|P_XP_Y)\\ &= \mathbb{E}_{P_XP_Y}\left[g\left(\frac{dP_{XY}}{dP_XP_Y}\right)\right], \end{aligned} \tag{2}$$

where $\frac{dP_{XY}}{dP_XP_Y}$ is the Radon-Nikodym derivative, and $g:(0,\infty)\to\mathbb{R}$ is, as in (1) a convex function with $g(1)=0$. Shannon mutual information is a particular cases of (1) for which $g(x)=x\log x$.

A. *Dependence Graphs*

Consider N i.i.d samples (X_i,Y_i), $1\le i\le N$ drawn from the probability measure P_{XY}, defined on the space $\mathcal{X}\times\mathcal{Y}$. Define the sets $\mathbf{X}=\{X_1,X_2,...,X_N\}$ and $\mathbf{Y}=\{Y_1,Y_2,...,Y_N\}$. The dependence graph $G(X,Y)$ is a directed bipartite graph, consisting of two sets of nodes V and U with cardinalities denoted as $|V|$ and $|U|$, and the set of edges E_G. Each point in the sets $\mathbf{X}$ and $\mathbf{Y}$ are mapped to the nodes in the sets U and V, respectively, using the hash function H, described as follows.

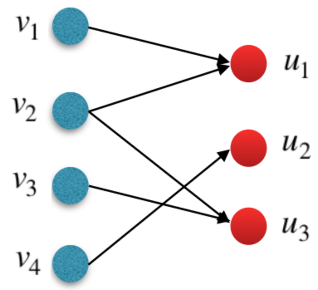

Fig. 1. Sample dependence graph with 4 and 3 respective distinct hash values of $\mathbf{X}$ and $\mathbf{Y}$ nodes, and the corresponding dependency edges.

A vector valued hash function H is defined in a similar way as defined in [18]. First, define the vector valued hash function $H_1:\mathbb{R}^d\to\mathbb{Z}^d$ as

$$H_1(x) = [h_1(x_1), h_1(x_2), ..., h_1(x_d)], \tag{3}$$

where x_i denotes the ith component of the vector x. In (3), each scalar hash function $h_1(x):\mathbb{R}\to\mathbb{Z}$ is given by

$$h_1(x) = \left\lfloor \frac{x}{\epsilon} \right\rfloor, \tag{4}$$

for a fixed $\epsilon>0$, where $\lfloor y\rfloor$ denotes the floor function (the smallest integer value less than or equal to y). Let $\mathcal{F}:=\{1,2,..,F\}$, where $F:=c_HN$ and c_H is a fixed tunable integer. We define a random hash function $H_2:\mathbb{Z}^d\to\mathcal{F}$ with a uniform density on the output and consider the combined hashing function

$$H(x) := H_2(H_1(x)), \tag{5}$$

which maps the points in $\mathbb{R}^d$ to $\mathcal{F}$.

$H(x)$ reveals the index of the mapped vertex in $G(X,Y)$. The weights ω_i and ω'_j corresponding to the nodes v_i and u_j, and ω_{ij}, the weight of the edge (v_i,u_j), are defined as follows.

$$\omega_i = \frac{N_i}{N}, \qquad \omega'_j = \frac{M_j}{N}, \qquad \omega_{ij} = \frac{N_{ij}N}{N_iM_j}, \tag{6}$$

where N_i and M_j respectively are the the number of hash collisions at the vertices v_i and u_j, and N_{ij} is the number of joint collisions of the nodes (X_k,Y_k) at the vertex pairs (v_i,u_j). The number of hash collisions is defined as the number of instances of the input variables map to the same output value. In particular,

$$N_{ij} := \#\left\{(X_k,Y_k) \text{ s.t } H(X_k)=i \text{ and } H(Y_k)=j\right\}. \tag{7}$$

Figure 1 represents a sample dependence graph. Note that nodes and edges with zero collisions do not show up in the dependence graph. The following theorem states an upper bound on the number of vertices in V and U.

Lemma 2.1. *Cardinality of the sets U and V are upper bounded as $|V|\le O(\epsilon^{-d})$ and $|U|\le O(\epsilon^{-d})$, respectively.*

III. The Base Estimator of MI

A. Assumptions

The following are the assumptions we make on the probability measures and g:

A1. The support sets $\mathcal{X}$ and $\mathcal{Y}$ are bounded.

A2. The following supremum exists and is bounded:

$$\sup_{P_X P_Y} g\left(\frac{dP_{XY}}{dP_X P_Y}\right) \leq U.$$

Let x_D and x_C respectively denote the discrete and continuous components of the vector x. Also let $f_{X_C}(x_C)$ and $p_{X_D}(x_D)$ respectively denote density and pmf functions of these components associated with the probability measure P_X.

A3. Assume that the density functions $f_{X_C}(x_C)$, $f_{Y_C}(y_C)$, $f_{X_C Y_C}(x_C, y_C)$, and the conditional densities $f_{X_C|X_D}(x_C|x_D)$, $f_{Y_C|Y_D}(y_C|y_D)$, $f_{X_C Y_C|X_D Y_D}(x_C, y_C|x_D, y_D)$ are Hölder continuous.

Hölder continuous functions: Given a support set $\mathcal{X}$, a function $f : \mathcal{X} \to \mathbb{R}$ is called Hölder continuous with parameter $0 < \gamma \leq 1$, if there exists a positive constant G_f, possibly depending on f, such that

$$|f(y) - f(x)| \leq G_f \|y - x\|^{\gamma}, \tag{8}$$

for every $x \neq y \in \mathcal{X}$.

A4. Assume that the function g in (2) is Lipschitz continuous; i.e. g is Hölder continuous with $\gamma = 1$.

B. The Base Estimator of MI

For a fixed value of the hash parameter ϵ, we propose the following base estimator of MI (2) function based on the dependence graph:

$$\widehat{I}(X, Y) := \max \left\{ \sum_{e_{ij} \in E_G} \omega_i \omega'_j \widetilde{g}(\omega_{ij}) \, , 0 \right\}, \tag{9}$$

where the summation is over all edges $e_{ij} : (v_i \to u_j)$ of $G(X, Y)$ having non-zero weight and $\widetilde{g}(x) := \max\{g(x), U\}$.

When X and Y are strongly dependent, each point X_k hashed into the bucket (vertex) v_i corresponds to a unique hash value for Y_k in U. Therefore, asymptotically $\omega_{ij} \to 1$ and the mutual information estimation in (9) takes its maximum value. On the other hand, when X and Y are independent, each point X_k hashed into the bucket (vertex) v_i may be associated with different values of Y_k, and therefore asymptotically $\omega_{ij} \to \omega_j$ and the Shannon MI estimation tends to 0.

In Fig. 2 (a) and 2 (b) the corresponding Shannon MI will converge to 0 and $log(3)$ when X and Y are independent or completely dependent, respectively.

C. Convergence Results

In the following theorems we state upper bounds on the bias and variance rates of the proposed MI estimator (9). We define the notations $\mathbb{B}[\hat{T}] = \mathbb{E}[\hat{T}] - T$ for bias and $\mathbb{V}[\hat{T}] = \mathbb{E}[\hat{T}^2] - \mathbb{E}[\hat{T}]^2$ for variance of $\hat{T}$. The following theorem states an upper bound on the bias.

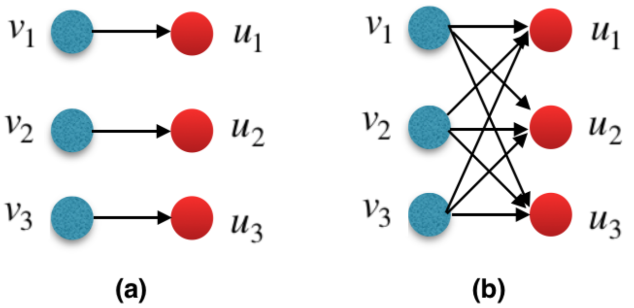

Fig. 2. Sample dependence graph with 3 distinct hash values of **X** and **Y** nodes, when X and Y have (a) complete dependence, (b) complete independence relation.

Theorem 3.1. *Let $d = d_X + d_Y$ be the dimension of the joint random variable (X, Y). Under the aforementioned assumptions **A1-A4**, and assuming that the density functions in **A3** have bounded derivatives up to the order $q \geq 0$, the following upper bound on the bias of the estimator in* (9) *holds*

$$\mathbb{B}\left[\widehat{I}(X, Y)\right] = \begin{cases} O(\epsilon^{\gamma}) + O\left(\frac{1}{N\epsilon^d}\right), & q = 0 \\ \sum_{i=1}^{q} C_i \epsilon^i + O(\epsilon^q) + O\left(\frac{1}{N\epsilon^d}\right) & q \geq 1, \end{cases} \tag{10}$$

where ϵ is the hash parameter in (4)*, γ is the smoothness parameter in* (8)*, and C_i are real constants.*

In (10), the hash parameter, ϵ needs to be a function of N to ensure that the bias converges to zero. For the case of $q = 0$, the optimum bias is achieved when $\epsilon = \left(\frac{1}{N}\right)^{\gamma/(\gamma+d)}$. When $q \geq 1$, the optimum bias is achieved for $\epsilon = \left(\frac{1}{N}\right)^{1/(1+d)}$.

Theorem 3.2. *Under the assumptions **A1-A4** the variance of the proposed estimator can be bounded as*

$$\mathbb{V}\left[\widehat{I}(X, Y)\right] \leq O\left(\frac{1}{N}\right). \tag{11}$$

Further, the variance of the variable ω_{ij} is also upper bounded by $O(1/N)$.

IV. Ensemble Dependence Graph Estimator (EDGE)

Given the expression for the bias in Theorem 3.1, the ensemble estimation technique proposed in [17] can be applied to improve the convergence rate of the MI estimator (9). Assume that the densities in **A3** have continuous bounded derivatives up to the order q, where $q \geq d$. Let $\mathcal{T} := \{t_1, ..., t_T\}$ be a set of index values with $t_i < c$, where $c > 0$ is a constant. Let $\epsilon(t) := tN^{-1/2d}$. For a given set of weights $w(t)$ the weighted ensemble estimator is then defined as

$$\widehat{I}_w := \sum_{t \in \mathcal{T}} w(t) \widehat{I}_{\epsilon(t)}, \tag{12}$$

Algorithm 1: MI Dependence Graph Estimator

Input : N i.i.d samples (X_k, Y_k), $1 \leq k \leq N$.

for *each* $k \in 1:N$ **do**
  $i \leftarrow H(\mathbf{X_k})$
  $j \leftarrow H(\mathbf{Y_k})$
  $N_i \leftarrow N_i + 1$
  $M_j \leftarrow M_j + 1$
  $N_{ij} \leftarrow N_{ij} + 1$
$\omega_i \leftarrow N_i/N; \omega'_j \leftarrow M_j/N; \omega_{ij} \leftarrow N_{ij}N/N_iM_j;$
$\widehat{I} \leftarrow \max\left\{\sum_{e_{ij}} \omega_i\omega'_j\widetilde{g}(\omega_{ij}), 0\right\}$

Output : $\widehat{I}$

where $\widehat{I}_{\epsilon(t)}$ is the mutual information estimator with the parameter $\epsilon(t)$. Using (10), for $q > 0$ the bias of the weighted ensemble estimator (12) takes the form

$$\mathbb{B}(\hat{I}_w) = \sum_{i=1}^{q} CiN^{-\frac{i}{2d}} \sum_{t\in\mathcal{T}} w(t)t^i + O\left(\frac{t^d}{N^{1/2}}\right) + O\left(\frac{1}{N\epsilon^d}\right) \tag{13}$$

Given the form (13) we can select the weights $w(t)$ to force to zero the slowly decaying terms in (13), i.e. $\sum_{t\in\tau} w(t)t^i = 0$ subject to the constraint that $\sum_{t\in\tau} w(t) = 1$. In particular we have the following theorem:

Theorem 4.1. *For $T > d$ let w_0 be the solution to:*

$$\begin{aligned} \min_{w} \quad & \|w\|_2 \\ \textit{subject to} \quad & \sum_{t\in\mathcal{T}} w(t) = 1, \\ & \sum_{t\in\mathcal{T}} w(t)t^i = 0, i \in \mathbb{N}, i \leq d. \end{aligned} \tag{14}$$

Then the MSE rate of the ensemble estimator $\widehat{I}_{w_0}$ is $O(1/N)$.

V. Online Learning of Mutual Information

Using a similar approach as the one developed in [18] for divergence estimation, we propose an efficient online mutual information estimator. Observed is the joint data stream $\{(X_k, Y_k)\}_{k>0}$. A brute force batch approach to computing the mutual information estimation for the updated data set recomputes the estimate over entire data set and requires $O(N)$ per update. Below we propose an algorithm that only requires $O(1)$ time complexity on average. The proposed update method exploits the fact that as a new pair (X_{N+1}, Y_{N+1}) comes in, only two nodes and one edge are affected in the bipartite dependency graph. However, note that if the hashing parameter ϵ is not updated at each step, by Theorem 3.1, the bias will not be optimally reduced. Hence obtaining an exact update in $O(1)$ is not trivial since the update of ϵ requires $O(N)$ time complexity. Similarly to [18] and we update ϵ only when $N+1$ is a power of 2. Similar to Theorem 3.1 of [18], we can then show that the MSE rate of this algorithm is of optimal order $O(1/N)$ and the amortized complexity of each update is $O(1)$.

VI. Proof Sketch

In this section we provide a proof sketch of the bias bound stated in Theorem 3.1, and briefly discuss the proof of the variance bound and the ensemble estimation MSE bound of Theorems 3.2 and 4.1. The proof of Theorem 3.2 is based on resampling and Efron-Stein inequality following a similar argument to that of [18]. Theorem 4.1 can be proved by directly applying the ensemble estimation theorem in [17] (Theorem 4) to the bias and variance bounds in (10) and (11). The complete proofs of the theorems and lemmas, are provided in the Appendix as follows:

- Appendix. A: Proofs of Lemma 2.1 and Theorem 3.1.
- Appendix. B: Proof of Theorem 3.2.
- Appendix. C: Proof of Theorem 4.1.

The bias proof is based on analyzing the hash function defined in (5). The proof consists of two main steps: 1) Finding the expectation of hash collisions of H_1; and 2) Analyzing the collision error of H_2. An important point about H_1 and H_2 is that collision of H_1 plays a crucial role in our estimator, while the collision of H_2 adds extra bias to the estimator. The following lemma states an upper bound on the bias error caused by H_2.

Lemma 6.1. *The bias error due to collision of H_2 is upper bounded as*

$$\mathbb{B}_H \leq O\left(\frac{1}{\epsilon^d N}\right). \tag{15}$$

Secondly, similar to [23], a key idea in proving Theorem 3.1 is connecting the edge weights in (9) to the Radon-Nikodym derivative $\frac{dP_{XY}}{dP_XP_Y}$. This fact is stated as follows:

Lemma 6.2. *Under the assumptions **A1-A4**, and assuming that the density functions in **A3** have bounded derivatives up to order $q \geq 0$ we have:*

$$\mathbb{E}\left[\omega_{ij}\right] = \frac{dP_{XY}}{dP_XP_Y} + \mathbb{B}(N, \epsilon, q, \gamma), \tag{16}$$

where

$$\mathbb{B}(N, \epsilon, q, \gamma) := \begin{cases} O(\epsilon^\gamma) + O\left(\frac{1}{N\epsilon^d}\right), & q = 0 \\ \sum_{i=1}^{q} C_i\epsilon^i + O(\epsilon^q) + O\left(\frac{1}{N\epsilon^d}\right), & q \geq 1, \end{cases} \tag{17}$$

and C_i are real constants.

VII. Numerical Validation

In this section we use simulated data to compare the proposed estimator to the competing MI estimators Ensemble KDE (EKDE) [15], and generalized KSG [16]. Both of these estimators work on mixed continuous-discrete variables.

Fig. 3, shows the MSE estimation rate of Shannon MI between the continuous random variables X and Y having the relation $Y = X + aN_U$, where X is a $2D$ Gaussian random variable with the mean $[0, 0]$ and covariance matrix $C = I_2$. Here I_d denote the d-dimensional identity matrix. N_U is a uniform random vector with the support $\mathcal{N}_U = [0, 1] \times [0, 1]$.

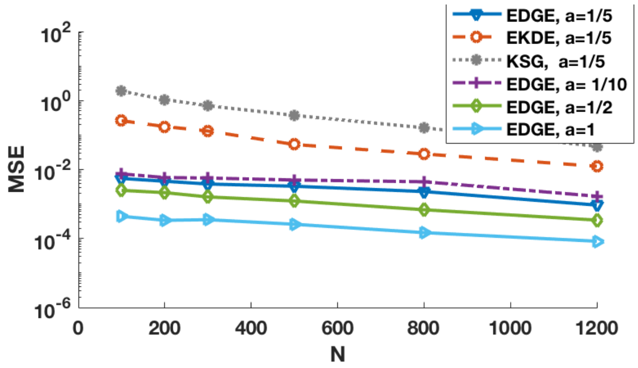

Fig. 3. MSE comparison of EDGE, EDKE and KSG Shannon MI estimators. X is a $2D$ Gaussian random variable with the expectation $[0, 0]$ and covariance matrix $C = I_2$. $Y = X + aN_U$, where N_U is a uniform noise. The MSE rates of EDGE, EKDE and KSG are compare for $a = 1/5$. Further, the MSE rate of EDGE is investigated for various noise levels of $a = 1/10$, $a = 1/2$ and $a = 1$.

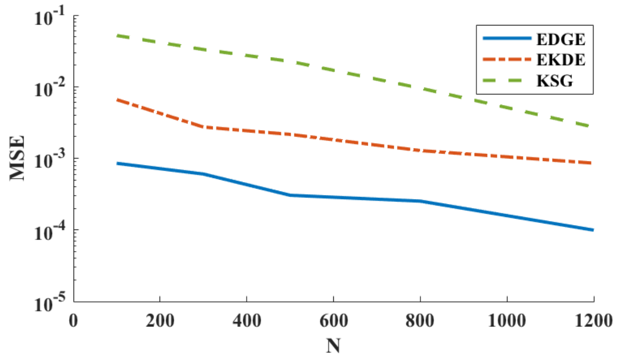

Fig. 4. MSE comparison of EDGE, EDKE and KSG Shannon MI estimators. $X \in \{1, 2, 3, 4\}$, and each $X = x$ is associated with multivariate Gaussian random vector Y, with $d = 4$, the expectation $[x/2, 0, 0, 0]$ and covariance matrix $C = I_4$.

We compute the MSE of each estimator for different sample sizes. The MSE rates of EDGE, EKDE and KSG are compared for $a = 1/5$. Further, the MSE rate of EDGE is investigated for noise levels of $a = \{1/10, 1/5, 1/2, 1\}$. As the dependency between X and Y increases the MSE rate becomes slower.

Fig. 4, shows the MSE estimation rate of Shannon MI between a discrete random variables X and a continuous random variable Y. We have $X \in \{1, 2, 3, 4\}$, and each $X = x$ is associated with multivariate Gaussian random vector Y, with $d = 4$, the expectation $[x/2, 0, 0, 0]$ and covariance matrix $C = I_4$. In general in Figures 3 and 4, EDGE has better performance than EKDE and KSG estimators.

VIII. Conclusion

In this paper we proposed a fast non-parametric estimation method for MI based on random hashing, dependence graphs, and ensemble estimation. Remarkably, the proposed estimator has linear computational complexity and attains optimal (parametric) rates of MSE convergence. We provided bias and variance convergence rate, and validated our results by numerical experiments.

References

[1] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[2] H. Liu, L. Liu, and H. Zhang, "Ensemble gene selection for cancer classification," *Pattern Recognition*, vol. 43, no. 8, pp. 2763–2772, 2010.

[3] A. Kraskov, H. Stögbauer, R. G. Andrzejak, and P. Grassberger, "Hierarchical clustering using mutual information," *EPL (Europhysics Letters)*, vol. 70, no. 2, p. 278, 2005.

[4] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, "Feature selection: A data perspective," *ACM Computing Surveys (CSUR)*, vol. 50, no. 6, p. 94, 2017.

[5] K. R. Moon, M. Noshad, S. Y. Sekeh, and A. O. Hero, "Information theoretic structure learning with confidence," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pp. 6095–6099, IEEE, 2017.

[6] J. P. Pluim, J. A. Maintz, and M. A. Viergever, "Mutual-information-based registration of medical images: a survey," *IEEE Transactions on Medical Imaging*, vol. 22, no. 8, pp. 986–1004, 2003.

[7] S. Krishnaswamy, M. H. Spitzer, M. Mingueneau, S. C. Bendall, O. Litvin, E. Stone, D. Peér, and G. P. Nolan, "Conditional density-based analysis of t cell signaling in single-cell data," *Science*, vol. 346, no. 6213, p. 1250689, 2014.

[8] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical Review E*, vol. 69, no. 6, p. 066138, 2004.

[9] Y. Moon, B. Rajagopalan, and U. Lall, "Estimation of mutual information using kernel density estimators," *Physical Review E*, vol. 52, no. 3, p. 2318, 1995.

[10] N. Kwak and C.-H. Choi, "Input feature selection by mutual information based on parzen window," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 12, pp. 1667–1671, 2002.

[11] G. A. Darbellay and I. Vajda, "Estimation of the information by an adaptive partitioning of the observation space," *IEEE Transactions on Information Theory*, vol. 45, no. 4, pp. 1315–1321, 1999.

[12] S. Khan, S. Bandyopadhyay, A. R. Ganguly, S. Saigal, D. J. Erickson III, V. Protopopescu, and G. Ostrouchov, "Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data," *Physical Review E*, vol. 76, no. 2, p. 026209, 2007.

[13] K. Kandasamy, A. Krishnamurthy, B. Poczos, L. Wasserman, *et al.*, "Nonparametric von mises estimators for entropies, divergences and mutual informations," in *Advances in Neural Information Processing Systems*, pp. 397–405, 2015.

[14] S. Singh and B. Póczos, "Exponential concentration of a density functional estimator," in *Advances in Neural Information Processing Systems*, pp. 3032–3040, 2014.

[15] K. R. Moon, K. Sricharan, and A. O. Hero III, "Ensemble estimation of mutual information," *Proceedings of the IEEE Intl Symp. on Information Theory (ISIT), Aachen, June 2017.*

[16] W. Gao, S. Kannan, S. Oh, and P. Viswanath, "Estimating mutual information for discrete-continuous mixtures," in *Advances in Neural Information Processing Systems*, pp. 5988–5999, 2017.

[17] K. R. Moon, K. Sricharan, K. Greenewald, and A. O. Hero, "Improving convergence of divergence functional ensemble estimators," in *IEEE International Symposium Inf Theory*, pp. 1133–1137, IEEE, 2016.

[18] M. Noshad and A. O. Hero III, "Scalable hash-based estimation of divergence measures," *Proceedings of the 22nd Conference on Artificial Intelligence and Statistics, Canary Islands, March 2018, arXiv:1801.00398.*

[19] Y. Zhang, K. Huang, G. Geng, and C. Liu, "Fast kNN graph construction with locality sensitive hashing," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 660–674, 2013.

[20] Q. Lv, W. Josephson, Z. Wang, M. Charikar, and K. Li, "Multi-probe LSH: efficient indexing for high-dimensional similarity search," in *Proceedings of the 33rd International Conference on Very Large Data Bases*, pp. 950–961, VLDB Endowment, 2007.

[21] Y. Polyanskiy and Y. Wu, "Strong data-processing inequalities for channels and bayesian networks," in *Convexity and Concentration*, pp. 211–249, Springer, 2017.

[22] I. Csiszár, "Generalized cutoff rates and renyi's information measures," *IEEE Tran on Information Theory*, vol. 41, no. 1, pp. 26–34, 1995.

[23] M. Noshad, K. R. Moon, S. Y. Sekeh, and A. O. Hero III, "Direct estimation of information divergence using nearest neighbor ratios," *Proc of the IEEE Intl Symp. on Information Theory (ISIT), Aachen, June 2017.*

A. BIAS PROOF

In this section we provide the proofs of Lemma 2.1 and Theorem 3.1.

Let $\left\{\widetilde{X}_i\right\}_{i=1}^{L_X}$ and $\left\{\widetilde{Y}_i\right\}_{i=1}^{L_Y}$ respectively denote distinct outputs of H_1 with the N i.i.d points X_k and Y_k as input. Then according to [18] (Lemma 4.1), we have

$$L_X \le O\big(\epsilon^{-d}\big), \quad L_Y \le O\big(\epsilon^{-d}\big). \tag{18}$$

Proof of Lemma 2.1. The number of distinct inputs of H_2 is greater than or equal to the number of its distinct outputs. So, $|V| \le L_X$ and $|U| \le L_Y$. Using the bounds in (18) completes the proof. ■

We define the following events:

$$\begin{aligned}
E_{ij} &: \text{The event that there is an edge between the vertices } v_i \text{ and } u_j.\\
E_{v_i}^{>k} &: \text{The event that there are at least } k+1 \text{ vectors from } \left\{\widetilde{X}_i\right\}_{i=1}^{L_X} \text{ that map to } v_i \text{ using } H_2\ .\\
E_{v_i}^{=k} &: \text{The event that there are exactly } k \text{ vectors from } \left\{\widetilde{X}_i\right\}_{i=1}^{L_X} \text{ that map to } v_i \text{ using } H_2\ .\\
E_{v_i}^{\le k} &: \text{The event that there are at most } k \text{ vectors from } \left\{\widetilde{X}_i\right\}_{i=1}^{L_X} \text{ that map to } v_i \text{ using } H_2\ .
\end{aligned} \tag{19}$$

$E_{u_i}^{>k}$, $E_{u_i}^{=k}$ and $E_{u_i}^{\le k}$ are defined similarly. Further, let for any event E, $\overline{E}$ denote the complementary event. Let $E_{ij}^{=k} := E_{v_i}^{=k} \cap E_{u_i}^{=k}$, $E_{ij}^{\le k} := E_{v_i}^{\le k} \cap E_{u_i}^{\le k}$ and $E_{ij}^{>k} := E_{v_i}^{>k} \cup E_{u_i}^{>k}$. Finally, we define $E^{=k} := \left(\cap_{i=k}^{L_X} E_{v_i}^{=k}\right) \cap \left(\cap_{j=k}^{L_Y} E_{u_j}^{=k}\right)$, which represent the event of no collision for $k = 1$.

Consider the notation $\widetilde{I}(X,Y) := \sum_{e_{ij}\in E_G} \omega_i \omega'_j \widetilde{g}(\omega_{ij})$ (Notice the difference from the definition in (9)). We can derive its expectation as

$$\begin{aligned}
\mathbb{E}\left[\widetilde{I}(X,Y)\right] &= \mathbb{E}\left[\sum_{e_{ij}\in E_G} \omega_i\omega'_j\widetilde{g}(\omega_{ij})\right]\\
&= \mathbb{E}\left[\sum_{i,j\in\mathcal{F}} \mathbb{1}_{E_{ij}}\omega_i\omega'_j\widetilde{g}(\omega_{ij})\right]\\
&= \sum_{i,j\in\mathcal{F}} P\left(E_{ij}^{\le 1}\right)\mathbb{E}\left[\mathbb{1}_{E_{ij}}\omega_i\omega'_j\widetilde{g}(\omega_{ij})\middle|E_{ij}^{\le 1}\right]\\
&\quad + \sum_{i,j\in\mathcal{F}} P\big(E_{ij}^{>1}\big)\,\mathbb{E}\left[\mathbb{1}_{E_{ij}}\omega_i\omega'_j\widetilde{g}(\omega_{ij})\middle|E_{ij}^{>1}\right]
\end{aligned} \tag{20}$$

Note that the second term in (20) is the bias due to collision of H_2, and as mentioned in Lemma 6.1, we denote this term by $\mathbb{B}_H$.

A. Bias Due to Collision

In the following we provide a lemma required for proving the bound on $\mathbb{B}_H$ stated in Lemma 6.1.

Lemma 8.1. *$P(E_{ij}^{=1}|E_{ij})$ is given by*

$$P(E_{ij}^{=1}|E_{ij}) = 1 - O\left(\frac{1}{\epsilon^d N}\right). \tag{21}$$

Proof. Let $\widetilde{\mathbf{X}} = \widetilde{\mathbf{x}}$ and $\widetilde{\mathbf{Y}} = \widetilde{\mathbf{y}}$ respectively abbreviate the equations $\widetilde{X}_1 = \widetilde{x}_1, ..., \widetilde{X}_{L_X} = \widetilde{x}_{L_X}$ and $\widetilde{Y}_1 = \widetilde{y}_1, ..., \widetilde{Y}_{L_Y} = \widetilde{y}_{L_Y}$. Let $\widetilde{\mathbf{x}} := \{\widetilde{x}_1, \widetilde{x}_2, ..., \widetilde{x}_{L_X}\}$ and $\widetilde{\mathbf{y}} := \{\widetilde{y}_1, \widetilde{y}_2, ..., \widetilde{y}_{L_Y}\}$. Define $\widetilde{\mathbf{z}} := \widetilde{\mathbf{x}} \cup \widetilde{\mathbf{y}}$ and $L_Z := |\widetilde{\mathbf{z}}|$.

$$P(E_{ij}^{=1}|E_{ij}) = \sum_{\widetilde{\mathbf{x}},\widetilde{\mathbf{y}}} P\left(\widetilde{\mathbf{X}} = \widetilde{\mathbf{x}}, \widetilde{\mathbf{Y}} = \widetilde{\mathbf{y}}|E_{ij}\right) P(E_{ij}^{=1}|E_{ij}, \widetilde{\mathbf{X}} = \widetilde{\mathbf{x}}, \widetilde{\mathbf{Y}} = \widetilde{\mathbf{y}}). \tag{22}$$

Define $a = 2$ for the case $i \neq j$ and $a = 1$ for the case $i = j$. Then we have

$$
\begin{aligned}
P(E_{ij}^{=1}|E_{ij}) &= \sum_{\widetilde{\mathbf{x}},\widetilde{\mathbf{y}}} P\Big(\widetilde{\mathbf{X}}=\widetilde{\mathbf{x}},\widetilde{\mathbf{Y}}=\widetilde{\mathbf{y}}|E_{ij}\Big)\, O\left(\left(\frac{F-a}{F}\right)^{L_Z-a}\right)\\
&= \sum_{\widetilde{\mathbf{x}},\widetilde{\mathbf{y}}} P\Big(\widetilde{\mathbf{X}}=\widetilde{\mathbf{x}},\widetilde{\mathbf{Y}}=\widetilde{\mathbf{y}}|E_{ij}\Big)\left(1-O\Big(\frac{L_Z}{F}\Big)\right)\\
&\le \sum_{\widetilde{\mathbf{x}},\widetilde{\mathbf{y}}} P\Big(\widetilde{\mathbf{X}}=\widetilde{\mathbf{x}},\widetilde{\mathbf{Y}}=\widetilde{\mathbf{y}}|E_{ij}\Big)\left(1-O\Big(\frac{L_X+L_Y}{F}\Big)\right)\\
&= \sum_{\widetilde{\mathbf{x}},\widetilde{\mathbf{y}}} P\Big(\widetilde{\mathbf{X}}=\widetilde{\mathbf{x}},\widetilde{\mathbf{Y}}=\widetilde{\mathbf{y}}|E_{ij}\Big)\left(1-O\Big(\frac{1}{\epsilon^d N}\Big)\right)\\
&= \left(1-O\Big(\frac{1}{\epsilon^d N}\Big)\right)\sum_{\widetilde{\mathbf{x}},\widetilde{\mathbf{y}}} P\Big(\widetilde{\mathbf{X}}=\widetilde{\mathbf{x}},\widetilde{\mathbf{Y}}=\widetilde{\mathbf{y}}|E_{ij}\Big)\\
&= \left(1-O\Big(\frac{1}{\epsilon^d N}\Big)\right),
\end{aligned}
\tag{23}
$$

where in the fourth line we have used (18). ■

Proof of 6.1. N_i' and M_j' respectively are defined as the number of the input points $\mathbf{X}$ and $\mathbf{Y}$ mapped to the buckets $\widetilde{X}_i$ and $\widetilde{Y}_j$ using H_1. Define $\mathcal{A}_i := \Big\{j : H_2(\widetilde{X}_j) = i\Big\}$ and $\mathcal{B}_i := \Big\{j : H_2(\widetilde{Y}_j) = i\Big\}$. For each i we can rewrite N_i and M_i as

$$
N_i = \sum_{j=1}^{L_X} \mathbb{1}_{\mathcal{A}_i}(j) N_j', \quad M_i = \sum_{j=1}^{L_Y} \mathbb{1}_{\mathcal{B}_i}(j) M_j'. \tag{24}
$$

Thus,

$$
\begin{aligned}
\mathbb{B}_H &\le \sum_{i,j\in\mathcal{F}} P\big(E_{ij}^{\geq 1}\big)\,\mathbb{E}\left[\mathbb{1}_{E_{ij}}\omega_i\omega_j'\widetilde{g}(\omega_{ij})\big|E_{ij}^{\geq 1}\right]\\
&= \sum_{i,j\in\mathcal{F}} P\big(E_{ij}^{\geq 1}\big)\Big(P\big(E_{ij}|E_{ij}^{\geq 1}\big)\,\mathbb{E}\left[\omega_i\omega_j'\widetilde{g}(\omega_{ij})\big|E_{ij}^{\geq 1},E_{ij}\right] + P\big(\overline{E_{ij}}|E_{ij}^{\geq 1}\big)\,\mathbb{E}\left[\omega_i\omega_j'\widetilde{g}(\omega_{ij})\big|E_{ij}^{\geq 1},\overline{E_{ij}}\right]\Big)\\
&= \sum_{i,j\in\mathcal{F}} P(E_{ij})\,P\big(E_{ij}^{\geq 1}|E_{ij}\big)\,\mathbb{E}\left[\omega_i\omega_j'\widetilde{g}(\omega_{ij})\big|E_{ij}^{\geq 1},E_{ij}\right] &(25)\\
&\le O\Big(\frac{U}{\epsilon^d N}\Big)\sum_{i,j\in\mathcal{F}} P(E_{ij})\,\mathbb{E}\left[\omega_i\omega_j'\big|E_{ij}^{\geq 1},E_{ij}\right] &(26)\\
&= O\Big(\frac{U}{\epsilon^d N^3}\Big)\sum_{i,j\in\mathcal{F}} P(E_{ij})\,\mathbb{E}\left[N_iM_j\big|E_{ij}^{\geq 1},E_{ij}\right]\\
&= O\Big(\frac{U}{\epsilon^d N^3}\Big)\sum_{\widetilde{\mathbf{x}},\widetilde{\mathbf{y}}} p_{\widetilde{\mathbf{X}},\widetilde{\mathbf{Y}}}(\widetilde{\mathbf{x}},\widetilde{\mathbf{y}})\sum_{i,j\in\mathcal{F}} P(E_{ij})\,\mathbb{E}\left[N_iM_j\Big|E_{ij}^{\geq 1},E_{ij},\widetilde{\mathbf{X}}=\widetilde{\mathbf{x}},\widetilde{\mathbf{Y}}=\widetilde{\mathbf{y}}\right]\\
&= O\Big(\frac{U}{\epsilon^d N^3}\Big)\sum_{\widetilde{\mathbf{x}},\widetilde{\mathbf{y}}} p_{\widetilde{\mathbf{X}},\widetilde{\mathbf{Y}}}\sum_{i,j\in\mathcal{F}} P(E_{ij})\,\mathbb{E}\left[\left(\sum_{r=1}^{L_X}\mathbb{1}_{\mathcal{A}_i}(r)N_r'\right)\left(\sum_{s=1}^{L_Y}\mathbb{1}_{\mathcal{B}_j}(s)M_s'\right)\Bigg|E_{ij}^{\geq 1},E_{ij},\widetilde{\mathbf{X}}=\widetilde{\mathbf{x}},\widetilde{\mathbf{Y}}=\widetilde{\mathbf{y}}\right] &(27)\\
&= O\Big(\frac{U}{\epsilon^d N^3}\Big)\sum_{\widetilde{\mathbf{x}},\widetilde{\mathbf{y}}} p_{\widetilde{\mathbf{X}},\widetilde{\mathbf{Y}}}\sum_{i,j\in\mathcal{F}} P(E_{ij})\sum_{r=1}^{L_X}\sum_{s=1}^{L_Y}\mathbb{E}\left[\big(\mathbb{1}_{\mathcal{A}_i}(r)\big)\big(\mathbb{1}_{\mathcal{B}_j}(s)\big)\Big|E_{ij}^{\geq 1},E_{ij},\widetilde{\mathbf{X}}=\widetilde{\mathbf{x}},\widetilde{\mathbf{Y}}=\widetilde{\mathbf{y}}\right]\mathbb{E}\left[N_r'M_s'|E_{ij},\widetilde{\mathbf{X}}=\widetilde{\mathbf{x}},\widetilde{\mathbf{Y}}=\widetilde{\mathbf{y}}\right]\\
&= O\Big(\frac{U}{\epsilon^d N^3}\Big)\sum_{\widetilde{\mathbf{x}},\widetilde{\mathbf{y}}} p_{\widetilde{\mathbf{X}},\widetilde{\mathbf{Y}}}\sum_{i,j\in\mathcal{F}} P(E_{ij})\sum_{r=1}^{L_X}\sum_{s=1}^{L_Y} P\Big(r\in\mathcal{A}_i, s\in\mathcal{B}_j\big|E_{ij}^{\geq 1},E_{ij},\widetilde{\mathbf{X}}=\widetilde{\mathbf{x}},\widetilde{\mathbf{Y}}=\widetilde{\mathbf{y}}\Big)\,\mathbb{E}\left[N_r'M_s'|E_{ij},\widetilde{\mathbf{X}}=\widetilde{\mathbf{x}},\widetilde{\mathbf{Y}}=\widetilde{\mathbf{y}}\right], &(28)
\end{aligned}
$$

where in (25) we have used the Bayes rule, and the fact that $\widetilde{g}(\omega_{ij}) = 0$ conditioned on the event $\overline{E_{ij}}$. In (26) we have used the bound in Lemma 8.1, and the upper bound on $\widetilde{g}(\omega_{ij})$. Equation (27) is due to (24). Now we simplify $P\Big(r\in\mathcal{A}_i, s\in\mathcal{B}_j\big|E_{ij}^{\geq 1},E_{ij},\widetilde{\mathbf{X}}=\widetilde{\mathbf{x}},\widetilde{\mathbf{Y}}=\widetilde{\mathbf{y}}\Big)$ in (28) as follows. First assume that $\tilde{X}_r \neq \tilde{Y}_s$.

$$P\Big(r \in \mathcal{A}_i, s \in \mathcal{B}_j \big| E_{ij}^{>1}, E_{ij}, \widetilde{\mathbf{X}} = \widetilde{\mathbf{x}}, \widetilde{\mathbf{Y}} = \widetilde{\mathbf{y}}\Big) \leq P\Big(r \in \mathcal{A}_i, s \in \mathcal{B}_j \big| E_{v_i}^{>1}, E_{u_j}^{>1}, \widetilde{\mathbf{X}} = \widetilde{\mathbf{x}}, \widetilde{\mathbf{Y}} = \widetilde{\mathbf{y}}\Big)$$
$$= P\Big(r \in \mathcal{A}_i \big| E_{v_i}^{>1}, \widetilde{\mathbf{X}} = \widetilde{\mathbf{x}}\Big) P\Big(s \in \mathcal{B}_j \big| E_{u_j}^{>1}, \widetilde{\mathbf{Y}} = \widetilde{\mathbf{y}}\Big), \tag{29}$$

where the second line is because the hash function H_2 is random and independent for different inputs. $P\Big(r \in \mathcal{A}_i \big| E_{v_i}^{>1}, \widetilde{\mathbf{X}} = \widetilde{\mathbf{x}}\Big)$ in (29) can be written as

$$P\Big(r \in \mathcal{A}_i \big| E_{v_i}^{>1}, \widetilde{\mathbf{X}} = \widetilde{\mathbf{x}}\Big) = \frac{P\Big(r \in \mathcal{A}_i, E_{v_i}^{>1} \big| \widetilde{\mathbf{X}} = \widetilde{\mathbf{x}}\Big)}{P\Big(E_{v_i}^{>1} \big| \widetilde{\mathbf{X}} = \widetilde{\mathbf{x}}\Big)}. \tag{30}$$

We first find $P\Big(E_{v_i}^{>1} \big| \widetilde{\mathbf{X}} = \widetilde{\mathbf{x}}\Big)$:

$$\begin{aligned} P\Big(E_{v_i}^{>1} \big| \widetilde{\mathbf{X}} = \widetilde{\mathbf{x}}\Big) &= 1 - P\Big(E_{v_i}^{=0} \big| \widetilde{\mathbf{X}} = \widetilde{\mathbf{x}}\Big) - P\Big(E_{v_i}^{=1} \big| \widetilde{\mathbf{X}} = \widetilde{\mathbf{x}}\Big) \\ &= 1 - \left(\frac{F-1}{F}\right)^{L_X} - \left(\frac{L_X}{F}\left(\frac{F-1}{F}\right)^{L_x - 1}\right) \\ &= \frac{L_X^2}{2F^2} + o\left(\frac{L_X^2}{2F^2}\right). \end{aligned} \tag{31}$$

Next, we find $P\Big(r \in \mathcal{A}_i, E_{v_i}^{>1} \big| \widetilde{\mathbf{X}} = \widetilde{\mathbf{x}}\Big)$ in (30) as follows.

$$\begin{aligned} P\Big(r \in \mathcal{A}_i, E_{v_i}^{>1} \big| \widetilde{\mathbf{X}} = \widetilde{\mathbf{x}}\Big) &= P\Big(E_{v_i}^{>1} \big| r \in \mathcal{A}_i, \widetilde{\mathbf{X}} = \widetilde{\mathbf{x}}\Big) P\Big(r \in \mathcal{A}_i \big| \widetilde{\mathbf{X}} = \widetilde{\mathbf{x}}\Big) \\ &= \left(1 - \left(\frac{F-1}{F}\right)^{L_X - 1}\right)\left(\frac{1}{F}\right) = O\left(\frac{L_X}{F^2}\right) \end{aligned} \tag{32}$$

Thus, using (31) and (32) yields

$$P\Big(r \in \mathcal{A}_i \big| E_{v_i}^{>1}, \widetilde{\mathbf{X}} = \widetilde{\mathbf{x}}\Big) = O\left(\frac{1}{L_X}\right). \tag{33}$$

Similarly, we have

$$P\Big(s \in \mathcal{B}_j \big| E_{u_j}^{>1}, \widetilde{\mathbf{Y}} = \widetilde{\mathbf{y}}\Big) = O\left(\frac{1}{L_Y}\right). \tag{34}$$

Now assume the case $\tilde{X}_r = \tilde{Y}_s$. Then since $H_2(\tilde{X}_r) = H_2(\tilde{Y}_s)$, we can simplify $P\Big(r \in \mathcal{A}_i, s \in \mathcal{B}_j \big| E_{ij}^{>1}, E_{ij}, \widetilde{\mathbf{X}} = \widetilde{\mathbf{x}}, \widetilde{\mathbf{Y}} = \widetilde{\mathbf{y}}\Big)$ in (28) as

$$P\Big(r \in \mathcal{A}_i, s \in \mathcal{B}_j \big| E_{ij}^{>1}, E_{ij}, \widetilde{\mathbf{X}} = \widetilde{\mathbf{x}}, \widetilde{\mathbf{Y}} = \widetilde{\mathbf{y}}\Big) = \delta_{ij} P\Big(r \in \mathcal{A}_i \big| E_{v_i}^{>1}, \widetilde{\mathbf{X}} = \widetilde{\mathbf{x}}, \widetilde{\mathbf{Y}} = \widetilde{\mathbf{y}}\Big). \tag{35}$$

Recalling the definition $\widetilde{\mathbf{z}} := \widetilde{\mathbf{x}} \cup \widetilde{\mathbf{y}}$ and $L_Z := |\widetilde{\mathbf{z}}|$, similar to

$$P\Big(r \in \mathcal{A}_i \big| E_{v_i}^{>1}, \widetilde{\mathbf{X}} = \widetilde{\mathbf{x}}, \widetilde{\mathbf{Y}} = \widetilde{\mathbf{y}}\Big) = O\left(\frac{1}{L_Z}\right). \tag{36}$$

By using equations (29), (33), (34) and (36) in (28), we can write the following upper bound for the bias estimator due to collision.

$$
\begin{aligned}
\mathbb{B}_H &\leq O\left(\frac{U}{\epsilon^d N^3}\right) \sum_{\widetilde{\mathbf{x}},\widetilde{\mathbf{y}}} p_{\widetilde{\mathbf{X}},\widetilde{\mathbf{Y}}} \sum_{i,j\in\mathcal{F}} P(E_{ij}) \sum_{r=1}^{L_X}\sum_{s=1}^{L_Y} \mathbb{E}\left[N'_r M'_s | E_{ij}, \widetilde{\mathbf{X}}=\widetilde{\mathbf{x}}, \widetilde{\mathbf{Y}}=\widetilde{\mathbf{y}}\right]\left(O\left(\frac{1}{L_X L_Y}\right)+\delta_{ij} O\left(\frac{1}{L_Z}\right)\right)\\
&= O\left(\frac{U}{\epsilon^d N^3}\right) \sum_{\widetilde{\mathbf{x}},\widetilde{\mathbf{y}}} p_{\widetilde{\mathbf{X}},\widetilde{\mathbf{Y}}} \sum_{i,j\in\mathcal{F}} P(E_{ij})\, \mathbb{E}\left[\sum_{r=1}^{L_X} N'_r \sum_{s=1}^{L_Y} M'_s | E_{ij}, \widetilde{\mathbf{X}}=\widetilde{\mathbf{x}}, \widetilde{\mathbf{Y}}=\widetilde{\mathbf{y}}\right]\left(O\left(\frac{1}{L_X L_Y}\right)+\delta_{ij} O\left(\frac{1}{L_Z}\right)\right)\\
&= O\left(\frac{U}{\epsilon^d N^3}\right) \sum_{\widetilde{\mathbf{x}},\widetilde{\mathbf{y}}} p_{\widetilde{\mathbf{X}},\widetilde{\mathbf{Y}}} \sum_{i,j\in\mathcal{F}} P(E_{ij})\, N^2\left(O\left(\frac{1}{L_X L_Y}\right)+\delta_{ij} O\left(\frac{1}{L_Z}\right)\right)\\
&= O\left(\frac{U}{\epsilon^d N^3}\right) \sum_{\widetilde{\mathbf{x}},\widetilde{\mathbf{y}}} p_{\widetilde{\mathbf{X}},\widetilde{\mathbf{Y}}} \left(O\left(\frac{N^2}{L_X L_Y}\right)+O\left(\frac{N}{L_Z}\right)\right) \sum_{i,j\in\mathcal{F}} P(E_{ij})\\
&= O\left(\frac{U}{\epsilon^d N^3}\right) \sum_{\widetilde{\mathbf{x}},\widetilde{\mathbf{y}}} p_{\widetilde{\mathbf{X}},\widetilde{\mathbf{Y}}} \left(O\left(\frac{N^2}{L_X L_Y}\right)+O\left(\frac{N}{L_Z}\right)\right) \mathbb{E}\left[\sum_{i,j\in\mathcal{F}} \mathbb{1}_{E_{ij}}\right]\\
&\leq O\left(\frac{U}{\epsilon^d N^3}\right) \sum_{\widetilde{\mathbf{x}},\widetilde{\mathbf{y}}} p_{\widetilde{\mathbf{X}},\widetilde{\mathbf{Y}}} \left(O\left(\frac{N^2}{L_X L_Y}\right)+O\left(\frac{N}{L_Z}\right)\right)(L_X L_Y)\\
&\leq O\left(\frac{1}{\epsilon^d N}\right).
\end{aligned}
\tag{37}
$$

■

B. Bias without Collision

As mentioned in Section VI, a key idea in proving Theorem 3.1 is to show that the expectation of the edge weights ω_{ij} are proportional to the Radon-Nikodym derivative $dP_{XY}/dP_X P_Y$ at the points that correspond to the vertices v_i and u_j. Note that since $\omega_{ij} = N_{ij}N/N_i M_j$, and N_{ij}, N_i and N_j are not independent variables, deriving the expectation is not trivial. In the following we give a lemma that provides conditions under which the expectation of a function of random variables is close to the function of expectations of the random variables. We will use the following lemma to simplify $\mathbb{E}\left[\omega_{ij}\right]$.

Lemma 8.2. *Assume that $g(Z_1, Z_2, ..., Z_k) : \mathcal{Z}_1 \times ... \times \mathcal{Z}_k \to R$ is a Lipschitz continuous function with constant $H_g > 0$ with respect to each of variables $Z_i, 1 \leq i \leq k$. Let $\mathbb{V}[Z_i]$ and $\mathbb{V}[Z_i|X]$ respectively denote the variance and the conditional variance of each variable Z_i for a given variable X. Then we have*

$$
\textbf{\textit{a)}}\quad \left|\mathbb{E}\left[g(Z_1, Z_2, ..., Z_k)\right] - g(\mathbb{E}\left[Z_1\right], \mathbb{E}\left[Z_2\right], ..., \mathbb{E}\left[Z_k\right])\right| \leq H_g \sum_{i=1}^{k} \sqrt{\mathbb{V}[Z_i]}, \tag{38}
$$

$$
\textbf{\textit{b)}}\quad \left|\mathbb{E}\left[g(Z_1, Z_2, ..., Z_k)\,|X\right] - g(\mathbb{E}\left[Z_1|X\right], \mathbb{E}\left[Z_2|X\right], ..., \mathbb{E}\left[Z_k|X\right])\right| \leq H_g \sum_{i=1}^{k} \sqrt{\mathbb{V}[Z_i|X]}. \tag{39}
$$

Proof.

$$
\begin{aligned}
|\mathbb{E}\left[g(Z_1,Z_2,...,Z_k)\right]-g(\mathbb{E}\left[Z_1\right],\mathbb{E}\left[Z_2\right],...,\mathbb{E}\left[Z_k\right])| &= |\mathbb{E}\left[g(Z_1,Z_2,...,Z_k)-g(\mathbb{E}\left[Z_1\right],\mathbb{E}\left[Z_2\right],...,\mathbb{E}\left[Z_k\right])\right]| \\
&\le \mathbb{E}\left[|g(Z_1,Z_2,...,Z_k)-g(\mathbb{E}\left[Z_1\right],\mathbb{E}\left[Z_2\right],...,\mathbb{E}\left[Z_k\right])|\right] \qquad (40)\\
&\le \mathbb{E}[|g(Z_1,Z_2,...,Z_k)-g(\mathbb{E}\left[Z_1\right],Z_2,...,Z_k)+\\
&\quad + g(\mathbb{E}\left[Z_1\right],Z_2,...,Z_k)-g(\mathbb{E}\left[Z_1\right],\mathbb{E}\left[Z_2\right],...,Z_k)\\
&\quad + ...\\
&\quad + g(\mathbb{E}\left[Z_1\right],\mathbb{E}\left[Z_2\right],...,\mathbb{E}\left[Z_{k-1}\right],Z_k)-g(\mathbb{E}\left[Z_1\right],\mathbb{E}\left[Z_2\right],...,\mathbb{E}\left[Z_k\right])\,|]\\
&\le \mathbb{E}\left[\left|g(Z_1,Z_2,...,Z_k)-g(\mathbb{E}\left[Z_1\right],Z_2,...,Z_k)\right|\right]\\
&\quad + \mathbb{E}\left[\left|g(\mathbb{E}\left[Z_1\right],Z_2,...,Z_k)-g(\mathbb{E}\left[Z_1\right],\mathbb{E}\left[Z_2\right],...,Z_k)\right|\right]\\
&\quad + ...\\
&\quad + \mathbb{E}\left[\left|g(\mathbb{E}\left[Z_1\right],...,\mathbb{E}\left[Z_{k-1}\right],Z_k)-g(\mathbb{E}\left[Z_1\right],...,\mathbb{E}\left[Z_k\right])\right|\right] \qquad (41)\\
&\le H_g\mathbb{E}\left[|Z_1-\mathbb{E}\left[Z_1\right]|\right]+H_g\mathbb{E}\left[|Z_2-\mathbb{E}\left[Z_2\right]|\right]+...+H_g\mathbb{E}\left[|Z_k-\mathbb{E}\left[Z_k\right]|\right] \qquad (42)\\
&\le H_g\sum_{i=1}^{k}\sqrt{\mathbb{V}[Z_i]}. \qquad (43)
\end{aligned}
$$

In (40) and (41) we have used triangle inequalities. In (42) we have applied Lipschitz condition, and finally in (43) we have used Cauchy-Schwarz inequality. Since the proofs of parts (a) and (b) are similar, we omit the proof of part (b). ■

Lemma 8.3. *Define $\nu_{ij}=N_{ij}/N$, and recall the definitions $\omega_{ij}=N_{ij}N/N_iN_j$, $\omega_i=N_i/N$, and $\omega'_j=N_j/N$. Then we can write*

$$
\mathbb{E}\left[\omega_{ij}\right]=\frac{\mathbb{E}\left[\nu_{ij}\right]}{\mathbb{E}\left[\omega_i\right]\mathbb{E}\left[\omega'_j\right]}+O\left(\sqrt{\frac{1}{N}}\right) \qquad (44)
$$

Proof. The proof follows by Lemma 8.2 and variance results in Lemma 8.7. ■

Let x_D and x_C respectively denote the discrete and continuous components of the vector x, with dimensions d_D and d_C. Also let $f_{X_C}(x_C)$ and $p_{X_D}(x_D)$ respectively denote density and pmf functions of these components associated with the probability measure P_X. Let $S(x,r)$ be the set of all points that are within the distance $r/2$ of x in each dimension i, i.e.

$$
S(x,r):\{x|\forall i\le d, |X_i-x_i|<r/2\}\,. \qquad (45)
$$

Denote $P_r(x):=P(x\in S(x,r))$. Then we have the following lemma.

Lemma 8.4. *Let $r<s_{\mathcal{X}}$, where $s_{\mathcal{X}}$ is the smallest possible distance in the discrete components of the support set, $\mathcal{X}$. Under the assumption **A3**, and assuming that the density functions in **A3** have bounded derivatives up to the order $q\ge 0$, we have*

$$
P_r(x)=P(X_D=x_D)r^{d_C}\left(f(x_C|x_D)+\mu(r,\gamma,q,\mathbf{C}_X)\right), \qquad (46)
$$

where

$$
\mu(r,\gamma,q,\mathbf{C}_X):=\begin{cases}O(r^\gamma)\,, & q=0\\ \sum_{i=1}^{q}C_ir^i+O(r^q)\,, & q\ge 1.\end{cases} \qquad (47)
$$

In the above equation, $\mathbf{C}_X:=(C_1,C_2,...,C_q)$, and C_i are real constants depending on the probability measure P_X.

Proof. The proof is straightforward by using (8) for the case $q=0$ (similar to (27)-(29) in [18]), and using the Taylor expansion of $f(x_C|x_D)$ for the case $q\ge 1$ (similar to (36)-(37) in [18]). ■

Lemma 8.5. *Let $H(x)=i, H(y)=j$. Under the assumptions **A1-A3**, and assuming that the density functions in **A3** have bounded derivatives up to the order $q\ge 0$, we have*

$$
\mathbb{E}\left[\omega_{ij}|E_{ij}^{\le 1}\right]=\frac{dP_{XY}}{dP_XP_Y}(x,y)+\mu(\epsilon,\gamma,q,\mathbf{C}'_{XY})+O\left(\frac{1}{\sqrt{N}}\right), \qquad (48)
$$

where $\mu(\epsilon,\gamma,q,\mathbf{C}'_{XY})$ is defined in (47).

Proof. Define $\nu_{ij}=N_{ij}/N$, and recall the definitions $\omega_{ij}=N_{ij}N/N_iN_j$, $\omega_i=N_i/N$, and $\omega'_j=N_j/N$. Using Lemma 8.2 we have

$$\mathbb{E}\left[\omega_{ij}|E_{ij}^{\leq 1}\right]=\frac{\mathbb{E}\left[\nu_{ij}|E_{ij}^{\leq 1}\right]}{\mathbb{E}\left[\omega_i|E_{ij}^{\leq 1}\right]\mathbb{E}\left[\omega'_j|E_{ij}^{\leq 1}\right]}+O\left(\frac{1}{\sqrt{N}}\right) \tag{49}$$

Assume that $H(x)=i$. Let $\mathcal{X}$ have d_C and d_D continuous and discrete components, respectively. Also let $\mathcal{Y}$ have d'_C and d'_D continuous and discrete components, respectively. Then we can write

$$\begin{aligned}\mathbb{E}\left[\omega_i|E_{ij}^{\leq 1}\right]&=\frac{1}{N}\mathbb{E}\left[N_i|E_{ij}^{\leq 1}\right]\\&=P(X\in S(x,\epsilon))\\&=P(X_D=x_D)\epsilon^{d_C}\left(f(x_C|x_D)+\mu(\epsilon,\gamma,q,\mathbf{C}_X)\right),\end{aligned} \tag{50}$$

where in the third line we have used Lemma 8.4. Similarly we can write

$$\begin{aligned}\mathbb{E}\left[\omega'_j|E_{ij}^{\leq 1}\right]&=P(Y_D=y_D)\epsilon^{d'_C}\left(f(y_C|y_D)+\mu(\epsilon,\gamma,q,\mathbf{C}_X)\right),\\ \mathbb{E}\left[\nu_{ij}|E_{ij}^{\leq 1}\right]&=P(X_D=x_D,Y_D=y_D)\epsilon^{(d_C+d'_C)}\left(f(x_C,y_C|x_D,y_D)+\mu(\epsilon,\gamma,q,\mathbf{C}_{XY})\right).\end{aligned} \tag{51}$$

Using (50) and (51) in (49) results in

$$\mathbb{E}\left[\omega_{ij}|E_{ij}^{\leq 1}\right]=\frac{P(X_D=x_D)P(Y_D=y_D)f(x_C|x_D)f(y_C|y_D)}{P(X_D=x_D,Y_D=y_D)f(x_C,y_C|x_D,y_D)}+\mu(\epsilon,\gamma,q,\mathbf{C}'_{XY})+O\left(\frac{1}{\sqrt{N}}\right), \tag{52}$$

where $\mathbf{C}'_{XY}$ depends only on P_{XY}. Now note that using Lemma 8.4, $\frac{dP_{XY}}{dP_XP_Y}(x,y)$ can be simplified as

$$\frac{dP_{XY}}{dP_XP_Y}(x,y)=\frac{\frac{dP_{XY,r}}{dr}(x,y)}{\frac{dP_{X,r}P_{Y,r}}{dr}(x,y)}=\frac{P(X_D=x_D)P(Y_D=y_D)f(x_C|x_D)f(y_C|y_D)}{P(X_D=x_D,Y_D=y_D)f(x_C,y_C|x_D,y_D)}+\mu(\epsilon,\gamma,q,\mathbf{C}''_{XY}). \tag{53}$$

Finally, using (53) in (52) gives

$$\mathbb{E}\left[\omega_{ij}|E_{ij}^{\leq 1}\right]=\frac{dP_{XY}}{dP_XP_Y}(x,y)+\mu(\epsilon,\gamma,q,\widetilde{\mathbf{C}}_{XY})+O\left(\frac{1}{\sqrt{N}}\right), \tag{54}$$

where $H(x)=i, H(y)=j$. ∎

Proof of Lemma 6.2. Lemma 6.2 is a simple consequence of Lemma 8.5. We have

$$\mathbb{E}\left[\omega_{ij}\right]=P\left(E_{ij}^{\leq 1}\right)\mathbb{E}\left[\omega_{ij}|E_{ij}^{\leq 1}\right]+P\left(E_{ij}^{>1}\right)\mathbb{E}\left[\omega_{ij}|E_{ij}^{>1}\right]. \tag{55}$$

Recall the definitions $\widetilde{\mathbf{X}}:=\left(\widetilde{X}_1,\widetilde{X}_2,...,\widetilde{X}_{L_X}\right)$ and $\widetilde{\mathbf{Y}}:=\left(\widetilde{Y}_1,\widetilde{Y}_2,...,\widetilde{Y}_{L_Y}\right)$ as the mapped $\mathbf{X}$ and $\mathbf{Y}$ points through H_1. Let $\widetilde{\mathbf{Z}}:=\widetilde{\mathbf{X}}\cup\widetilde{\mathbf{Y}}$ and $L_Z:=|\widetilde{\mathbf{Z}}|$. We first find $P\left(E_{ij}^{\leq 1}\right)$ as follows. For a fixed set $\widetilde{\mathbf{Z}}$ we have

$$\begin{aligned}P\left(E_{ij}^{\leq 1}\right)&=P\left(E_{v_i}^{=0}\cap E_{u_j}^{=0}\right)+P\left(E_{v_i}^{=0}\cap E_{u_j}^{=1}\right)+P\left(E_{v_i}^{=1}\cap E_{u_j}^{=0}\right)+P\left(E_{v_i}^{=1}\cap E_{u_j}^{=1}\right)\\&=\frac{(F-2)^{L_Z}}{F^{L_Z}}+\frac{L_Y(F-2)^{L_Z-1}}{F^{L_Z}}+\frac{L_X(F-2)^{L_Z-1}}{F^{L_Z}}+\frac{L_YL_X(F-2)^{L_Z-2}}{F^{L_Z}}\\&=1-O\left(\frac{L_Z}{F}\right)\\&\leq 1-O\left(\frac{L_X+L_Y}{F}\right)\\&=1-O\left(\frac{1}{\epsilon^dN}\right).\end{aligned} \tag{56}$$

Now note that the second term in (55) is the bias due to collision of H_2, and similar to (37) it is upper bounded by $O\big(\frac{1}{\epsilon^d N}\big)$. Thus, (56) and (55) give rise to

$$\mathbb{E}\left[\omega_{ij}\right] = \frac{dP_{XY}}{dP_X P_Y}(x, y) + \mu(\epsilon, \gamma, q, \widetilde{\mathbf{C}}_{XY}) + O\bigg(\frac{1}{\sqrt{N}}\bigg) + O\bigg(\frac{1}{\epsilon^d N}\bigg). \tag{57}$$

which completes the proof.

■

In the following lemma we make a relation between the bias of an estimator and the bias of a function of that estimator.

Lemma 8.6. *Assume that $g(x) : \mathcal{X} \to \mathbb{R}$ is infinitely differentiable. If $\widehat{Z}$ is a random variable estimating a constant Z with the bias $\mathbb{B}[\widehat{Z}]$ and the variance $\mathbb{V}[\widehat{Z}]$, then the bias of $g(\widehat{Z})$ can be written as*

$$\mathbb{E}\left[g(\widehat{Z}) - g(Z)\right] = \sum_{i=1}^{\infty} \xi_i \Big(\mathbb{B}\Big[\widehat{Z}\Big]\Big)^i + O\bigg(\sqrt{\mathbb{V}\Big[\widehat{Z}\Big]}\bigg), \tag{58}$$

where ξ_i are real constants.

Proof.

$$\begin{aligned}
\mathbb{E}\left[g\Big(\widehat{Z}\Big) - g(Z)\right] &= g\Big(\mathbb{E}\Big[\widehat{Z}\Big]\Big) - g(Z) + \mathbb{E}\left[g\Big(\widehat{Z}\Big) - g\Big(\mathbb{E}\Big[\widehat{Z}\Big]\Big)\right] \\
&= \sum_{i=1}^{\infty}\Big(\mathbb{E}\Big[\widehat{Z}\Big] - Z\Big)^i \frac{g^{(i)}(Z)}{i!} + O\Big(\mathbb{E}\Big[\Big|g\Big(\widehat{Z}\Big) - g\Big(\mathbb{E}\Big[\widehat{Z}\Big]\Big)\Big|\Big]\Big) \\
&= \sum_{i=1}^{\infty} \xi_i \Big(\mathbb{B}\Big[\widehat{Z}\Big]\Big)^i + O\bigg(\sqrt{\mathbb{V}\Big[\widehat{Z}\Big]}\bigg).
\end{aligned} \tag{59}$$

In the second line we have used Taylor expansion for the first term, and triangle inequality for the second term. In the third line we have used the definition $\xi_i := g^{(i)}(Z)/i!$, and the Cauchy-Schwarz inequality for the second term.

■

In the following we obtain the expectation of the first term in (20) and prove Theorem 3.1.

Proof of Theorem 3.1. Recall that N'_i and M'_j respectively are defined as the number of the input points $\mathbf{X}$ and $\mathbf{Y}$ mapped to the buckets $\widetilde{X}_i$ and $\widetilde{Y}_j$ using H_1. Similarly, N'_{ij} is defined as the number of input pairs $(\mathbf{X}, \mathbf{Y})$ mapped to the bucket pair $\Big(\widetilde{X}_i, \widetilde{Y}_j\Big)$ using H_1. Define the notations $r(i) := H_2^{-1}(i)$ for $i \in \mathcal{F}$ and $s(x) := H_1(x)$ for $x \in \mathcal{X} \cup \mathcal{Y}$. Then from (52) since there is no collision of mapping with H_2 into v_i and u_j we have

$$\mathbb{E}\left[\frac{N'_{s(x)s(y)}N}{N'_{s(x)}N'_{s(y)}}\right] = \frac{dP_{XY}}{dP_X P_Y}(x, y) + \mu(\epsilon, \gamma, q, \widetilde{\mathbf{C}}_{XY}) + O\bigg(\frac{1}{\sqrt{N}}\bigg), \tag{60}$$

By using (56) and defining $\tilde{h}(x) = \tilde{g}(x)/x$ we can simplify the first term of (20) as

$$
\begin{aligned}
\sum_{i,j\in\mathcal{F}} P\Big(E_{ij}^{\leq 1}\Big)\,\mathbb{E}\left[\mathbb{1}_{E_{ij}}\omega_i\omega'_j\widetilde{g}(\omega_{ij})\Big|E_{ij}^{\leq 1}\right] &= \left(1-O\Big(\frac{1}{\epsilon^d N}\Big)\right)\sum_{i,j\in\mathcal{F}}\mathbb{E}\left[\mathbb{1}_{E_{ij}}\omega_i\omega'_j\widetilde{g}(\omega_{ij})\Big|E_{ij}^{\leq 1}\right]\\
&= \sum_{i,j\in\mathcal{F}}\mathbb{E}\left[\mathbb{1}_{E_{ij}}\frac{N_iM_j}{N^2}\widetilde{g}\Big(\frac{N_{ij}N}{N_iM_j}\Big)\Big|E_{ij}^{\leq 1}\right]+O\Big(\frac{1}{\epsilon^d N}\Big)\\
&= \sum_{i,j\in\mathcal{F}}\mathbb{E}\left[\mathbb{1}_{E_{ij}}\frac{N'_{r(i)}M'_{r(j)}}{N^2}\widetilde{g}\left(\frac{N'_{r(i)r(j)}N}{N'_{r(i)}M'_{r(j)}}\right)\right]+O\Big(\frac{1}{\epsilon^d N}\Big)\\
&= \sum_{i,j\in\mathcal{F}}\mathbb{E}\left[\mathbb{1}_{E_{ij}}\frac{N'_{r(i)r(j)}}{N}\widetilde{h}\left(\frac{N'_{r(i)r(j)}N}{N'_{r(i)}M'_{r(j)}}\right)\right]+O\Big(\frac{1}{\epsilon^d N}\Big)\\
&= \frac{1}{N}\sum_{i,j\in\mathcal{F}}\mathbb{E}\left[N'_{r(i)r(j)}\widetilde{h}\left(\frac{N'_{r(i)r(j)}N}{N'_{r(i)}M'_{r(j)}}\right)\right]+O\Big(\frac{1}{\epsilon^d N}\Big) \qquad (61)\\
&= \frac{1}{N}\mathbb{E}\left[\sum_{i,j\in\mathcal{F}}N'_{r(i)r(j)}\widetilde{h}\left(\frac{N'_{r(i)r(j)}N}{N'_{r(i)}M'_{r(j)}}\right)\right]+O\Big(\frac{1}{\epsilon^d N}\Big)\\
&= \frac{1}{N}\mathbb{E}\left[\sum_{i=1}^{N}\widetilde{h}\left(\frac{N'_{s(X)s(Y)}N}{N'_{s(X)}M'_{s(Y)}}\right)\right]+O\Big(\frac{1}{\epsilon^d N}\Big)\\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\widetilde{h}\left(\frac{N'_{s(X)s(Y)}N}{N'_{s(X)}M'_{s(Y)}}\right)\right]+O\Big(\frac{1}{\epsilon^d N}\Big)\\
&= \mathbb{E}_{(X,Y)\sim P_{XY}}\left[\mathbb{E}\left[\widetilde{h}\left(\frac{N'_{s(X)s(Y)}N}{N'_{s(X)}M'_{s(Y)}}\right)\Big|X=x,Y=y\right]\right]+O\Big(\frac{1}{\epsilon^d N}\Big)\\
&= \mathbb{E}_{(X,Y)\sim P_{XY}}\left[\frac{dP_{XY}}{dP_XP_Y}\right]+\mu(\epsilon,\gamma,q,\overline{\mathbf{C}}_{XY})+O\Big(\frac{1}{\sqrt{N}}\Big)+O\Big(\frac{1}{\epsilon^d N}\Big). \qquad (62)
\end{aligned}
$$

(63)

(61) is due to the fact that $N'_{r(i)r(j)}=0$ if there is no edge between v_i and u_j. Also, (62) is due to (60).

From (62) and (20) we obtain

$$
\mathbb{E}\left[\widetilde{I}(X,Y)\right]=\mathbb{E}\left[\sum_{e_{ij}\in E_G}\omega_i\omega'_j\widetilde{g}(\omega_{ij})\right]=\mathbb{E}_{(X,Y)\sim P_{XY}}\left[\frac{dP_{XY}}{dP_XP_Y}\right]+\mu(\epsilon,\gamma,q,\overline{\mathbf{C}}_{XY})+O\Big(\frac{1}{\sqrt{N}}\Big)+O\Big(\frac{1}{\epsilon^d N}\Big). \tag{64}
$$

Finally using Lemma 8.6 results in (10).

■

A. VARIANCE PROOF

In this section we first prove bounds on the variances of the edge and vertex weights and then we provide the proof of Theorem 3.2.

Lemma 8.7. *Under the assumptions* ***A1-A4****, the following variance bounds hold true.*

$$
\mathbb{V}[\omega_i]\leq O\Big(\frac{1}{N}\Big),\qquad \mathbb{V}[\omega'_j]\leq O\Big(\frac{1}{N}\Big),\qquad \mathbb{V}[\omega_{ij}]\leq O\Big(\frac{1}{N}\Big),\qquad \mathbb{V}[\nu_{ij}]\leq O\Big(\frac{1}{N}\Big). \tag{65}
$$

Proof. Here we only provide the variance proof of ω_i. The variance bounds of ω'_j, ω_{ij} and ν_{ij} can be proved in the same way. The proof is based on Efron-Stein inequality. Define $Z_i:=(X_i,Y_i)$. For using the Efron-Stein inequality on $\mathbf{Z}:=(Z_1,...,Z_N)$, we consider another independent copy of $\mathbf{Z}$ as $\mathbf{Z}':=(Z'_1,...,Z'_N)$ and define $\mathbf{Z}^{(i)}:=(Z_1,...,Z_{i-1},Z'_i,Z_{i+1},...,Z_N)$. Define $\omega_i(\mathbf{Z})$ as the weight of vertex v_i in the dependence graph constructed by the set $\mathbf{Z}$. By applying Efron-Stein inequality [23] we have

$$
\begin{aligned}
\mathbb{V}[\omega_i] &\le \frac{1}{2}\sum_{i=1}^{N}\mathbb{E}\left[\left(\omega_i(\mathbf{Z}) - \omega_i\left(\mathbf{Z}^{(j)}\right)\right)^2\right] \\
&= \frac{1}{2N^2}\sum_{i=1}^{N}\mathbb{E}\left[\left(N_i(\mathbf{Z}) - N_i\left(\mathbf{Z}^{(j)}\right)\right)^2\right] \\
&\le \frac{1}{2N^2}O(N) \\
&\le O\left(\frac{1}{N}\right).
\end{aligned} \tag{66}
$$

In the third line we have used the fact that the absolute value of $N_i(\mathbf{Z}) - N_i\left(\mathbf{Z}^{(j)}\right)$ is at most 1.

■

Proof of Theorem 3.2 . We follow similar steps as the proof of Lemma 8.7. Define $\widehat{I}_g(\mathbf{Z})$ as the mutual information estimation using the set $\mathbf{Z}$. By applying Efron-Stein inequality we have

$$
\begin{aligned}
\mathbb{V}\left[\widehat{I}(X,Y)\right] &\le \frac{1}{2}\sum_{k=1}^{N}\mathbb{E}\left[\left(\widehat{I}(\mathbf{Z}) - \widehat{I}(\mathbf{Z}^{(k)})\right)^2\right] \\
&\le \frac{N}{2}\mathbb{E}\left[\left(\sum_{e_{ij}\in E_G}\omega_i(\mathbf{Z})\,\omega'_j(\mathbf{Z})\,\widetilde{g}(\omega_{ij}(\mathbf{Z})) - \sum_{e_{ij}\in E_G}\omega_i\left(\mathbf{Z}^{(k)}\right)\omega'_j\left(\mathbf{Z}^{(k)}\right)\widetilde{g}(\omega_{ij})\left(\mathbf{Z}^{(k)}\right)\right)^2\right] \\
&= \frac{N}{2N^4}\mathbb{E}\left[\left(\sum_{e_{ij}\in E_G}N_i(\mathbf{Z})\,M_j(\mathbf{Z})\,\widetilde{g}\left(\frac{N_{ij}(\mathbf{Z})\,N}{N_i(\mathbf{Z})\,M_j(\mathbf{Z})}\right) - \sum_{e_{ij}\in E_G}N_i\left(\mathbf{Z}^{(k)}\right)M_j\left(\mathbf{Z}^{(k)}\right)\widetilde{g}\left(\frac{N_{ij}\left(\mathbf{Z}^{(k)}\right)N}{N_i\left(\mathbf{Z}^{(k)}\right)M_j\left(\mathbf{Z}^{(k)}\right)}\right)\right)^2\right]
\end{aligned} \tag{67}
$$

$$
\le \frac{1}{2N^3}\mathbb{E}\left[\left(\Sigma_{n_1} + \Sigma_{n_2} + \Sigma_{m_1} + \Sigma_{m_1} + D_{n_1m_1} + D_{n_2m_2}\right)^2\right]. \tag{68}
$$

Note that in equation (68), when (X_k, Y_k) is resampled, at most two of N_i for $i \in \mathcal{F}$ are changed exactly by one (one decrease and the other increase). The same statement holds true for M_j. Let these vertices be v_{n_1}, v_{n_2}, v_{m_1} and v_{m_2}. Also the pair collision counts N_{ij} are fixed except possibly $N_{n_1m_1}$ and $N_{n_2m_2}$ that may change by one. So, in the fourth line Σ_{n_1} and Σ_{n_2} account for the changes in MI estimation due to the changes in N_{n_1} and N_{n_2}, and Σ_{m_1} and Σ_{m_2} account for the changes in M_{m_1} and v_{m_2}, respectively. Finally $D_{n_1m_1}$ and $D_{n_2m_2}$ account for the changes in MI estimation due to the changes in $N_{n_1m_1}$ and $N_{n_2m_2}$. For example, Σ_{n_1} is precisely defined as follows:

$$
\Sigma_{n_1} := \sum_{j:e_{mj}\in E_G} N_m M_j \widetilde{g}\left(\frac{N_{mj}N}{N_m N_j}\right) - (N_m+1)M_j\widetilde{g}\left(\frac{N_{mj}N}{(N_m+1)M_j}\right) \tag{69}
$$

where we have used the notations N_i and $N_i^{(k)}$ instead of $N_i(\mathbf{Z})$ and $N_i(\mathbf{Z}^{(k)})$ for simplicity. Now note that by assumption **A4** we have

$$
\begin{aligned}
\left|\widetilde{g}\left(\frac{N_{mj}N}{N_m M_j}\right) - \widetilde{g}\left(\frac{N_{mj}N}{(N_m+1)M_j}\right)\right| &\le G_g\left|\frac{N_{mj}N}{N_m M_j} - \frac{N_{mj}N}{(N_m+1)M_j}\right| \\
&\le O\left(\frac{N_{mj}N}{N_m^2 M_j}\right).
\end{aligned} \tag{70}
$$

Thus, using (70), Σ_{n_1} can be upper bounded as follows

$$
\Sigma_{n_1} \le \sum_{j:e_{mj}\in E_G} O\left(\frac{N_{mj}N}{N_m^2}\right) = O\left(\frac{N}{N_m}\right) \le O(N). \tag{71}
$$

It can similarly be shown that N_{n_2}, Σ_{m_1}, Σ_{m_2}, $D_{n_1m_1}$ and $D_{n_2m_2}$ are upper bounded by $O(N)$. Thus, (68) simplifies as follows

$$\mathbb{V}\left[\widehat{I}(X,Y)\right] \leq \frac{36O(N^2)}{2N^3} = O(\frac{1}{N}). \tag{72}$$

■

C. Optimum MSE Rates of EDGE

In this short section we prove Theorem 4.1.

Proof of Theorem 4.1. The proof simply follows by using the ensemble theorem in ([17], Theorem 4) with the parameters $\psi_i(t) = t^i$ and $\phi_{i,d}(N) = N^{-i/2d}$ for the bias result in Theorem 3.1. Thus, the following weighted ensemble estimator (EDGE) can achieve the optimum parametric MSE convergence rate of $O(1/N)$ for $q \geq d$.

$$\widehat{I}_w := \sum_{t \in \mathcal{T}} w(t)\widehat{I}_{\epsilon(t)}, \tag{73}$$

■